Mamba-MSQNet: A Fast and Efficient Model for Animal Action Recognition

1st Edoardo Fazzari
The Biorobotics Institute &
Department of Excellence in Robotics and AI
Scuola Superiore Sant'Anna
Pisa, Italy
edoardo.fazzari@santannapisa.it

2nd Donato Romano
The Biorobotics Institute &
Department of Excellence in Robotics and AI
Scuola Superiore Sant'Anna
Pisa, Italy
donato.romano@santannapisa.it

3rd Fabrizio Falchi
The Biorobotics Institute
Scuola Superiore Sant'Anna &
Inf. Science and Technologies Inst. "A.Faedo"
Consiglio Nazionale delle Ricerche
Pisa, Italy
fabrizio.falchi@cnr.it

4th Cesare Stefanini
The Biorobotics Institute &
Department of Excellence in Robotics and AI
Scuola Superiore Sant'Anna
Pisa, Italy
cesare.stefanini@santannapisa.it

***Abstract*—Animal action recognition is crucial for assessing animal well-being in agriculture and environmental monitoring. Recent advancements in this field rely on computer vision technologies. However, many current applications are restricted to recognizing actions within a single animal species or a limited set of actions, resulting in highly specific models and lacking generality. When addressing a broader range of actions and species, transformer models are typically required, which demand significant computational and processing power, potentially limiting their practical use. In this work, we introduce a deep learning model based on selective state spaces designed to reduce the computational cost of MSQNet, the current state-of-the-art model for action recognition in the Animal Kingdom dataset. Our approach achieves superior results with fewer parameters and lower FLOPs, thereby enhancing efficiency without compromising performance. Code available on https://github.com/edofazza/mamba-msqnet.**

***Index Terms*—animal action recognition, deep learning, selective state spaces, computer vision, mamba, msqnet**

I. Introduction

Animal Action Recognition (AAR) is the automated process of identifying and classifying specific animal behaviors through the analysis of video or sensor data [1]. By monitoring and analyzing these behaviors, researchers can gain deeper insights into the physical and psychological well-being of animals, allowing for timely interventions in cases of distress or illness [2]. Recognizing signs of pain, discomfort, or unusual behavior patterns in livestock, for instance, can lead to quicker diagnoses and treatments, thereby improving their overall health and quality of life [3]. Additionally, pest control in agriculture benefits from such studies, as understanding the behavior of pest species can aid in developing effective pest control strategies, helping farmers manage crop damage and reduce the need for harmful pesticides [4].

In recent years, Animal Action Recognition has significantly advanced due to developments in machine learning and computer vision, allowing the identification of a wide range of actions, from simple movements to complex behavioral patterns [5]. These technological advancements have led to more accurate and efficient behavior analysis, facilitating real-time monitoring and intervention. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been particularly influential, enabling the processing of vast amounts of data and the extraction of intricate features relevant to specific actions [6]. However, existing applications of animal action recognition face several limitations: 1) the datasets used are often small and contain a limited number of actions, reducing the models' generalization ability and transferability [7]; 2) there is a limited number and diversity of animals represented, as many existing datasets are designed to study specific groups of animals, such as only cows or only sheep [8], [9]. Recently, a public dataset called Animal Kingdom [10] was released to address these limitations. Animal Kingdom includes over 50 hours of clips showcasing 850 different species and 140 action classes, all labeled to facilitate multi-action recognition.

The state-of-the-art in action classification for Animal Kingdom is represented by the Multi-modal Semantic Query Network (MSQNet) [11], a transformer-based object detection framework effectively utilizes both visual and textual

This is a post-print version. Published version can be found at 2024 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), DOI: 10.1109/MetroAgriFor63043.2024.10948802

This research was carried out in the framework of the EU H2020 FETOPEN Project 'Robocoenosis - ROBOts in cooperation with a bioCOENOSIS' [899520]. Funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

modalities to enhance the representation of action classes, thereby improving mean average precision (mAP). However, the complexity and resource demands of transformer-based networks can limit their practical use in real-world scenarios with limited resources, for this reason we thought about how to lightening this model.

Recently, Gu and Dao [12] introduced a more efficient approach that substitutes transformers with a linear-time sequence modeling technique using selective state spaces blocks, termed Mamba. Mamba offers faster inference and improved performance with lower computational demands. Originally designed to replace transformer layers in Foundation Models (FMs) for Large Language Models (LLMs), Mamba's efficiency has rapidly made it a valuable tool in computer vision applications as well [13].

