Zero-Shot Action Recognition with Transformer-based Video Semantic Embedding

Keval Doshi and Yasin Yilmaz
University of South Florida
4202 E Fowler Ave, Tampa, FL 33620
{kevaldoshi, yasiny}@usf.edu

Abstract

While video action recognition has been an active area of research for several years, zero-shot action recognition has only recently started gaining traction. In this work, we propose a novel end-to-end trained transformer model which is capable of capturing long range spatiotemporal dependencies efficiently, contrary to existing approaches which use 3D-CNNs. Moreover, to address a common ambiguity in the existing works about classes that can be considered as previously unseen, we propose a new experimentation setup that satisfies the zero-shot learning premise for action recognition by avoiding overlap between the training and testing classes. The proposed approach significantly outperforms the state of the arts in zero-shot action recognition in terms of the the top-1 accuracy on UCF-101, HMDB-51 and ActivityNet datasets.

1. Introduction

Several visual recognition tasks, such as image classification and video action recognition, have made tremendous progress in recent years, thanks to the availability of extensively annotated datasets and enhanced deep learning architectures. However, collecting and annotating video samples for every possible interaction between objects is impractical, therefore recognizing previously unseen actions remains a challenging task. On the other hand, humans are exceptionally good at recognizing new categories without seeing any visual samples. For example, if a person is familiar with *ice skating* and understands the concept of *dancing*, (s)he will have no trouble recognizing the action of *ice dancing*. In the recent literature, this problem has drawn considerable attention and is known as Zero-Shot Learning (ZSL) for video action recognition. While several approaches have shown promising results in the image domain, zero-shot video action recognition remains largely unexplored.

In the existing research, ZSL is characterized as a classification problem in which a model is trained on a collection of known classes and then uses semantic attributes to identify unknown classes. Most existing approaches employ a 3D Convolutional Neural Network (3D-CNN) to extract visual features from videos. Because the utilized 3D-CNNs are pretrained on a variety of large-scale datasets, there is no obvious demarcation between what defines seen and unseen classes. Several recent techniques, for example, are pretrained on the Kinetics-400/600/700 [5, 25, 37] datasets and evaluated on the UCF-101 [38], HMDB-51 [28], Olympics [39] and ActivityNet [11] datasets. However, as shown in Fig. 1, several classes that are considered as "unseen" are already present in the Kinetics dataset, which clearly violates the zero-shot paradigm. On the other hand, even a human is incapable of recognizing *archery* as an activity if he has never seen a *bow* and *arrow* before and is oblivious to the concept of *shooting*. Thus, it can be argued that it is redundant to include activities in the test set that are significantly dissimilar to the activities present in the training set. Furthermore, the existing evaluation setup comprises of randomly splitting the test dataset's classes into half and evaluating the proposed approach on that set. However, this causes an unfair comparison since modifying the split can significantly alter the result. To this end, we propose a new ZSL experimentation framework for action recognition, that addresses these issues and thus would provide guidance for future algorithm design.

Since the majority of existing works employ 2D or 3D convolutions as the principal operator for learning spatiotemporal features from videos, they suffer from several challenges [2]. First, inductive biases such as local connectivity, translation invariance, and a locally restricted receptive field severely limit the learning capability of convolutional models on large datasets. Moreover, convolutional kernels are incapable of capturing spatiotemporal relations that span multiple time instances [2]. Finally, even with advances in hardware acceleration, training and evaluating deep CNNs on large video datasets remain computationally

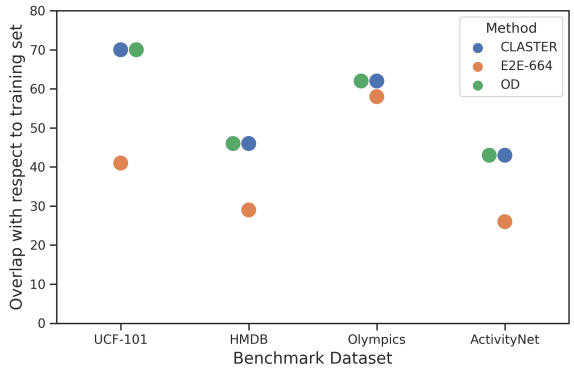

Figure 1. In the recent state-of-the-art approaches CLASTER [19], E2E-664 [4], OD [29], there is a significant overlap between the training classes and the testing classes. The vertical axis shows the percentage of test classes in the benchmark datasets (UCF-101, HMDB, Olympics, ActivityNet) that overlap with the training classes in the Kinetics dataset, which is significantly larger than the benchmark datasets used for testing. Since the test classes are supposed to be previously unseen, such high-percentage overlap (30%-70%) violates the ZSL premise.

expensive.

