Actor-agnostic Multi-label Action Recognition with Multi-modal Query

Anindya Mondal[123*], Sauradip Nag[125*], Joaquin M Prada[14], Xiatian Zhu[123], Anjan Dutta[1234*]
[1]University of Surrey, [2]CVSSP, [3]Surrey Institute for People-Centred AI,
[4]School of Veterinary Medicine, [5]iFlyTek-Surrey Joint Research Center on AI
{a.mondal, s.nag, j.prada, xiatian.zhu, anjan.dutta}@surrey.ac.uk,

Abstract

Existing action recognition methods are typically actor-specific *due to the intrinsic topological and apparent differences among the actors. This requires actor-specific pose estimation (*e.g.*, humans vs. animals), leading to cumbersome model design complexity and high maintenance costs. Moreover, they often focus on learning the visual modality alone and single-label classification whilst neglecting other available information sources (*e.g.*, class name text) and the concurrent occurrence of multiple actions. To overcome these limitations, we propose a new approach called 'actor-agnostic multi-modal multi-label action recognition,' which offers a unified solution for various types of actors, including humans and animals. We further formulate a novel Multi-modal Semantic Query Network (MSQNet) model in a transformer-based object detection framework (*e.g.*, DETR), characterized by leveraging visual and textual modalities to represent the action classes better. The elimination of actor-specific model designs is a key advantage, as it removes the need for actor pose estimation altogether. Extensive experiments on five publicly available benchmarks show that our MSQNet consistently outperforms the prior arts of actor-specific alternatives on human and animal single- and multi-label action recognition tasks by up to 50%. Code is made available at https://github.com/mondalanindya/MSQNet.*

1. Introduction

Action recognition has been extensively studied, focusing on humans as the actors [30, 29, 16, 32, 53, 14, 25, 2, 6], benefiting a variety of applications, *e.g.*, healthcare [5, 75], virtual and augmented reality [33], and many more [27]. While majority of research has concentrated on humans, there is potential for action recognition to be applied to animals as well [39]. However, recognizing actions and behaviors in animals presents a challenging task. Animals often exhibit different shapes, sizes, and appearances, as illustrated in Figure 1. As a result, it becomes necessary to develop more sophisticated and customized designs tailored to the unique characteristics of each animal actor. One approach to achieve this is by incorporating specific pose information of the actors [39]. Consequently, the ultimate solution becomes tailored to each specific animal actor. Furthermore, most existing methods for action recognition focus on single-label classification, which means they are designed to assign a single action label to a given video. However, in real-world scenarios, multiple actions may occur within a single video, making the task more complex. These methods rely exclusively on video data for model training and inference. As a result, the textual information contained within action class names, often represented as discrete encoded numbers, is often disregarded, even though it could provide valuable context and information.

In order to address the limitations mentioned above, we propose a new problem formulation – *multi-modal multi-label learning for actor-agnostic action recognition*. This novel problem setting aims to utilize multiple sources of information, including visual and textual data, to predict multiple action labels for each video. An essential prerequisite is for the model to remain independent of the actor's identity, thus ensuring its broader applicability and ease of deployment without relying heavily on specific actor traits. This encourages the advancement of sophisticated action recognition models that offer practical benefits in terms of computation and cost-effectiveness. In this direction, we introduce a novel Multi-modal Semantic Query Network (MSQNet), inspired by the Transformer based detection framework (*e.g.*, DETR [35]). By treating multi-label action classification as a specialized form of object detection, MSQNet eliminates the reliance on explicit localization and actor-specific information like actor's pose, thus making it *actor-agnostic*. Importantly, we design a multi-modal semantic query learning scheme to incorporate visual and textual information using a pretrained vision-language model (*e.g.*, CLIP [18]). This approach allows us to combine and utilize both visual and textual data in a trainable manner.

*Authors have equal contributions.

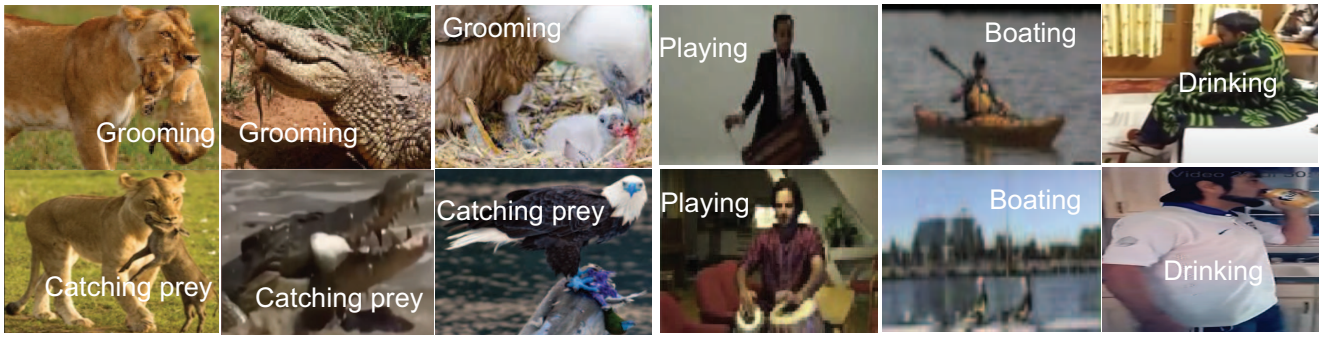

Figure 1. Illustration of large action variation across different actors (*e.g.*, animals and humans). Such differences often motivate the development of actor-specific action recognition models, such as using actor-specific pose estimation [39]

.

This results in a more comprehensive and precise representation of action classes, while simultaneously acquiring the distinct characteristics of actors directly from the training data without requiring actor-specific elements.