In this study, we tackle the challenge of reducing the computational overhead associated with the transformer layers and encoder in MSQNet. To address this issue, we introduce Mamba blocks and integrate VideoMamba, a pretrained video encoder leveraging these optimized blocks. This modification results in a refined architecture we termed Mamba-MSQNet, where we ensure seamless integration with MSQNet's other components. Our approach significantly reduces model parameters and FLOPs while demonstrating a slight performance enhancement over the original MSQNet.

II. Related Work

A. Animal Action Recognition

Application of Animal Action Recognition (AAR) in agriculture and environmental monitoring has been extensively studied in the literature for its ability to distinguish animal behaviors and draw conclusions related to their health [14]. With the advances in deep learning, there has been a growing interest in applying neural networks to this field. Typically, deep learning models in AAR are used on two types of data: sensor data or images and videos.

Sensor data is usually collected from accelerometers, gyroscopes, and Global Navigation Satellite System (GNSS) devices, which are attached to animals using collar tags or other equipment [15]. This data can be classified using various deep learning techniques. For instance, a simple multi-layer perceptron can be employed, as demonstrated by [16] in their analysis of cattle behaviors such as drinking, grazing, walking, ruminating, and resting. Alternatively, convolutional and recurrent neural networks have been used for cow monitoring [17]. Recently, convolutional neural networks (CNNs) have become particularly prominent. A notable example is a CNN based on two-channel temporal and spatial (TCTS) networks, which was introduced to leverage 3D sensor data for identifying behaviors in lactating sows [18].

Image and video data can be effectively analyzed using Computer Vision strategies. The most straightforward approach is to employ a convolutional neural network for direct action classification [19]. However, advanced techniques often leverage object detection or pose estimation for more nuanced insights. For instance, utilizing YOLO (You Only Look Once) has facilitated the investigation of abnormal behaviors in pigs, such as ear and tail biting [20]. Additionally, object detection enables the quantification of time animals spend foraging and feeding, providing valuable insights into their health. On the other hand, pose estimation, through the identification of skeletal structures overlaid on the animal, can not only aid in identifying discrete behavioral motifs but also capture a hierarchical representation of their usage [21].

Although the importance of the research topic, not many public datasets exists for action recognition [2], [22], [23]. The most important is Animal Kingdom [10] which consist of video clips from multiple animals over a large number of possible actions (further details in III). The authors of this dataset proposed a possible solution to based on some convolutional architecture, which although performed poorly. The state-of-the-art was achieved subsequently by MSQNet [11], a multi-modal architecture that leverages three modalities: video information, frame (i.e., image) information and text information from the multi-class labels. The video information is sent to the video encoder, which extracts the spatio-temporal features, the CLIP encoders processes the image and the textual information which are lately merged by the query encoder in order to make the multi-modal query sent to the multi-modal transformer decoder which received in inputs the video encoding and transform this information to make multi-label classification with a feed-forward network.

B. Mamba Architectures

The Transformer architecture has become almost ubiquitous in foundation models, finding applications across numerous research fields [24]. Each Transformer layer is composed of an attention mechanism and a feed-forward neural network [25]. The attention module allows the model to focus on various parts of the input sequence by computing a weighted sum of input features, enabling it to capture dependencies regardless of their distance within the sequence. However, the quadratic complexity of the attention module makes Transformers computationally demanding for processing long sequences. To address this limitation, State Space Models (SSMs) have been introduced [26], specifically designed to model long-range dependencies with linear complexity. Leveraging SSMs, Mamba [12] has been developed. Mamba features a data-dependent SSM layer and a selection mechanism using parallel scan, allowing it to surpass Transformers in processing long sequences.

Although Mamba [12] was originally introduced for language processing, it has found applications in computer vision as well. Zhu et al. [27] introduced Vision Mamba (Vim), which utilizes bidirectional SSM for data-dependent global visual context modeling and position embedding for location-aware visual understanding. Remarkably, Vim achieves the same level of power as Vision Transformers without relying on attention mechanisms, while maintaining subquadratic-time computation and linear memory complexity. Initially, Vim was limited to classification, detection, and segmentation tasks on images. However, with the introduction of VideoMamba

[13], its capabilities were extended to videos, transitioning Vim from 2D to 3D. VideoMamba offers numerous pretrained models that outperform Transformer architectures in short- and long-term video understanding and in multi-modality video understanding. Figure 1 details the components of a Transformer layer, Mamba and Vim.

