# DUAL-PHASE MSQNET FOR SPECIES-SPECIFIC ANIMAL ACTIVITY RECOGNITION

*An Yu* *Jeremy Varghese* *Ferhat Demirkiran* *Peter Buonaiuto* *Xin Li* *Ming-Ching Chang*

{ ayu, jvarghese2, fdemirkiran, pmbuonaiuto, xli48, mchang2 }@albany.edu

University at Albany – State University of New York, NY 12222, USA

## ABSTRACT

We present an effective method to detect and recognize multiple animal activities that can advance wildlife conservation and ecological research. While current activity recognition models focus on human actions, we emphasize the need for species-specific designs to accommodate the diverse and complex movements of animals. Our approach involves a dual-phase process: first identifying the species, then recognizing its activities using the cutting-edge Multi-modal Semantic Query Network (MSQNet), a Transformer-based object detector. By customizing and training our models meticulously, we address the challenges posed by the wide range of animal behaviors and physical characteristics. This highlights the importance of dedicated action recognition systems for non-human subjects. Our method achieves an impressive multi-label average precision (MAP) score of 72.5% on the Animal Kingdom dataset, demonstrating precise animal activity recognition capabilities that benefit wildlife conservation and ecological studies. This performance placed our team among the top contributions to the ICME 2024 MMVRAC challenge. Our research paves the way for real-time observation, recognition, and classification of animal behaviors in wildlife studies. Code is available at https://github.com/casperious/DP_MSQNet.

***Index Terms*—** Animal action recognition, animal identification, Animal Kingdom dataset, MMVRAC.

## 1. INTRODUCTION

Animal action recognition from videos plays a pivotal role in the observation and interpretation of animal behavior, serving as a fundamental tool for biological and ecological studies [1]. This capability has tangible benefits in fields such as conservation [2], animal welfare [3, 4], and study of human-animal interactions [5]. Moreover, it significantly contributes to wildlife monitoring and protection initiatives. Through accurate recognition and analysis of animal movements, researchers and conservationists can effectively track endangered species, enhance our understanding of animal behaviors in their natural environments, improve livestock management practices, and improve conditions for animals held in captivity. Additionally, this research opens avenues for exploring evolutionary links in behavior and cognition that span across different species, including humans.

Action recognition has predominantly focused on human subjects, driving extensive research and advances in diverse applications such as healthcare [6], security [7] and sports analytics [8]. The ability to automatically identify and understand human actions from video footage has led to innovative solutions that improve safety measures, enable advanced healthcare monitoring systems, and promote interactive technologies, significantly advancing these domains.

Extending human action recognition technologies to animals adds complexity due to species diversity. Each animal species has unique physical traits and behaviors, necessitating specialized models for accurate identification and behavior recognition. Traditional models designed for humans are inadequate due to these differences. Utilizing animal-specific recognition can improve scientific understanding and enrich educational and research experiences. The AnimalKingdom dataset, which includes 50 hours of annotated footage, supports the recognition of specific animal behaviors.

In this paper, we present a novel animal action recognition pipeline that starts with species identification from video data, followed by action recognition for precise behavioral analysis. Fig. 1 provides an overview. Our approach uses species-specific models to account for the unique movements and behaviors of each species. This dual-phase approach allows for a more refined and precise interpretation of animal actions. We employ a general model for initial species identification and specialized models for detailed action analysis. This strategy not only improves recognition accuracy but also optimizes efficiency. Our methodology was successfully tested in the 2024 ICME Grand Challenge: Multi-Modal Video Reasoning and Analyzing Competition (MMVRAC), where it ranked among the top entries, demonstrating its effectiveness and innovation.