Our contributions are summarized as follows: **(1)** We introduce a new problem formulation of *multi-modal multi-label learning for actor-agnostic action recognition* in unconstrained videos. Using a single model architecture for various action tasks minimizes the cumbersome need for actor-specific design, improving model generalization and maintenance efficiency. **(2)** To tackle this problem, we design a novel *Multi-modal Semantic Query Network* (MSQNet) model that casts the multi-label action classification problem into a multi-modal target detection task in the elegant Transformer encoder-decoder framework. It is characterized by a principled vision-language information fusion design for creating richer label queries so that more accurate action classes can be recognized eventually, without requiring actor-specific pose information. **(3)** Through rigorous experimentation on five publicly available benchmarks, we demonstrate that our MSQNet consistently surpasses previous more complex, actor-specific alternatives for human and animal multi-label action recognition tasks.

2. Related Work

Action Recognition: Accurate encoding of spatial and motion information is crucial for recognizing actions in unconstrained videos. Early attempts at video understanding used a combination of 2D or 3D convolution and sequential models in order to capture the spatial and temporal information [10, 17, 61]. Recently, researchers have proposed vision transformer-based models [2, 37, 66], which effectively consider long-range spatio-temporal relationships and have comfortably surpassed their convolutional counterparts. While earlier models mostly consider unimodal solutions, recent works, such as ActionCLIP [57], and XCLIP [40] adopted multi-modal approach by utilizing CLIP and driving it for video understanding. However, all existing works are *actor-specific*, *i.e.*, they consider actions either by humans [2, 45] or by animals [39]. We aim to address this limitation by addressing an *actor-agnostic* action recognition problem, which, to our knowledge, is the first of its kind.

Vision Transformers (ViTs): Inspired by the success of attention-based Transformer [55] models in NLP, Dosovitskiy *et al*. adapted the framework for image classification and named it Vision Transformer (ViT) [13]. With the success of ViT, many others came out with their frameworks focusing on efficient training [51], shifted window-based self-attention [36], deeper architectures [52], self-supervised pretraining [9], etc. Following them, Carion *et al*. [8] considered the CNN feature maps within a classical Transformer encoder-decoder architecture to design an end-to-end object detection framework called DETR [8]. This was improved further by several DETR-like object detectors focusing on training [76, 35] and data efficiency [59]. Self-attention has also been explored in dense prediction tasks like image segmentation, where hierarchical pyramid ViT [58], progressive upsampling, and multi-level feature aggregation [71], masking based predictions [48]. In addition to these works on the image domain, Transformers have been adopted on top of convolutional feature maps for action localization and recognition [20], video classification [60], and group activity recognition [19], which were extended in pure Transformer based models considering spatio-temporal attention [6, 50]. In this paper, we have adapted the Transformers as part of our video encoder to consider fine-grained features and their spatial and temporal relationships to model actor-agnostic action classification tasks. Besides, we take advantage of the DETR framework for multi-label action recognition, which has never been attempted in previous action models.

Vision Language Models: Numerous applications have demonstrated the high effectiveness of large-scale pretraining of image-text representations, including but not limited to, text-to-image retrieval [62], image captioning [65], vi-

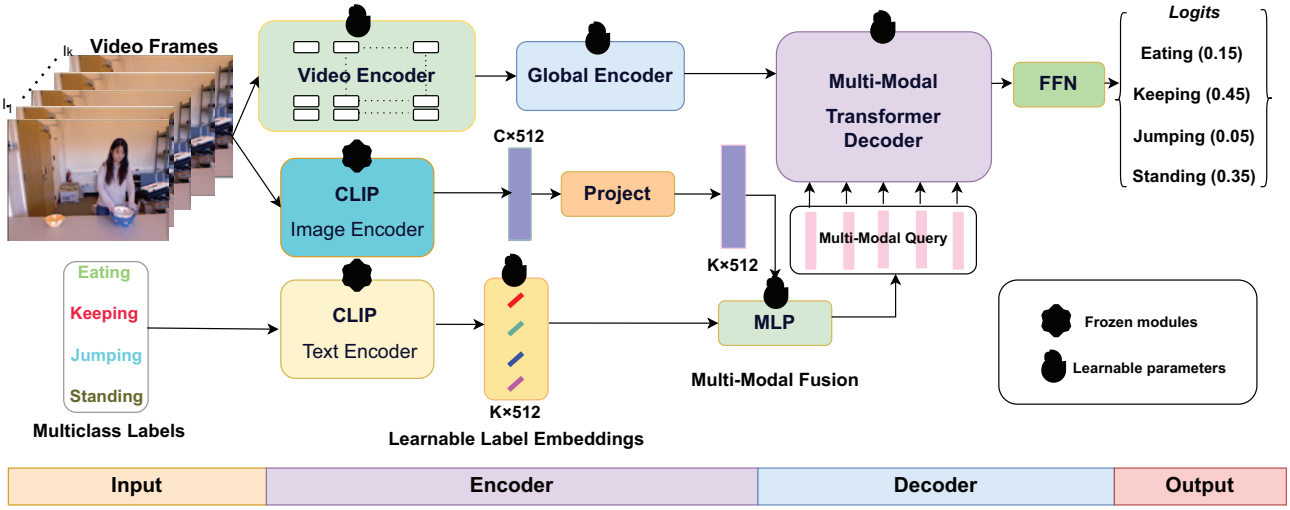

Figure 2. Overview of our MSQNet model for multi-label action recognition using multi-modal query. It has three key components: a spatio-temporal video encoder, a vision-language query encoder and a multi-modal decoder. The video encoder extracts the spatio-temporal features from an input video, the query encoder merges the visual and textual information, and the multi-modal decoder transforms the video encoding to make multi-label classification with a feed-forward network (FFN). Note that two separate visual encoders are employed for distinct purposes. The video encoder extracts spatio-temporal features from videos, while the CLIP image encoder captures complementary features aligned with textual content.

sual question answering [1], few and zero-shot recognition [68, 73], object detection [21, 4, 74], and image segmentation [12, 31, 72]. As a result of their success, foundational vision-language models such as CLIP [44] and ALIGN [23] have become quite popular in the computer vision community. However, when attempting to transfer this knowledge from vision-language models to videos, challenges arise due to the limited availability of temporal information at the image level. To address this problem, recent research such as [57, 40, 42] have attempted to adapt the popular CLIP model to videos by incorporating additional learnable components, including self-attention layers, textual or vision prompts, etc. In contrast to these existing models, our approach involves using pretrained vision-language models to create multi-modal semantic queries that can be plugged into the Transformer decoder network for extracting the key features from the video encoder for actor-agnostic multi-label classification.