III. Method

A. Animal Kingdom Dataset

The Animal Kingdom Multi-Label Action Recognition dataset [10] comprises 30,100 clips totaling 50 hours of recording. This dataset identifies 140 action classes organized into 16 categories: affection, aggressive, communication, death, defensive, feeding, general, life cycle, maintenance, movement, prey, resting, sexual, shelter, social, and transport. It features 850 unique species, including mammals, reptiles, amphibians, birds, fishes, and insects, providing high variability. The diversity is further enhanced by presenting the same animal species performing different actions and in varied scenarios across multiple clips.

B. Mamba-MSQNet

In our Mamba implementation of MSQNet [11], we focused on substituting all trainable transformer components of the network. Consequently, the CLIP image and text encoders [28] were preserved and left unmodified, but we replaced the *video encoder* and the *multi-modal transformer decoder*.

The MSQNet [11] achieves its best results using the TimeSformer [29] video encoder, which leverages self-attention for spatiotemporal feature learning directly from sequences of frame-level patches. This encoder, pretrained on the Kinetics-400 (K400) dataset [30], is a crucial, trainable part of the MSQNet network [11]. In our research, we completely replaced this component with VideoMamba [13]. Unlike TimeSformer, which employs divided spatiotemporal attention, VideoMamba uses Vision Mamba (VIM) layers [27] to process spatiotemporal tokens with linear complexity. To ensure a fair comparison with the standard MSQNet [11], VideoMamba [13] was also pretrained on K400 [30] and achieved an accuracy of 81.9% on video inputs of 16 frames, slightly higher than TimeSformer's 78% [29].

The multi-modal transformer decoder was replaced with a more effective system comprising three key components: 1) a Multi-Layer Perceptron (MLP) that integrates the fused information from the projected image-encoded data and learnable label embedding with the global encoder output, adding a second layer of information fusion; 2) this fused output is then processed through 16 Mamba blocks. This number follows the recommendation by the Mamba block's inventor [12], which suggests using two Mamba blocks for each head in a transformer block. Since the MSQNet transformer decoder has 8 heads, 16 Mamba blocks are used; 3) finally, an additional MLP layer adjusts the output to match the dimensionality of the fully-connected network (FFN), producing the final logits used for classification. The complete Mamba-MSQNet architecture is illustrated in Figure 2.

TABLE I
Comparison in terms of floating point operations (FLOPs) and parameters between MSQNet [11] and our Mamba implementation.

Model	FLOPs (T)	PARAM (M)
MSQNet [11]	0.677	252
Mamba-MSQNet (our)	**0.487**	**190**

The benefits of our architecture, in terms of FLOPs and parameters, compared to the standard version, are detailed in Table I. Our model achieves a reduction of 190G FLOPs, which is a 29% decrease overall. Excluding the CLIP image encoder [28], which we did not modify and which accounts for 281G FLOPs, the reduction in FLOPs for the remaining architecture is 48%. Similarly, the total number of parameters is reduced by 25%, and by 38% when excluding the image encoder.

The training procedure we used for Mamba-MSQNet closely followed the methodology outlined by the authors of the standard MSQNet [11]. Each video frame set consisted of 16 frames, randomly selected from each clip, and underwent the same augmentation transformations as MSQNet. The primary and only difference lay in the number of training epochs. While the standard MSQNet was trained for 100 epochs, we adopted a two-phase training approach for Mamba-MSQNet. In the first phase, we froze the video encoder and trained the model for 100 epochs. In the second phase, we unfroze the encoder and continued training for an additional 150 epochs. This strategy was implemented to prevent the propagation of gradients caused by the initial random initialization of other layers, which can lead to feature distortion [31].

IV. Results

To evaluate the multi-label classification of actions in video clips, we employed mean Average Precision (mAP) as our primary evaluation metric. This metric has been previously used to assess the performance of the standard MSQNet [11] and is recommended by the authors of the Animal Kingdom dataset [10]. mAP is particularly advantageous for multi-class problems like action recognition because it considers performance across all classes, including those with imbalanced distributions or rare occurrences. This makes it a more comprehensive evaluation metric compared to overall accuracy, which can be misleading in scenarios with imbalanced class distributions.