Motivated by these observations, we propose leveraging self-attention architectures, in particular a spatiotemporal transformer model, for extracting semantic embeddings from videos. The self-attention approach, in contrast to convolutional kernels, can capture long-range dependencies and is permutation invariant while being computationally efficient during training and inference. Our end-to-end trained transformer based approach is able to learn semantically more separable features as compared to 3D-CNNs. Our contributions can be summarized as follows:

- *A novel end-to-end transformer to learn visual-semantic representations.* To the best of our knowledge, this is the first work to propose a *spatiotemporal transformer* for *zero-shot video action recognition* although there are existing transformer methods for zero-shot image tasks.

- *A thorough analysis of the shortcomings in the existing ZSL framework for action recognition.* We propose a new evaluation framework that satisfies the ZSL criterion for action recognition.

- *An extensive evaluation of the proposed transformer method on several benchmark datasets.* The proposed method outperforms several state-of-the-art approaches by a wide margin under different experimental setups.

2. Related Works

Video action recognition has been extensively studied over the past several years [6, 12–14, 36, 45]. In contrast, ZSL for video action recognition has only recently started gaining attention. Broadly, ZSL can be classified into the inductive setting [3, 21, 33, 50, 52], where the test data is completely unknown during training, and the transductive setting [1, 31, 43, 44, 47–49], where the test data without class labels is available. In this work, we only focus on the inductive setting.

Existing approaches have been predominantly dependent on word embeddings to tackle the problem of ZSL. Specifically, these approaches use a pretrained model to extract visual features from training videos and map them to a semantic space and hypothesize that a good robust generalization on the semantic space can lead to improved performance on unseen classes [3, 4, 15–17, 19, 21, 47–50]. For extracting the visual features, most recent approaches propose using a 3D-CNN, which takes 16 frames sampled from a video as input. In [4], Brattoli et al. propose training a C3D [40] and a R(2+1)D model [41] in an end-to-end fashion for ZSL. On the other hand, Gowda et al. [19] propose a reinforcement learning based clustering approach, which uses a two-stream I3D [6] model for learning visual features. However, we argue that for realistic applications, sampling 16 frames from a video might not always be sufficient, especially when it comes to large-scale datasets. In [6, 22], it is shown that increasing the number of input frames only marginally increases the performance of convolutional models. On the contrary, recent transformer based approaches [2, 51] have shown considerable performance gains when number of input frames are increased from 8 to 96, especially on tasks that require longer temporal reasoning. While earlier approaches use hand-crafted semantic features [24], recent works have primarily use Word2Vec [30] for generating the semantic embeddings from the class labels. However, such approaches are prone to the domain shift problem, which occurs when a model trained on the seen semantic labels is unable to generalize well to the unseen class labels [21].

Recently, Brattoli et al. [4] propose an extension on the work of Roitberg et al. [35] and formulate a novel evaluation protocol for satisfying the ZSL paradigm, which involves removing certain classes from the training set which overlap with the test set by using semantic embedding matching. However, in this paper, we show that such an approach fails to remove a significant portion of the overlapping classes, thus still violates the ZSL premise. Alternatively, Gowda et al. [20] propose a deterministic "TruZe" split for the UCF-101 [38] and HMDB-51 [28] datasets, by manually removing all classes which overlap with the Kinetics-400 dataset. While it is a promising approach, we show that it quickly becomes obsolete since several recent approaches use Kinetics-600/700 for training, which includes a major-

ity of the classes from the "TruZe" test set. Furthermore, neither of the proposed approaches remove classes which are significantly different from any of the classes in the training set. Hence, we propose a novel test set (Section 3.3) which addresses all of the above mentioned issues.

3. Proposed Approach

In this section, we present the proposed approach for zero-shot video action recognition. We begin by carefully defining the problem formulation. Next, we introduce a novel end-to-end trained transformer model, *ViSET (Video Semantic Embedding via transformer)*, which leverages divided space-time attention to learn improved semantically differentiable visual features (Fig. 2). Then, we present a new data split for more suitable ZSL experimentation. Finally, we propose to enhance the available semantic embeddings using class descriptions instead of labels.