3. Multi-modal Semantic Query Network

We introduce the MSQNet, a vision language model (VLM) in the transformer architecture, to perform multi-label multi-modal action classification in an *actor-agnostic* way. As shown in Figure 2, our model consists of three components: (1) a *spatio-temporal video encoder* that takes into account fine-grained spatial and motion cues, (2) a *multi-modal query encoder* that combines information from both video and action class-specific sources, and (3) a *multi-modal decoder* that employs multi-headed self-attention and encoder-decoder attention mechanisms to transform the video encoding. We start with an overview of the video encoder in Section 3.1, followed by a description of multi-modal query encoder in Section 3.2 and a sketch of our multi-modal decoder in Section 3.3. Finally, we outline our learning objective in Section 3.4.

3.1. Spatio-temporal Video Encoder

Consider a video $V \in \mathbb{R}^{T\times 3\times H\times W}$ of spatial dimension $H \times W$ with T sampled frames. Following the existing video Transformer models [6, 2], each frame is divided into N non-overlapping square patches of size $P \times P$, with the total number of patches being $N = HW/P^2$. We flatten these patches into vectors and represented those vectors as $\mathbf{x}_{(p,t)} \in \mathbb{R}^{3P^2}$, where $p = 1, \ldots, N$ denoting spatial locations and $t = 1, \ldots, T$ depicting an index over frames. We then map each patch $\mathbf{x}_{(p,t)}$ into an embedding vector $\mathbf{z}^{(0)}_{(p,t)} \in \mathbb{R}^{D'}$ by a projection layer $W_{\text{emb}} \in \mathbb{R}^{3P^2\times D'}$:

$$\mathbf{z}^{(0)}_{(p,t)} = W_{\text{emb}}\mathbf{x}_{(p,t)} + \mathbf{e}^{pos}_{(p,t)} \tag{1}$$

where $\mathbf{e}^{\text{pos}}_{(p,t)} \in \mathbb{R}^{D'}$ represents a learnable positional embedding to encode the spatio-temporal position of each patch. The resulting sequence of embedded vectors $\mathbf{z}^{(0)}_{(p,t)}$ ($p = 1, \ldots, N$, and $t = 1, \ldots, T$) represents the input of the Transformer encoder [13]. Following most Transformers, we add in the first position of the sequence a special learnable vector $\mathbf{z}^{(0)}_{(0,0)} \in \mathbb{R}^{D'}$ depicting the embedding of the global token. From the video encoder with L_v number

of layers, we thus obtain the patch level representation at each layer l as:

$$\mathbf{z}_{(p,t)}^{(l)} = f_{\theta_v}^{(l)}(\mathbf{z}_{(p,t)}^{(l-1)}), \quad l \in 1, \ldots, L_v \tag{2}$$

where $f_{\theta_v}^{(l)}$ is the l-th layer of the video encoder. Finally, to obtain a global frame level representation, all the patch tokens from each of the frames are averaged and then projected to a dimension D using a linear projection layer (also called *global encoder*) $\mathbf{W}_{\text{out}} \in \mathbb{R}^{D \times D'}$

$$\mathbf{v}_t = \mathbf{W}_{\text{out}} \mathbf{z}_t^{(L_v)}, \tag{3}$$

where $\mathbf{z}_t^{(L_v)} = \text{AvgPool}([\mathbf{z}_{(0,t)}^{(L_v)}, \ldots, \mathbf{z}_{(N,t)}^{(L_v)}])$, $\mathbf{v}_t$ is the output representation of frame t and $\mathbf{z}_{(0,0)}^{(L_v)}$ is the global token from the output sequence of the last layer of the video encoder. The sequence representing the video V comprises the global token $\mathbf{z}_{(0,0)}^{(L_v)}$ and also the frame level representations $[\mathbf{v}_1, \ldots, \mathbf{v}_N]$, and takes the form of $[\mathbf{z}_{(0,0)}^{(L_v)}, \mathbf{v}_1, \ldots, \mathbf{v}_N]$, which we write as $\mathcal{F} = [\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_N]$ by abuse of notations.

3.2. Multi-modal Query Encoder

Given a training video $V \in \mathbb{R}^{T \times 3 \times H \times W}$ with multi-class action labels Y, we construct multi-modal query for our Transformer decoder network. The multi-modal query is formed by fusing the learnable label embedding and video-specific embedding. In our case, the learnable label embedding for a class is a D dimensional learnable vector, depicted as $\mathcal{Q}_l \in \mathbb{R}^{K \times D}$ for a dataset, where K is the total number of classes in that dataset. In our training, we initialize $\mathcal{Q}_l$ with the text embeddings of the corresponding classes. For obtaining those D dimensional text embeddings, a pretrained text encoder is used, *e.g.*, a 12 layer CLIP [44] model (for CLIP B/16 variant) with an embedding size of $D = 512$. For obtaining the video embedding, we employ the CLIP [18] image encoder (CLIP B/16 variant) on the T frames independently as a batch of images and produce frame-level embeddings of dimension D''. These frame level embeddings are average pooled to obtain a video embedding $\mathcal{Q}_v \in \mathbb{R}^{D''}$. Note that we use the CLIP image encoder to extract complementary video features that are aligned with the textual content. This serves a different purpose from the spatio-temporal video encoder described in Section 3.1. In order to form a multi-modal query, we concatenate $\mathcal{Q}_l$ and $\mathcal{Q}_v$ and deploy a linear projection with weights $\mathbf{W}_{\text{que}} \in \mathbb{R}^{D \times (D + D'')}$ resulting the multi-modal query $\mathcal{Q}_0 = \mathbf{W}_{\text{que}}[\mathcal{Q}_l, \mathcal{Q}_v]$, where $[\cdot, \cdot]$ denotes the concatenation operation.