Table II presents the results of MSQNet [11] and our proposed Mamba-MSQNet. Our implementation achieved a mAP of 74.2%, slightly higher than the 73.1% obtained by the standard version. Although the improvement in classification performance is modest, the differences highlighted in Table I demonstrate that our model achieves comparable performance to the standard transformer-based network with fewer parameters and lower FLOPs. This is achieved through the introduction of linear-time sequence modeling with Selective State Spaces, marking a significant advancement in efficiency,

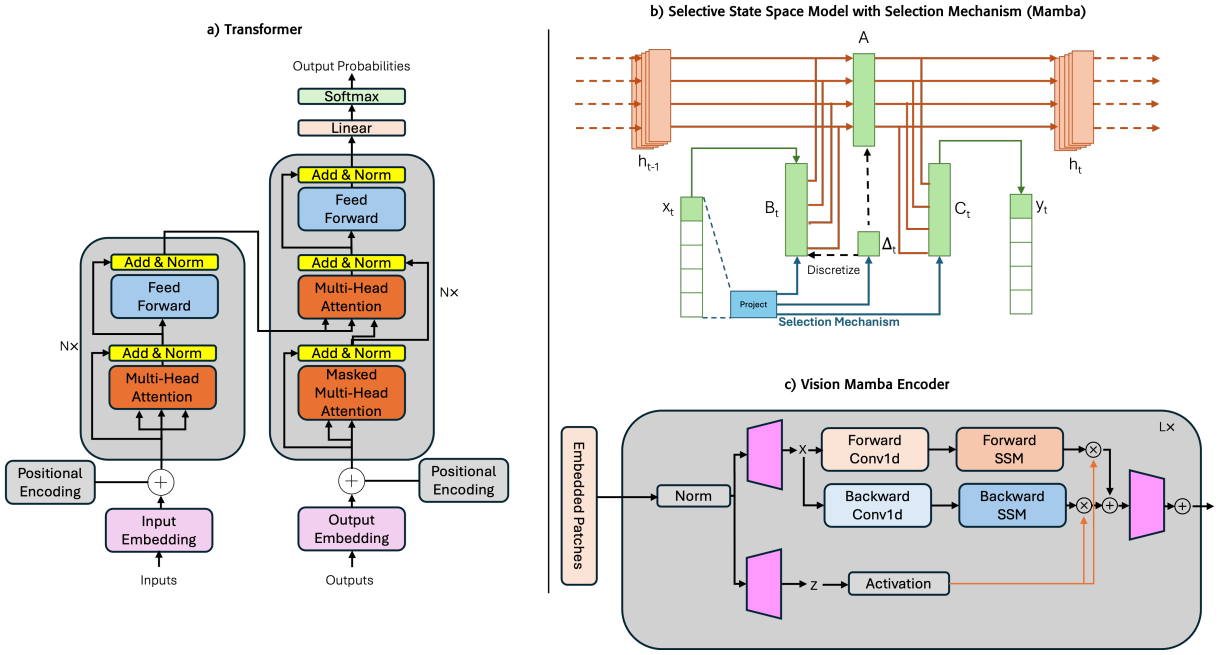

Fig. 1. Architecture details for the Transformer layer (a) [25], the Selective Space Model layer with Selection Mechanism employed for Mamba [12] (b), and the encoder layer structure for Vim [27] (c). The pink trapezoids in Video Mamba encoder are projection layers.

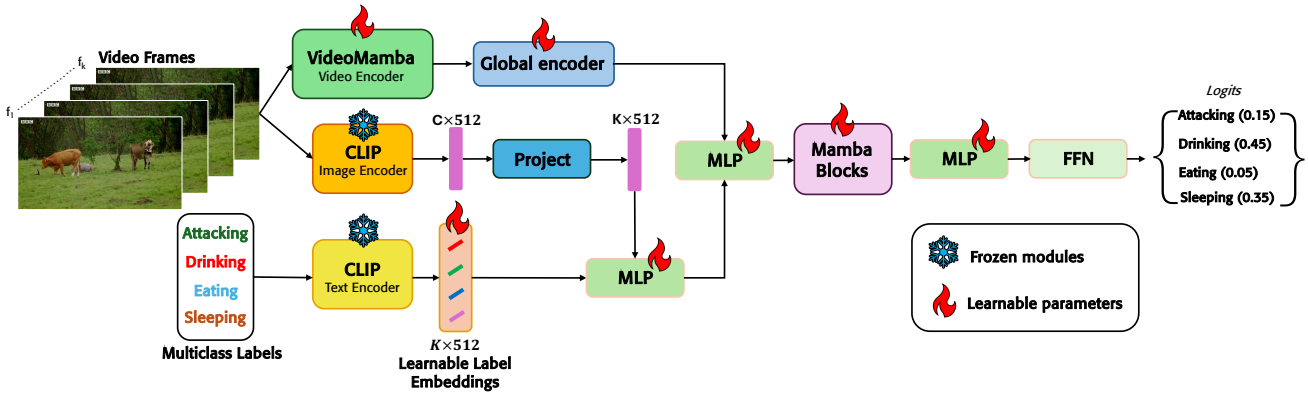

Fig. 2. Architecture of our Mamba-MSQNet for multi-modal multi-label action recognition. The actions indicated in the logits and multiclass labels blocks are just few examples of the 140 actions present in Animal Kingdom [10]. Also, it is worth to notice that the CLIP text encoder is performed only once to obtain the initial label embeddings.