Contribution of this paper includes the following:

- We introduce a dual-phase, species-specific pipeline for effective action recognition capable of identifying animal actions across a wide range of species. Starting with animal identification from video, our approach makes use of the state-of-the-art Multi-modal Semantic Query Network (MSQNet), which is tailored toward a novel actor-agnostic

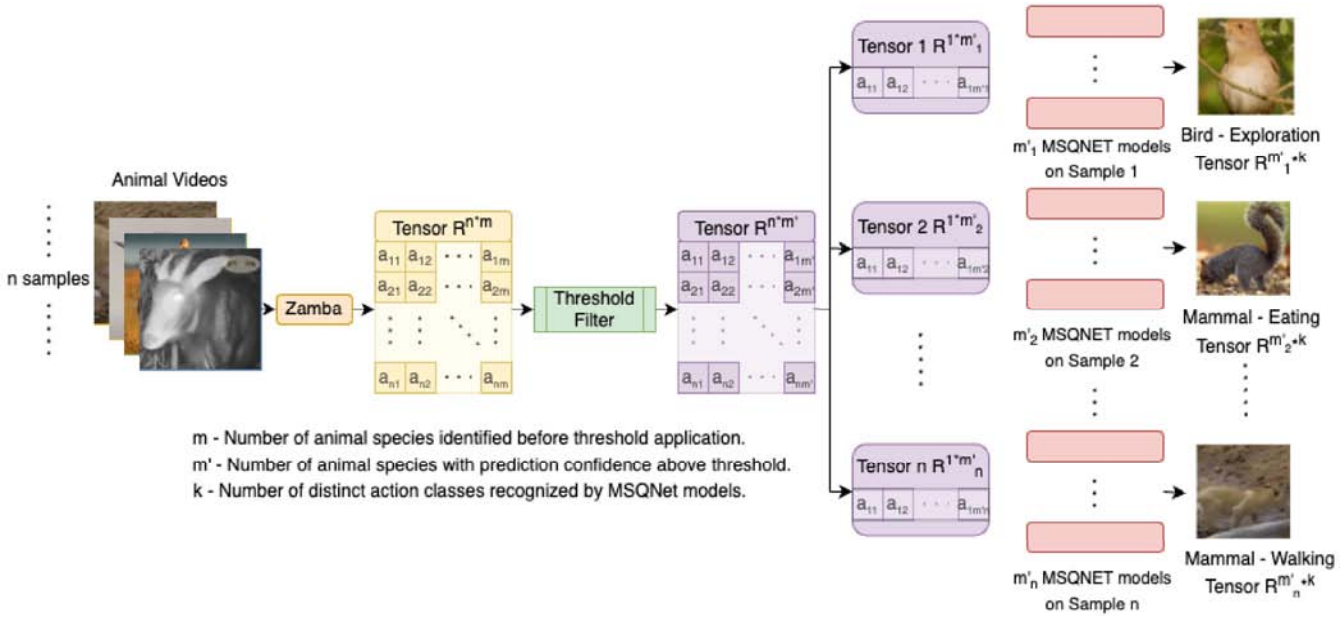


**Fig. 1**. The Adaptive Dual-Phase MSQNet Pipeline: Initiating with Zamba's Sequential Video Analysis for Animal Detection, Proceeding to Species-Specific Action Recognition, and Finalizing with Weighted Action Class Prediction.

recognition pipeline. The final animal action class is determined through robust weighted prediction.

- Our method achieves an MAP score of 72.5% on the Animal Kingdom dataset. Our results is among top teams in the 2024 ICME MMVRAC Grand Challenge.

## 2. RELATED WORK

**Animal identification from images:** The shift from traditional machine learning (ML) to deep learning has transformed animal identification. Initially, ML methods relied on handcrafted features and classifiers like support vector machines (SVMs) [9], but struggled with natural image variability. The emergence of deep learning, particularly Convolutional Neural Networks (CNNs), significantly boosted accuracy over traditional ML models [10, 11]. CNNs excel in distinguishing among species and achieving high accuracy in identification tasks. With the development of CNN architectures and the use of transfer learning, challenges stemming from image variability have been addressed. In [12], Mask R-CNN with a pre-trained ResNet101 backbone is used to identify two animal species. In [13], transfer learning with pre-trained networks is applied to classify 26 animal species on ImageNet. This highlights the benefits of fine-tuning pre-trained models for the ability to identify diverse species.