Discussion: Initially, the proposed model design is independent of poses, ensuring that our solution is not limited to specific actors. Subsequently, the amalgamation of textual data and visual embeddings yields a comprehensive representation of actions, enhancing the model's expressive capacity. Lastly, incorporating general textual embedding enables the model to exhibit zero-shot capabilities.

3.3. Multi-modal Decoder

After obtaining the spatio-temporal features $\mathcal{F}$ of input video from the video encoder, we consider the multi-modal semantic queries $\mathcal{Q}_0 \in \mathbb{R}^{K \times D}$ from the multi-modal query encoder. We then perform self- and cross-attention to pool action-specific features from the spatio-temporal video representation using multi-layer Transformer decoders. We use the standard Transformer architecture, with a multi-head self-attention (MultiHeadSA) module, a cross-attention (MultiHeadCA) module, and a position-wise feed-forward network (FFN). Each decoder layer l updates the queries $\mathcal{Q}_{l-1}$ from the outputs of its previous layer as follows:

$$\mathcal{Q}_l^{(1)} = \text{MultiHeadSA}(\tilde{\mathcal{Q}}_{l-1}, \tilde{\mathcal{Q}}_{l-1}, \mathcal{Q}_{l-1}), \tag{4}$$

$$\mathcal{Q}_l^{(2)} = \text{MultiHeadCA}(\tilde{\mathcal{Q}}_l^{(1)}, \tilde{\mathcal{F}}, \mathcal{F}), \tag{5}$$

$$\mathcal{Q}_l = \text{FFN}(\mathcal{Q}_l^{(2)}), \tag{6}$$

where the tilde denotes the original vectors modified by adding position encodings, $\mathcal{Q}_l^{(1)}$ and $\mathcal{Q}_l^{(2)}$ are two intermediate variables. For the sake of simplicity, we exclude the parameters of the MultiHead attention and FFN functions, which are identical to those in the standard Transformer decoder [55]. Each label embedding $\mathcal{Q}_{l-1,k} \in \mathbb{R}^D, k \in \{1, \ldots, K\}$, evaluates the spatio-temporal frame features $\tilde{\mathcal{F}}$ to find where to attend and then combine with the features of interest. This results in a better category-related feature for label embedding. The label embedding is then updated with this new feature. This process is repeated for each layer of the decoder network. As a result, the label embeddings $\mathcal{Q}_k$ are updated layer by layer and progressively injected with contextualized information from the input video via self- and cross-attention. In this way, the embeddings can be learned end-to-end from data and model label correlations implicitly.

Feature projection: To perform single-label classification, we require our model to be confident in the correct action label. For multi-label classification, we consider each predicted label as a binary classification problem. To achieve this, we project the feature representation of each class from the L-th layer of the Transformer decoder $\mathcal{Q}_{L,k} \in \mathbb{R}^D$ onto a linear projection layer. This step is followed by applying an activation function σ, which is implemented as Softmax for single-label tasks and as a Sigmoid function for multi-label action classification tasks:

$$p_k = \sigma(W_k \mathcal{Q}_{L,k} + b_k), \tag{7}$$

where $W_k \in \mathbb{R}^D$, $\mathbf{W} = [W_1, ..., W_K]^T \in \mathbb{R}^{K \times D}$, and $b_k \in \mathbb{R}$, $b = [b_1, \ldots, b_K]^T \in \mathbb{R}^K$ are the linear layer parameters and $p = [p_1, \ldots, p_K]^T \in \mathbb{R}^K$ are the probabilities of each class. We see p as a function mapping an input video to class probabilities.

3.4. Learning Objective

Given a video V, our objective is to train our model in a way so that the predicted probability for each action class $p = [p_1, \ldots, p_K]^T \in \mathbb{R}^K$ matches with the ground truth. Thus we train our model using the categorical cross-entropy loss as the final learning objective.

$$\mathcal{L} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} y_{ij} \log(p_{ij}), \tag{8}$$

where n is the number of samples or observations and K is the number of classes. Specifically, on a single-label dataset, such as Thumos14, we use the "classical" cross-entropy loss useful for multi-class single-label settings, whereas, for other datasets which are multi-label, we use binary cross-entropy loss.

4. Experiments

Datasets: We evaluate both single-label and multi-label action recognition datasets: (1) **Thumos14** [22], a single-label dataset, consists of 13,000 videos from 20 classes, with 1,010 validation and 1500 untrimmed test videos. (2) **Hockey** [47], a multi-label dataset, has 12 activities across 36 videos. (3) **Charades** [46] contains 66,500 annotations for 157 actions and is divided into 7,986 training and 1,863 validation multi-label videos. (4) **Animal Kingdom** [39] is a large multi-label dataset with over 50 hours of footage featuring wild animals from different classes and environments. It consists of 30,000 video sequences, including over 850 species from Mammalia, Aves, Reptilia, Amphibia, Pisces, and Insects. (5) **HMDB51** [28] is a comprehensive compilation of original videos from various sources, including films and online videos. It comprises 6,766 video clips spanning 51 different action categories, such as "jump", "kiss", "laugh" etc. Each class consists of at least 101 clips.
Training details: We train our model for 100 epochs with a cosine decay scheduler and an initial learning rate of 0.00001 using Adam optimizer [26]. Unless otherwise stated, we set the number of frames for training to 16. We use the BCEWithLogitsLoss for Animal Kingdom, Charades Hockey, Volleyball, and the CrossEntropyLoss for Thumos14 and HMDB51.
Evaluation metrics: Following the existing protocol, we use accuracy [41] as the evaluation metric for Hockey, Volleyball, Thumos14 and HMDB51. For Animal Kingdom and Charades datasets, we use the mean average precision (mAP) [49] for performance measurement.