TABLE II
RESULTS IN TERMS OF MEAN AVERAGE PRECISION (MAP).

Model	mAP
CARe-X3D [10]	25.2
MSQNet [11]	73.1
Mamba-MSQNet (our)	**74.2**

especially if we consider that the entire architecture structure was unchanged.

V. DISCUSSION AND CONCLUSION

In recent years, the use of deep learning to understand animal actions has become a vibrant area of research. Understanding animal behavior is essential for agricultural activities such as disease identification and prevention, analyzing abnormal and violent behavior, and monitoring pests. However, many studies lack generalizability, and the only model demonstrating significant value, MSQNet [11], requires substantial computational power for inferring multi-label actions, limiting its usability in resource-constrained environments.

To address this limitation, we introduce an innovative approach by incorporating the latest advancements in Selective State Models into MSQNet [11], utilizing Mamba as an alternative to Transformer layers. We are the first to employ a multi-modal Mamba architecture for animal action recognition. This new architecture, named Mamba-MSQNet, enhances MSQNet [11] by replacing the TimeSformer video encoder [29] with a VideoMamba [13] encoder pretrained on

the same dataset, and substituting the transformer decoder with Mamba blocks.

Our results demonstrate a slight increase in the mean average precision of our model. More significantly, we achieved a 48% reduction in floating point operations and a 38% decrease in model parameters, excluding the CLIP encoders [28]. This improved efficiency is crucial for applications in agriculture and environmental monitoring, where computational resources are often limited. The linear-time sequence modeling with Selective State Spaces enables real-time or near-real-time processing, making our model more feasible for deployment in field conditions.

This capability is crucial for tasks such as real-time animal behavior monitoring, crop health assessment, and environmental surveillance, where timely and accurate action recognition can lead to more responsive and informed decision-making. Our advancements not only contribute to the field of action recognition but also open up new possibilities for practical applications in challenging and resource-limited scenarios, thanks to the introduction of Mamba layers.

In future work, we aim to replace the CLIP encoders with Mamba-based models to further accelerate computation. However, current research on Mamba-based models that can effectively mimic CLIP's performance is limited [32]. Despite this, our Mamba-MSQNet already presents a promising solution for enhancing precision and reducing computational demands.

ACKNOWLEDGEMENT

This research was partially conducted within the scope of the H2020 FETOPEN Project 'Robocoenosis – ROBOts in cooperation with a bioCOENOSIS' [899520]. The funder played no role in the design of the study, data collection or analysis, the decision to publish, or the manuscript preparation.

REFERENCES

[1] E. Fazzari, D. Romano, F. Falchi, and C. Stefanini, "Animal behavior analysis methods using deep learning: A survey," 2024.

[2] J. M. Graving, D. Chae, H. Naik, L. Li, B. Koger, B. R. Costelloe, and I. D. Couzin, "Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning," *Elife*, vol. 8, p. e47994, 2019.

[3] S. Manoharan, "Embedded imaging system based behavior analysis of dairy cow," *Journal of Electronics*, vol. 2, no. 02, pp. 148–154, 2020.

[4] Q. A. Mendoza, L. Pordesimo, M. Neilsen, P. Armstrong, J. Campbell, and P. T. Mendoza, "Application of machine learning for insect monitoring in grain facilities," *AI*, vol. 4, no. 1, pp. 348–360, 2023.

[5] S. R. Nilsson, N. L. Goodwin, J. J. Choong, S. Hwang, H. R. Wright, Z. C. Norville, X. Tong, D. Lin, B. S. Bentzley, N. Eshel *et al.*, "Simple behavioral analysis (simba)–an open source toolkit for computer classification of complex social behaviors in experimental animals," *BioRxiv*, pp. 2020–04, 2020.

[6] C. Segalin, J. Williams, T. Karigo, M. Hui, M. Zelikowsky, J. J. Sun, P. Perona, D. J. Anderson, and A. Kennedy, "The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice," *Elife*, vol. 10, p. e63720, 2021.