**Animal identification in video:** The process of identifying animals in video data typically involves two phases: frame extraction and animal classification using a sequence of images. CNNs are widely used to extract features from these frames. To handle spatial-temporal features within sequences of frames, architectures such as Long Short-Term Memory (LSTM) [14], Gated Recurrent Units (GRU) [15], Transformers [16], and Mamba [17] offer significant advantages.

Self-attention is not limited to object detection; it finds utility in dense prediction such as image segmentation as well. Various techniques, such as hierarchical pyramid vision transformer (ViT) [18], progressive upsampling [19], multi-level feature aggregation [20], and masking-based predictions [21] have been explored in this context. Moreover, Transformers have been adopted in diverse domains, including action localization and recognition [22], video classification [23], and group action recognition [24]. Notably, pure Transformer-based models with spatio-temporal attention have been developed to improve performance in these tasks.

**Action recognition:** Accurate encoding of spatial and motion information is crucial for detecting actions in unconfined video footage. Earlier approaches in video comprehension utilized a combination of 2D or 3D convolution alongside sequential models to understand spatial and temporal nuances [25]. Recent advances have introduced vision transformer-based models [26], prioritizing long-range context and showcasing superior performance. These models adeptly capture extensive spatio-temporal relationships, surpassing conventional convolutional counterparts.

**Vision language models:** Large-scale pre-training of image-text representations has proven highly effective across various tasks, including text-to-image retrieval [27], image captioning [28], visual question answering [29], object detection [30], and image segmentation [31]. This success has led to the widespread adoption of *foundation models* like Contrastive Language-Image Pre-Training (CLIP) [32] and A Large-scale ImaGe and Noisy-text embedding (ALIGN) [33]

within the computer vision community. However, extending this knowledge from vision-language models to videos presents challenges due to the limited temporal information available at the image level.

**Actor agnostic models.** While earlier approaches primarily focused on unimodal solutions, recent efforts such as ActionCLIP [34] and XCLIP [35] have embraced a multi-modal strategy by leveraging CLIP [32] for video comprehension. However, existing methodologies often concentrate on specific actors, either humans or animals. Noteably, the Multi-modal Semantic Query Network (MSQNet) [36] is an action recognition model that does not rely on specific actors. It builds a multi-modal query for the Transformer decoder network, which is trained on videos with multiple action labels. By incorporating Transformers into its video encoder, MSQNet captures fine-grained features and their relationships over time and space, allowing for actor-agnostic action classification. It utilizes the Detection with Transformers (DETR) framework for multi-label action recognition. In this paper, we train 7 MSQNet models on the individual species in the Animal Kingdom dataset [37] and demonstrate the capability to recognize multiple animal actions independent of the involved actors. Different animal species may behave differently due to their unique habits, skeletal structures, and other factors. Recognizing these actions is crucial for animal studies.

## 3. METHODOLOGY

Our method starts with the identification of animals in § 3.1, and proceeds to the recognition of specific animal behaviors in § 3.2. Finally, we calculate the weighted prediction in § 3.3.

### 3.1. Animal Identification

Given an input video with unknown animal type, we first conduct animal classification by analyzing frames across the temporal dimension. Specifically, we utilize the `TimeDistributedEfficientNet` model from **Project Zamba** [1] to leverage the robust feature extraction capabilities of EfficientNet [38] within a setting tailored for video classification. This model is tailored to process video frames sequentially, effectively utilizing EfficientNet as its backbone. With a `TimeDistributed` wrapper around EfficientNet, the model efficiently handles sequential data in videos, making it well-suited for analyzing animal videos.