3.1. Problem Definition

Traditionally, zero-shot action recognition has been defined as a classification problem, where given a training set of videos X^s and labels S from seen classes $\{(x_1^s, s_1), \ldots, (x_N^s, s_N)\}$, we aim to accurately classify a set of videos $X^u = \{x_1^u, \ldots, x_M^u\}$ from previously unseen classes $U = \{u_1, \ldots, u_M\}$, where N and M are the number of training and testing videos respectively. To satisfy the ZSL premise, there should be no overlap between the seen and unseen classes ($S \cap U = \varnothing$).

A broad generalization capability is needed to succeed in the ZSL action recognition problem. The challenge compared to the regular action recognition task lies in the fact that no direct mapping from the input videos to the output unseen class labels can be learned during training. Typically, semantic embeddings are used to bridge the input videos and the output unseen class labels, which consist of words. The idea behind this mainstream ZSL approach is to learn a semantic embedding model $f(x)$ for the input videos and choose the class that is semantically most similar.

3.2. Video Semantic Embedding via transformer (ViSET)

We consider the inductive ZSL approach for training our model, i.e., during training the model only has access to the videos and corresponding class labels from the seen classes. Most existing approaches extract visual features using pretrained 3D-CNNs. Recently, it was shown in [4] that end-to-end training a 3D-CNN performs significantly better than directly using a pretrained 3D-CNN model. However, the inability of 3D-CNNs to capture spatiotemporal information over a long time span make them unsuitable for large scale datasets such as Kinetics and ActivityNet. Hence, we propose a novel end-to-end trained transformer model, called *ViSET (Video Semantic Embedding via transformer)*. The overall structure of ViSET is shown in Fig. 2, which is based on the recently proposed TimeSformer [2] architecture. Specifically, we leverage the self-attention blocks from the Vision transformer (ViT) [10] model and space-time attention from the TimeSformer model. However, contrary to existing video transformers, we train our model to learn visual-semantic representations.

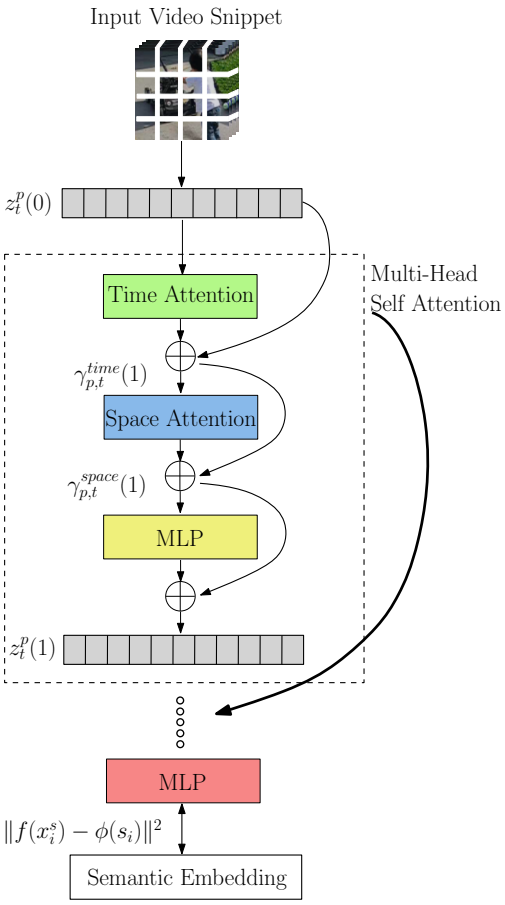

Figure 2. The proposed spatiotemporal transformer architecture, ViSET, for video semantic embedding.

Input: Even with a computationally efficient architecture, end-to-end training on the entire video is impractical due to GPU limitations. Following the existing video classification approaches, we first sample a clip y of F frames and size $H \times W$ from the input video x. The standard NLP transformer architecture [42] requires a 1D sequence of token embeddings as an input. Hence, as shown in Fig. 2, we first breakdown the entire video clip y into a sequence of 2D patches, $e_t^p \in \mathbb{R}^{3 \times P^2}$, where $p = 1, \ldots, N$ represents the spatial locations (i.e., patch index), $t = 1, \ldots, F$ denotes the temporal index, 3 is the number of color channels, and $P \times P$ is the patch size. Next, we flatten each patch e_t^p into $v_t^p \in \mathbb{R}^{3P^2}$ and linearly map it into a score vector using a

trainable linear projection $E \in \mathbb{R}^{q \times 3P^2}$:

$$z_t^p(0) = Ev_t^p + \mu_t^p, \tag{1}$$

where $\mu_t^p \in \mathbb{R}^q$ is a latent vector learned to encode the spatiotemporal position of each (p, t) pair. Following the NLP transformer BERT [9] we add a latent vector $z_0^0(0) \in \mathbb{R}^q$ for an additional fictional patch which will be learned to represent the score vector for the entire video by interacting with all patches through time and space self-attention. $\{z_0^0(0), z_t^p(0)\}_{p,t}$ is the input to the transformer model.