4.1. Supervised learning evaluation

Setting: This is the most conventional setting where a labeled dataset D_S with labels $Y_S = \{Y_i\}_{i=1}^n$ is available for model training.
Results: We present the comparative results in Table 1. For the Charades dataset, we compare our model with AFAC [69], MViT [15], and ActionCLIP [57]. In the case of the Thumos14 dataset, we have considered BMN [34], R-C3D [64], and SSN [70]. For Animal Kingdom, we only consider CARe [39] with two backbones (X3D and I3D), as no other method presented their results on it. For the Hockey dataset, we consider EO-SVM [7] and AFAC [69]. Finally, for HMDB51 in Table 2, we consider BIKE [63], R2+1D-BERT [24], and VideoMAE V2-g [56] for comparison. We have used two settings to evaluate MSQNet, one with learnable queries (*i.e.* only text cues, without the video cues) and the other with multi-modal queries, considering both the learnable queries and video cues. We utilized the backbone of ViT-B, C3D, and TimeSformer trained on the Kinetics-400 [25] and I3D for fair comparison. As seen in Table 1 and Table 2, our multi-modal setting surpasses all the previous approaches. Notably, the improvement is significant for the Animal Kingdom [39] dataset, suggesting that integrating vision-language information is beneficial for handling diverse actions across various species and genera. Furthermore, MSQNet's outstanding performance on the three human action datasets confirms the generalizability of our model, as it is independent of specific actors.

4.2. Zero-shot learning evaluation

Setting: In this setting, the model is trained on a source dataset D_{train} and tested directly on a target dataset D_{test}. The source dataset D_{train} contains samples belonging to source classes $Y_{\text{train}} = \{y_i\}_{i=0}^k$. The model is evaluated on the target dataset D_{test} with classes Y_{test}, such that $Y_{\text{train}} \bigcap Y_{\text{test}} = \phi$, *i.e.*, the action categories for training and testing remain disjoint. For this experiment, we have considered three datasets: (1) Thumos14 [22], (2) Charades [46], and (3) HMDB51 [28]. For Thumos14, we consider the dataset splits proposed by [38]. For HMDB51, we consider the zero-shot splits by [43]. In the case of Charades [46], since no such split is publicly available, we have defined our own random splits for conducting the zero-shot experiments, which we will make publicly available upon the acceptance of this paper. Specifically, for this experimental setting, we consider two different dataset splits for Thumos14, Charades, and HMDB51: (1) Y_{train} and Y_{test} respectively contain 75% and 25% of the total number of classes, (2) Y_{train} and Y_{test} respectively contain 50% and 50% of the total number of classes in the dataset. To ensure statistical significance, we have followed [38] and considered 10 different random splits of action classes with the aforementioned settings.

Charades [46]					**Thumos 14** [22]				
Method	Backbone	Pretrain	MMQ	*mAP*	Method	Backbone	Pretrain	MMQ	*Accuracy*
AFAC [69]	Nonlocal-101	-	No	44.20	SSN [70]	C3D	-	No	45.42
MViT [15]	SlowFast	K600	No	43.90	R-C3D [64]	C3D	-	No	57.19
ActionCLIP [57]	ViT-B	-	No	44.30	BMN [34]	C3D	S1M	No	62.12
MSQNet	ViT-B	K400	No	43.99	MSQNet	C3D	K400	No	67.47
MSQNet	TS	K400	No	44.11	**MSQNet**	TS	K400	Yes	**79.71**
MSQNet	TS	K400	Yes	**47.57**	**MSQNet**	TS	K400	Yes	**83.16**
Animal Kingdom [39]					**Hockey** [47]				
Method	Backbone	Pretrain	MMQ	*mAP*	Method	Backbone	Pretrain	MMQ	*Multilabel Accuracy*
CARe [39]	X3D	-	No	25.25	EO-SVM [7]	-	-	No	90.00
CARe [39]	I3D	-	No	16.48	AFAC [69]	CSN-152 [53]	-	No	96.30
MSQNet	I3D	-	No	**55.59**	MSQNet	I3D	-	No	93.29
MSQNet	TS	K400	No	**71.63**	**MSQNet**	TS	K400	No	93.45
MSQNet	TS	K400	Yes	**73.10**	**MSQNet**	TS	K400	Yes	**96.95**

Table 1. Comparing our MSQNet against the state-of-the-art in the supervised learning setting. The best results are in bold. `K400`: Kinetics-400; `K600`: Kinetics-600; `S1M`: Sports-1M; `TS`: TimeSformer; `MMQ`: Multi-modal Query.

Method	Backbone	Pretrain	MM	Accuracy
BIKE [63]	ViT	WIT-400M	Yes	84.31
R2+1D-BERT [24]	R(2+1)D	IG65M	No	85.10
VideoMAE V2-g [56]	ViT	K400/K600	No	88.10
MSQNet	TS	K400	Yes	**93.25**

Table 2. Supervised learning performance on HMDB51. The best results are highlighted in bold.

Split	Method	**Thumos 14**	**Charades**	**HMDB51**
Reported	VideoCOCA [67]	-	21.1	58.70
	CLIP-Hitchhiker [3]	-	25.8	-
	BIKE	-	-	61.40
50% Seen	Vanilla MSQNet	49.37	15.87	45.66
	Vanilla MSQNet + Text Init	53.76	18.30	51.22
	MSQNet	**63.98**	**30.91**	**59.24**
75% Seen	Vanilla MSQNet	52.02	17.43	48.37
	Vanilla MSQNet + Text Init	60.28	18.62	59.58
	MSQNet	**75.33**	**35.59**	**69.43**

Table 3. Comparing MSQNet against state-of-the-art in zero-shot setting. We report multilabel accuracy scores for Thumos 14, HMDB51, and mAP for the Charades dataset. The best results are in bold.