Our animal classification model consists of multiple layers. It starts with a linear layer that reduces the dimensionality of the backbone's output features. Dropout is applied for regularization purposes. Next, a ReLU activation function introduces non-linearity to the features. Another linear layer further processes these features. Finally, a Fully Connected (FC) layer predicts the animal class.

[1] https://zamba.drivendata.org/

### 3.2. Species-Specific Action Recognition

We adopt the state-of-the-art Multi-modal Semantic Query Network (MSQNet) [36], a vision-language model based on the Transformer [16] framework tailored for multi-label, multi-modal action classification. MSQNet integrates three critical elements: (1) a spatio-temporal video encoder that extracts spatial and motion details by segmenting and embedding video frames, (2) a multi-modal query encoder that merges video data with action-specific information through concatenation of label and video embeddings from a pre-trained CLIP [32] encoder, and (3) a multi-modal decoder that refines the video encoding using self-attention and encoder-decoder mechanisms. This streamlined process updates queries with contextual data from videos, aiming to accurately predict action class probabilities by employing categorical Binary Cross-Entropy (BCE) Logit loss for training. Our method streamlines the classification by first identifying the species through the Zamba model, and then passing the video through the relevant species-specific trained MSQNet model.

### 3.3. Weighted Prediction

We calculate the likelihood of each animal specie appearing in a video, leading to an estimation matrix for weighted prediction of the $k$ animal action classes. This matrix is a table where rows represent different videos, and columns correspond to the probabilities of each animal specie, with dimensions $R^{n \times m}$, where $n$ is the number of test videos and $m$ is the number of animal species. To refine our predictions, we use a threshold $\epsilon$ to dismiss probabilities below this value, resulting in a more confident prediction matrix for the presence of species in videos. The refined matrix focuses on species with confidence levels above $\epsilon$ with $m'$ representing the number of species confidently identified post-threshold application.

Algorithm 1 shows the calculation of weighted prediction combining the two estimations from animal identification and species-specific action recognition.

**Algorithm 1** Calculate weighted prediction

**Require:** Prediction result from zamba $R^{n \times m'}$, specific animal action recognition result $R^{m' \times k}$, $m'$ may vary for each video sample.

1: **for** $i = 1$ to $n$ **do**
2: $\quad pred_i \leftarrow R^{1 \times m'_i} \times R^{m'_i \times k}$ ▷ $pred_i$ is $R^{1 \times k}$
3: **end for**
4: $actionPred \leftarrow Concatenate(\text{ all } pred_i)$
5: **return** $actionPred$ ▷ $actionPred$ is $R^{n \times k}$

## 4. EXPERIMENTAL RESULTS

**Dataset:** We use the Animal Kingdom dataset [37] [2] as our primary data source. This extensive dataset encompasses over 50 hours of wildlife footage, showcasing diverse animals across various classes and ecosystems. It consists of 30,000 video sequences representing over 850 species, including Mammals, Reptiles, Amphibians, Birds, Marine Life, and Insects.

### 4.1. Implementation Details

We provide details for training the proposed animal identification and species-specific action recognition models.

**Animal identification:** The model is pre-trained on approximately 250,000 camera trap videos from regions in Central, West, and East Africa, supplemented by approximately 13,000 additional videos from camera traps located in Germany, as detailed in **Project Zamba**. For fine-tuning, we utilize the Animal Kingdom dataset. We employ early stopping to determine the optimal number of training epochs, using validation loss as the key criterion for this decision.

**Species-specific action recognition:** First, we conducted model pre-training on the K400 human action recognition dataset [39] [3]. Subsequently, the model is fine-tuned for each animal species over a span of 100 epochs, following the setting of MSQNet [36]. To streamline the dataset and eliminate redundancies, we generated lightweight CSV files. Further processing of the training data involved creating multiple MSQNet instances, each dedicated to training on a species-specific subset of the data.