[7] L. von Ziegler, O. Sturman, and J. Bohacek, "Big behavior: challenges and opportunities in a new era of deep behavior profiling," *Neuropsychopharmacology*, vol. 46, no. 1, pp. 33–44, 2021.

[8] N. Kleanthous, A. Hussain, W. Khan, J. Sneddon, and P. Liatsis, "Deep transfer learning in sheep activity recognition using accelerometer data," *Expert Systems with Applications*, vol. 207, p. 117925, 2022.

[9] J. Li, C. Xu, L. Jiang, Y. Xiao, L. Deng, and Z. Han, "Detection and analysis of behavior trajectory for sea cucumbers based on deep learning," *Ieee Access*, vol. 8, pp. 18 832–18 840, 2019.

[10] X. L. Ng, K. E. Ong, Q. Zheng, Y. Ni, S. Y. Yeo, and J. Liu, "Animal kingdom: A large and diverse dataset for animal behavior understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 023–19 034.

[11] A. Mondal, S. Nag, J. M. Prada, X. Zhu, and A. Dutta, "Actor-agnostic multi-label action recognition with multi-modal query," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2023, pp. 784–794.

[12] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[13] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "Videomamba: State space model for efficient video understanding," *arXiv preprint arXiv:2403.06977*, 2024.

[14] C. Chen, W. Zhu, and T. Norton, "Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning," *Computers and Electronics in Agriculture*, vol. 187, p. 106255, 2021.

[15] R. Arablouei, Z. Wang, G. J. Bishop-Hurley, and J. Liu, "Multimodal sensor data fusion for in-situ classification of animal behavior using accelerometry and gnss data," *Smart Agricultural Technology*, vol. 4, p. 100163, 2023.

[16] R. Arablouei, L. Wang, L. Currie, J. Yates, F. A. Alvarenga, and G. J. Bishop-Hurley, "Animal behavior classification via deep learning on embedded systems," *Computers and Electronics in Agriculture*, vol. 207, p. 107707, 2023.

[17] T.-H. Dang, N.-H. Dang, V.-T. Tran, and W.-Y. Chung, "A lorawan-based smart sensor tag for cow behavior monitoring," in *2022 IEEE Sensors*. IEEE, 2022, pp. 1–4.

[18] Z. Pan, H. Chen, W. Zhong, A. Wang, and C. Zheng, "A cnn-based animal behavior recognition algorithm for wearable devices," *IEEE Sensors Journal*, vol. 23, no. 5, pp. 5156–5164, 2023.

[19] M. Perez and C. Toler-Franklin, "Cnn-based action recognition and pose estimation for classifying animal behavior from videos: a survey," *arXiv preprint arXiv:2301.06187*, 2023.

[20] A. Alameer, S. Buijs, N. O'Connell, L. Dalton, M. Larsen, L. Pedersen, and I. Kyriazakis, "Automated detection and quantification of contact behaviour in pigs using deep learning," *biosystems engineering*, vol. 224, pp. 118–130, 2022.

[21] K. Luxem, P. Mocellin, F. Fuhrmann, J. Kürsch, S. R. Miller, J. J. Palop, S. Remy, and P. Bauer, "Identifying behavioral structure from deep variational embeddings of animal motion," *Communications Biology*, vol. 5, no. 1, p. 1267, 2022.

[22] E. Bocaj, D. Uzunidis, P. Kasnesis, and C. Z. Patrikakis, "On the benefits of deep convolutional neural networks on animal activity recognition," in *2020 International Conference on Smart Systems and Technologies (SST)*. IEEE, 2020, pp. 83–88.

[23] E. Fazzari, F. Carrara, F. Falchi, C. Stefanini, and D. Romano, "Using ai to decode the behavioral responses of an insect to chemical stimuli: towards machine-animal computational technologies," *International Journal of Machine Learning and Cybernetics*, pp. 1–10, 2023.

[24] S. Yang, O. Nachum, Y. Du, J. Wei, P. Abbeel, and D. Schuurmans, "Foundation models for decision making: Problems, methods, and opportunities," *arXiv preprint arXiv:2303.04129*, 2023.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *The International Conference on Learning Representations (ICLR)*, 2022.

[27] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[29] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 2, no. 3, 2021, p. 4.

[30] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017.

[31] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022.

[32] W. Huang, Y. Shen, and Y. Yang, "Clip-mamba: Clip pretrained mamba models with ood and hessian evaluation," *arXiv preprint arXiv:2404.19394*, 2024.