Results: We present the experimental results for zero-shot setting in Table 3. Unfortunately, we were unable to run VideoCOCA [67], CLIP-Hitchiker [3] and, BIKE [63] on our splits for the unavailability of their open-source codes. Therefore, we followed the settings mentioned in Section 4.2 to ensure a fair evaluation, and we still report their scores for all the datasets. While comparing, we consider different components of our MSQNet model in Table 3: (1) **Vanilla MSQNet:** MSQNet without text initialization and video embeddings; (2) **Vanilla MSQNet + Text Init.:** MSQNet with text initialisation, but without video embeddings; (3) **MSQNet:** our full model with text initialisation and video embeddings. We further compare MSQNet against their respective SoTA for all three datasets. The results in Table 3 demonstrate that our MSQNet model outperforms the baselines by a significant margin, highlighting the effectiveness of the different model components. Furthermore, the results emphasize the importance of video embeddings for achieving zero-shot capabilities in the two datasets.

4.3. Further analysis

Video encoder: In Table 4, we demonstrate the impact of various backbones on the performance of MSQNet. Our analysis focuses on state-of-the-art backbone architectures such as VideoMAE [50] and TimeSformer [6], which we use to initialize our video encoder. We leverage the Animal Kingdom and Charades datasets to assess the efficacy of these backbones. Our results reveal that the TimeSformer [6] backbone outperforms the VideoMAE [50] backbone on both datasets. The success of TimeSformer can be attributed to its "divided attention", which enables the network to attend separately to spatial and temporal features within each block, leading to enhanced video classification accuracy.

Backbone	**Animal Kingdom** (mAP)	**Charades** (mAP)
VideoMAE [50]	71.19	41.69
TimeSformer [6]	**73.10**	**47.57**

Table 4. Performance of MSQNet using different weights for the video encoder. The best scores are in bold.

# Frames	**AK** (mAP)	**Charades** (mAP)
8	67.74	43.06
10	**69.73**	**45.33**
16	**73.10**	**47.57**

Table 5. Performance of MSQNet with a different number of frames. The best and the second-best scores are in red and blue. AK: Animal Kingdom

Number of frames: In our MSQNet model, we have con-

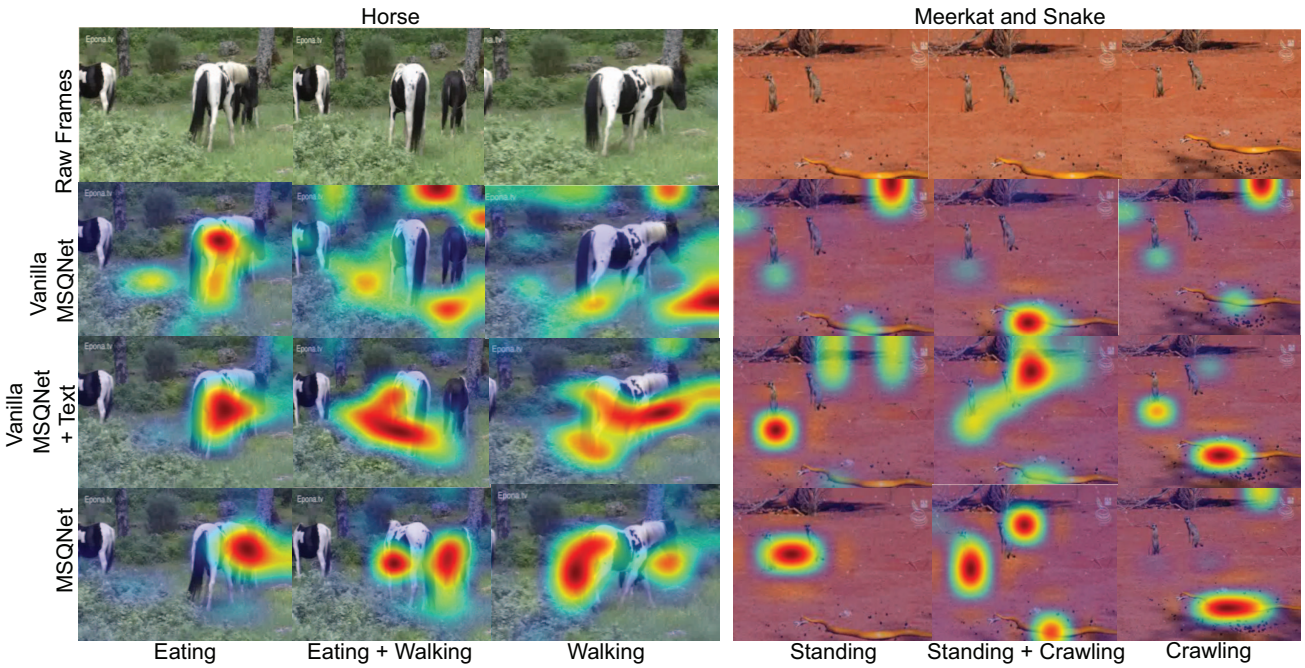

Figure 3. Attention rollout on sample videos from Animal Kingdom [39] showing raw frames, heatmap with only bare backbone, with uni-modal prompt, and MSQNet.

sidered videos with 16 (*i.e.*, $T = 16$) frames as a default setting. However, we have also experimented with clips of lengths 8 and 10 frames. As we can see from Table 5, using 16 frames gives us the best performance. This finding is attributed to the fact that sampling more frames provides a more extensive comprehension of intricate actions and events that can occur over an extended duration. Our finding is consistent over both human and animal action datasets, respectively.