We partitioned the input into seven distinct datasets, each corresponding to a different animal species in the Animal Kingdom dataset. Subsequently, specific models were trained and tested on these datasets. The selection of the model is determined by Zamba's prediction. If the likelihood $p_i$ of a video containing a particular animal species exceeds a threshold $\epsilon$, the corresponding MSQNet model is then used, as shown in the pipeline in Fig. 1. The depiction of pipeline results with video inputs is illustrated in Fig 3.

All experiments were conducted on a NVIDIA DGX Linux platform equipped with an A100 GPU, CUDA, and PyTorch 2.1.

### 4.2. Evaluation Metric

Fig. 2 illustrates the evaluation of the proposed dual-phase pipeline. For the evaluation of the animal identification module, we utilized the Zamba package and adopted Precision, Recall, and Accuracy as our evaluation metrics. The dataset was partitioned into training, validation, and testing sets with a 3:1:1 ratio. For evaluating our animal identification model, we employ the F1 score, while for both species-specific action recognition models and overall pipeline performance, we utilize the multi-label average precision (MAP) as our assessment metric.

[2] https://sutdcv.github.io/Animal-Kingdom/

[3] https://github.com/cvdfoundation/kinetics-dataset

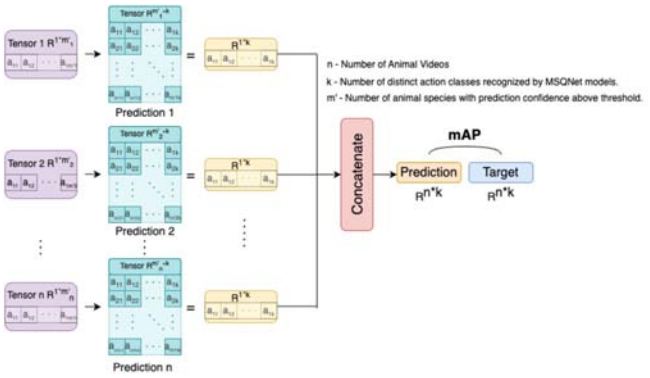


**Fig. 2**. Performance evaluation pipeline for the proposed model using multi-label average precision (MAP).

**Table 1**. Evaluation results on animal identification.

| Evaluation Score | Accuracy | Precision Score | Recall Score | F1 |
|---|---|---|---|---|
| Amphibian | 0.987 | 0.917 | 0.763 | 0.832 |
| Bird | 0.987 | **0.988** | **0.987** | 0.987 |
| Fish | **0.989** | 0.900 | 0.795 | 0.844 |
| Insect | 0.975 | 0.943 | 0.868 | 0.903 |
| Mammal | **0.989** | 0.959 | 0.911 | 0.934 |
| Reptile | 0.983 | 0.976 | 0.909 | 0.941 |
| Sea Animal | 0.987 | 0.900 | 0.613 | 0.729 |
| Overall | - | - | - | **0.89** |

### 4.3. Results

**Evaluation of Animal Identification.** On the test dataset, we achieved a macro F1 score of 0.89. The results of the test evaluation metrics are documented in Table 1. It can be observed that our approach has achieved high accuracy (over 97% across all seven categories. The precision and recall score of the *Bird* category is notably higher than those of others because birds often exhibit distinctive physical features and behaviors that are easily distinguishable from those of other animals, such as plumage color, flight patterns, and nesting activities. These unique attributes can make birds more recognizable for our proposed model.

**Evaluation of individual species action recognition.** In assessing the individual species action recognition module, we leveraged the MSQNet package and use MAP as evaluation metric. The bird species holds the highest MAP score of 82.81, surpassing all other species. Refer to Table 2 for comparison results.

**Evaluation of the whole pipeline.** We again use MAP as the evaluation metric for our pipeline. Since each module is trained independently, no additional training is required. The threshold $\epsilon$ for filtering the Zamba results is set to 0.9.


**Fig. 3**. Our pipeline accurately predicts the animal in the video to be a bird, and subsequently indicates. (a) **exploration** with 63% likelihood, (b) **eating** with 18% probability, and (c) **walking** with 12% chance. These behaviors were all confirmed in the labeled video.