Importance of multi-modal queries: The results presented in the final two rows for each dataset in Table 1 exemplify the advantage of integrating learnable queries (*i.e.* textual features) and visual features to augment the performance of our model. Through visual and textual cues, our model can achieve a more comprehensive understanding of the context surrounding a particular scene or action. This is particularly beneficial in animal datasets, where the CLIP image encoder can be employed to its full potential. This is because the CLIP has undergone pretraining on millions of image-text pairs [44], thereby gaining outstanding zero-shot capability. Consequently, our model excels in diverse scenarios, accommodating a wide range of actors.

Effect of Text Encoding: We evaluate the effect of text encoding in the supervised setting of our MSQNet model. We have tested two common text encoders: BERT [11] and CLIP [44]. In Table 6, we observe that the multi-modality (vision and language) learning based on CLIP is superior to the pure language model BERT. This is not surprising that the former boasts a more robust feature embedding capability, given its extensive pretraining with millions of image-text pairs. Further, it is noteworthy that their difference is not substantial, suggesting the robustness of our method in the text encoding component.

Text Encoder	**AK** (mAP)	**Charades** (mAP)
BERT [11]	69.81	43.57
CLIP [44]	**71.63**	**44.11**

Table 6. Effect of text encoding with our MSQNet.

4.4. Visualization

Qualitative analysis: We employed a feature attention visualization to examine the behavior of MSQNet, as shown in Figure 3. In row 2, we observed that the Vanilla MSQNet (the one with the video encoder alone concentrates on the background to classify the "Eating" and "Walking" actions for the "Horse" diagram. However, when we introduced textual information to the Vanilla MSQNet (row 3), the attention shifted entirely to the horse's body. We enhanced the model's performance by integrating the CLIP image encoder into the uni-modal MSQNet. The attention heatmap demonstrates that introducing a multi-modal prompt to MSQNet allows it to accurately classify "Eating" by emphasizing key features such as the horse, mouth, and grass. Meanwhile, when predicting "Walking" (row 4), the attention focuses on the horses' legs. Similarly, for

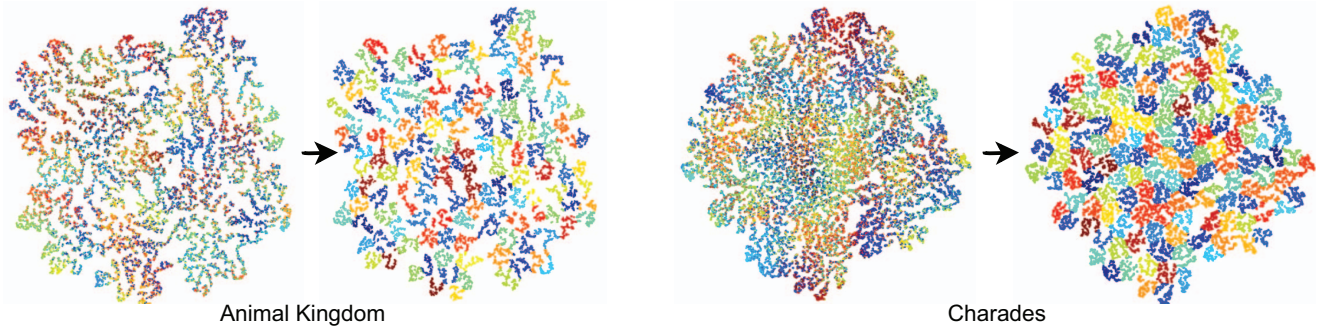

Figure 4. Video embeddings without and with the proposed multi-modal query learning on Animal Kingdom and Charades. Arrow shows the transition.

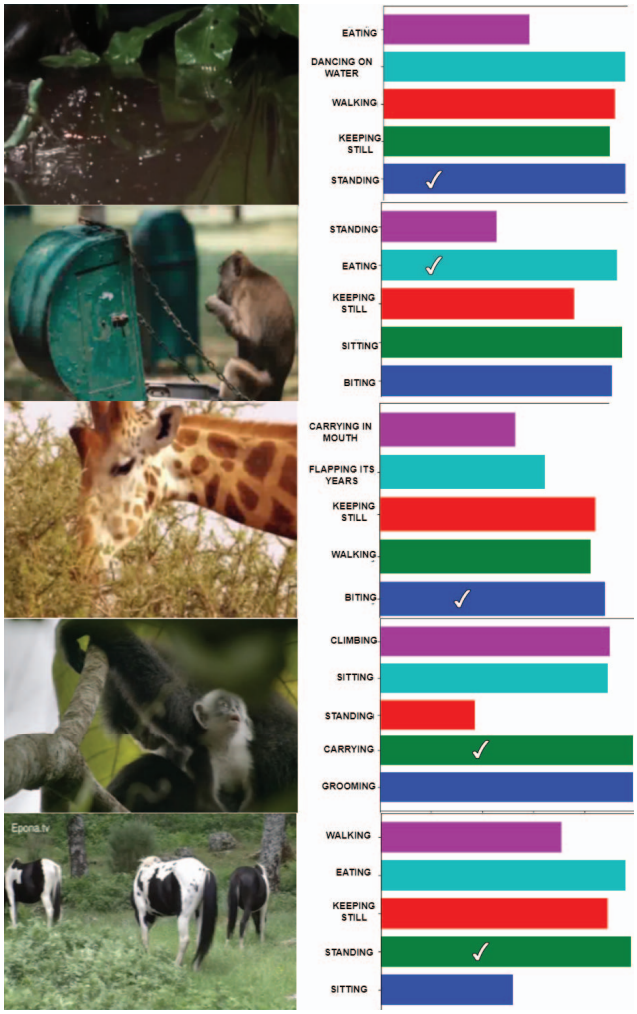

Figure 5. Confidence scores of top-5 classes predicted by our MSQNet. Correctly classified action classes are marked with a ✓.

the "Meerkat and Snake" diagram, MSQNet (row 4) correctly attends to the intended frame regions while identifying the actions "Standing", "Standing and Crawling", and "Crawling" more effectively than their uni-modal and bare video backbone counterparts. This confirms the effectiveness of our approach of leveraging pre-trained vision-language knowledge for accurate action classification.