**Table 2**. Evaluation results of species-specific action recognition in the MAP scores.

| - | Amphibian | Bird | Fish | Insect | Mammal | Reptile | Sea Animal |
|---|---|---|---|---|---|---|---|
| MAP | 71.28 | **82.81** | 81.84 | 63.72 | 66.70 | 72.41 | 62.11 |

**Table 3**. Comparing our pipeline against the state-of-the-art. TS: TimeSformer [40]; MMQ: Multi-modal Query [36]

| Method | Backbone | Pretrain | MMQ | MAP |
|---|---|---|---|---|
| CARe [37] | X3D | - | No | 25.25 |
| CARe [37] | I3D | - | No | 16.48 |
| MSQNet [36] | TS | K400 | No | 71.63 |
| MSQNet [36] | TS | K400 | Yes | **73.10** |
| **Ours** | TS | K400 | Yes | **72.50** |

Table 3 shows the comparison of our method against the state-of-the-art (SoTA) methods including the TimeSformer (TS) [40] and the Multi-Model Query (MMQ) [36]. For Animal Kingdom, we consider CARe [37] with two backbones (X3D and I3D) alongside MSQNet for scenarios without specific actors as outlined in MSQNet [36].

## 5. CONCLUSION

We introduce an innovative two-stage pipeline that adeptly integrates species identification and species-specific action recognition, significantly enhancing the accuracy of animal action recognition. This approach has demonstrated a multi-label average precision (MAP) of 72.5% on the Animal Kingdom dataset during the 2024 ICME Grand Challenge. Such progress promises to bolster efforts in wildlife monitoring, conservation, and the promotion of animal welfare, outperforming numerous current methodologies through the introduction of a groundbreaking feature for species classification.

**Future work** includes exploring multimodal learning advancements and integrating explanations of animal actions to improve model training. Furthermore, we intend to explore animal action segmentation, identifying the precise timing and nature of actions within a sequence. This ambitious trajectory will facilitate a deeper understanding of animal behaviors, paving the way for further research and applications in ecological studies and beyond.

**Acknowledgment:** This project is supported by NSF CCSS-2348046 and startup fund of Prof. Xin Li at UAlbany.

## 6. REFERENCES

[1] Timothy M Caro, *Behavioral ecology and conservation biology*, Oxford University Press, 1998.

[2] Alison L Greggor, Daniel T Blumstein, Bob Wong, and Oded Berger-Tal, "Using animal behavior in conservation management: a series of systematic reviews and maps," *Environmental Evidence*, vol. 8, no. 1, pp. 1–3, 2019.

[3] Genaro A Coria-Avila, James G Pfaus, Agustín Orihuela, Adriana Domínguez-Oliva, Nancy José-Pérez, Laura Astrid Hernández, and Daniel Mota-Rojas, "The neurobiology of behavior and its applicability for animal welfare: A review," *Animals*, vol. 12, no. 7, pp. 928, 2022.

[4] Jeremy N Marchant-Forde, "The science of animal behavior and welfare: Challenges, opportunities, and global perspective," *Frontiers in Veterinary Science*, vol. 2, pp. 16, 2015.

[5] James A Griffin, Karyl Hurley, and Sandra McCune, "Human-animal interaction research: Progress and possibilities," *Frontiers in Psychology*, vol. 10, 2019.

[6] Giovanni Diraco, Gabriele Rescio, Andrea Caroppo, Andrea Manni, and Alessandro Leone, "Human action recognition in smart living services and applications: context awareness, data availability, personalization, and privacy," *Sensors*, vol. 23, no. 13, pp. 6040, 2023.

[7] Marco Cristani, Ramachandra Raghavendra, Alessio Del Bue, and Vittorio Murino, "Human behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing*, vol. 100, pp. 86–97, 2013.

[8] Guangyu Zhu, Qingming Huang, Changsheng Xu, Liyuan Xing, Wen Gao, and Hongxun Yao, "Human behavior analysis for highlight ranking in broadcast racket sports video," *TMM*, vol. 9, no. 6, pp. 1167–1182, 2007.

[9] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010, pp. 3360–3367.

[10] Hung Nguyen, Sarah J Maclagan, Tu Dinh Nguyen, Thin Nguyen, Paul Flemons, Kylie Andrews, Euan G Ritchie, and Dinh Phung, "Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring," in *DSAA*. IEEE, 2017, pp. 40–49.

[11] Tibor Trnovszky, Patrik Kamencay, Richard Orjesek, Miroslav Benco, and Peter Sykora, "Animal recognition system based on convolutional neural network," *AEEE*, vol. 15, no. 3, pp. 517–525, 2017.

[12] Savyasachi Gupta, Dhananjai Chand, and Ilaiah Kavati, "Computer vision based animal collision avoidance framework for autonomous vehicles," in *CVIP*. Springer, 2021, pp. 237–248.

[13] Alexander Gomez Villa, Augusto Salazar, and Francisco Vargas, "Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks," *Ecological Informatics*, vol. 41, pp. 24–32, 2017.

[14] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv:1412.3555*, 2014.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[17] Albert Gu and Tri Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv:2312.00752*, 2023.

[18] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *CVPR*, 2022.

[19] Wang Yifan, Shihao Wu, Hui Huang, Daniel Cohen-Or, and Olga Sorkine-Hornung, "Patch-based progressive 3D point set upsampling," in *arXiv 1811.11286*, 2019.

[20] Yanhua Zhang, Ke Zhang, Jingyu Wang, Yulin Wu, and Wuwei Wang, "Multi-level feature aggregation and recursive alignment network for real-time semantic segmentation," in *arXiv 2402.02286*, 2024.

[21] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *arXiv 2112.09133*, 2023.

[22] Chenlin Zhang, Jianxin Wu, and Yin Li, "ActionFormer: Localizing moments of actions with transformers," in *arXiv 2202.07925*, 2022.

[23] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid, "ViViT: A video vision transformer," in *arXiv 2103.15691*, 2021.

[24] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek, "Actor-Transformers for group activity recognition," in *arXiv 2003.12737*, 2020.

[25] Wenbin Du, Yali Wang, and Yu Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE TIP*, vol. 27, no. 3, pp. 1347–1360, 2018.

[26] et al Dosovitskiy, Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale.," in *arXiv*, 2020.

[27] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang, "Image-text retrieval: A survey on recent research and development," 2022.

[28] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *arXiv 1411.4555*, 2015.

[29] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh, "VQA: Visual question answering," 2016.

[30] Ayoub Benali Amjoud and Mustapha Amrouch, "Object detection using deep learning, cnns and vision transformers: A review," *IEEE Access*, vol. 11, pp. 35479–35516, 2023.

[31] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos, "Image segmentation using deep learning: A survey," in *arXiv 2001.05566*, 2020.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[33] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *arXiv 2102.05918*, 2021.

[34] Mengmeng Wang, Jiazheng Xing, and Yong Liu, "Actionclip: A new paradigm for video action recognition," *arXiv:2109.08472*, 2021.

[35] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji, "X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval," in *MM*, New York, NY, USA, 2022, p. 638–647, ACM.

[36] Anindya Mondal, Sauradip Nag, Joaquin M Prada, Xiatian Zhu, and Anjan Dutta, "Actor-agnostic multi-label action recognition with multi-modal query," in *ICCV*, 2023.

[37] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu, "Animal kingdom: A large and diverse dataset for animal behavior understanding," in *CVPR*, 2022.

[38] Mingxing Tan and Quoc Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.

[39] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv:1705.06950*, 2017.

[40] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is space-time attention all you need for video understanding?," *ICML*, vol. 2, no. 3, pp. 4, 2021.