Further, inspired by [17], we visualize the confidence scores for the top-5 classes predicted by MSQNet, as shown in Figure 5 and https://i.imgur.com/GPoqH8C.gif. Notably, all the top-class predictions exhibit a strong correlation, indicating the robust generalization capabilities of our model.

Feature visualization : The plots in Figure 4 show the t-SNE [54] diagram of the video representations from the Animal Kingdom and Charades datasets. The visualizations demonstrate that the embeddings of the action classes become more distinguishable and meaningful after passing through the multi-modal transformer decoder. These findings suggest that the MSQNet is capable of accurately classifying actions, regardless of the characteristics of the datasets being used.

5. Conclusion

Our proposed MSQNet model utilizes visual and textual information from a pretrained vision-language model to accurately define action classes, eliminating the need for actor-specific design. We achieve improved model design and maintenance efficiency by framing the problem as a multi-modal target detection task within the Transformer decoder. Extensive experiments conducted on multiple benchmarks demonstrate the superiority of our approach over previous actor-specific alternatives for multi-label action recognition tasks involving both humans and animals as actors, including both fully-supervised and zero-shot scenarios.

In the future, we have plans to explore advancements in multi-modal learning techniques further and explore the integration of additional modalities, such as audio, to enhance our model's capabilities. Furthermore, we aim to extend our model to address action detection tasks, allowing for a more comprehensive video action understanding.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 3

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. VIVIT: A video vision transformer. In *ICCV*, 2021. 1, 2, 3

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker's guide to long video retrieval. *arXiv preprint arXiv:2205.08508*, 2022. 6

[4] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *NeurIPS*, 2022. 3

[5] Serge Belongie, Kristin Branson, Piotr Dollár, and Vincent Rabaud. Monitoring animal behavior in the smart vivarium. In *Measuring Behavior*. Wageningen The Netherlands, 2005. 1

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 1, 2, 3, 6

[7] Marc-André Carbonneau. *Multiple instance learning under real-world conditions*. PhD thesis, École de Technologie Supérieure, 2017. 5, 6

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[9] 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. 2

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2018. 7

[12] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 3

[13] 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. In *ICLR*, 2021. 2, 3

[14] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 1

[15] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 5, 6

[16] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, 2019. 1

[17] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 2017. 2, 8

[18] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-Adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 1, 4

[19] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek. Actor-transformers for group activity recognition. In *CVPR*, 2020. 2

[20] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, 2019. 2

[21] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 3

[22] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *CVIU*, 2017. 5, 6

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 3

[24] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *ECCV*, 2020. 5, 6

[25] 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 preprint arXiv:1705.06950*, 2017. 1, 5

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[27] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *IJCV*, 2022. 1

[28] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 5

[29] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, 2017. 1

[30] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *ECCV*, 2016. 1

[31] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 3

[32] Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE TPAMI*, 2020. 1

[33] Xuan Li, Hengxin Chen, Shengdong He, Xinrun Chen, Shuang Dong, Ping Yan, and Bin Fang. Action recognition based on multimode fusion for vr online platform. *VR*, 2023. 1

[34] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, 2019. 5, 6

[35] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *ICLR*, 2022. 1, 2

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[37] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 2

[38] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. In *ECCV*, 2022. 5

[39] 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. 1, 2, 5, 6, 7

[40] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022. 2, 3

[41] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. TorchMetrics - Measuring Reproducibility in PyTorch, Feb. 2022. 5

[42] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. In *NeurIPS*, 2022. 3

[43] Rui Qian, Yeqing Li, Zheng Xu, Ming-Hsuan Yang, Serge Belongie, and Yin Cui. Multimodal open-vocabulary video classification via pre-trained vision and language models. *arXiv preprint arXiv:2207.07646*, 2022. 5

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 4, 7

[45] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, 2023. 2

[46] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 5, 6

[47] Konstantin Sozykin, Stanislav Protasov, Adil Khan, Rasheed Hussain, and Jooyoung Lee. Multi-label class-imbalanced action recognition in hockey videos via 3d convolutional neural networks. In *SNPD*, 2018. 5, 6

[48] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 2

[49] Wanhua Su, Yan Yuan, and Mu Zhu. A relationship between the average precision and the area under the roc curve. In *ICTIR*, 2015. 5

[50] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 2, 6

[51] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2

[52] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021. 2

[53] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, 2019. 1, 6

[54] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2, 4

[56] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023. 5, 6

[57] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2, 3, 5, 6

[58] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 2

[59] Wen Wang, Jing Zhang, Yang Cao, Yongliang Shen, and Dacheng Tao. Towards data-efficient detection transformers. In *ECCV*, 2022. 2

[60] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2

[61] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S Yu. Spatiotemporal pyramid network for video action recognition. In *CVPR*, 2017. 2

[62] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *ICCV*, 2019. 2

[63] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *CVPR*, 2023. 5, 6

[64] Huijuan Xu, Abir Das, and Kate Saenko. Two-stream region convolutional 3d network for temporal activity detection. *IEEE TPAMI*, 2019. 5, 6

[65] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[66] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. 2

[67] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. VideoCoCa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022. 6

[68] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *ECCV*, 2022. 3

[69] Yanyi Zhang, Xinyu Li, and Ivan Marsic. Multi-label activity recognition using activity-specific features and activity correlations. In *CVPR*, 2021. 5, 6

[70] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017. 5, 6

[71] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 2

[72] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 3

[73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 3

[74] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 3

[75] Xiaokang Zhou, Wei Liang, I Kevin, Kai Wang, Hao Wang, Laurence T Yang, and Qun Jin. Deep-learning-enhanced human activity recognition for internet of healthcare things. *IEEE IoT*, 2020. 1

[76] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2