Overall Structure: The transformer model consists of L sequential encoding blocks, each of which includes A parallel self-attention heads. The first encoding block processes the input $\{z_0^0(0), z_t^p(0)\}$ in parallel in its self-attention heads, as explained next, and outputs $\{z_0^0(1), z_t^p(1)\}$ to the second encoding block for each patch (p, t). Similarly, each block l gets $\{z_0^0(l-1), z_t^p(l-1)\}$ and outputs $\{z_0^0(l), z_t^p(l)\}$. Finally, the output of the last block $z_0^0(L)$ is used to obtain the semantic embedding for the video. The role of $z_0^0(L)$ is to learn an effective summary score for the entire video, as an alternative to the straightforward approach of simply averaging $z_t^p(L)$ over all patches and frames. $z_0^0(L)$ is passed through an MLP with three hidden layers and ReLU activation function to obtain the semantic embedding $f(x)$ for video x.

Training and Inference: The whole transformer model is trained end-to-end by minimizing the loss function

$$C = \|f(x_i^s) - \phi(s_i)\|^2, \tag{2}$$

where $\phi(s_i)$ is the semantic embedding of the class description/label s_i for the training video x_i^s from a Sent2Vec model [32]. After the model is trained, for ZSL inference, the semantic embedding of the test video $f(x_i^u)$ is found using the transformer model, and the class closest to the video in the semantic space is chosen,

$$_jD_{\cos}(f(x_i^u), \phi(u_j)),$$

where $D_{\cos}$ is the cosine distance and $\phi(u_j)$ is the Sent2Vec semantic embedding of the class description u_j.

Class Descriptions: Due to the unavailability of class descriptions or attributes, existing methods in the literature alternatively use the class labels and Word2Vec to extract the semantic embeddings $\phi(u_j)$. However, we argue that such an approach would severely limit the performance of a model. Specifically, several class labels in all the datasets are not sufficiently distinctive and could refer to various different activities. For example, *diving* in UCF-101 could either refer to *scuba diving*, *cliff diving*, *sky diving* or *springboard diving*. Furthermore, using Word2Vec on multi-word labels is not an efficient technique since averaging the semantic embedding over several words might lose the context. This can be also seen in Fig. 3, where the semantic embedding of several multi-word class labels is mapped incorrectly. To circumvent these issues, we first manually annotate all the training and testing datasets with one line class descriptions. We then use a Sent2Vec [32] model to better capture the semantic information with respect to the context of the class description sentence. The annotated class descriptions are available in the supplementary material. In Section 4.4, we evaluate the contribution of class descriptions by comparing the performance of the proposed ViSET method with class labels and descriptions in an ablation study.

3.3. Fair Zero-Shot Action Recognition

While several recent works have shown promising results on the benchmark datasets, there are still several shortcomings. In contrast to earlier works, recent approaches [4, 19, 20, 34] use large-scale external datasets such as the Kinetics 400/600/700 dataset to pretrain the visual feature extractors. However, as shown in Fig. 1, there is a significant overlap between the seen and unseen classes, $S \cap U \neq \varnothing$, which clearly violates the ZSL paradigm. Recently, Brattoli et al. [4] proposed a novel training protocol which involves removing classes from the training set if

$$\min_{s_i \in S, u_j \in U} D_{\cos}(\phi(s_i), \phi(u_j)) < \tau, \tag{3}$$

where τ is set as 0.05. However, in Fig. 3, we show that even after applying such a constraint, there are several overlapping classes which are not removed. This can be partly attributed to semantic mismatch due to domain shift, where two very similar classes are called by slightly different names. For example, *blowing out candles* and *blowing candles* refer to the same class and yet the cosine distance between them in the semantic space is much greater than τ. Moreover, removing overlapping classes from the training set is a cumbersome approach since it requires retraining computationally expensive models for every individual test set. On the contrary, it would be far easier to remove overlapping classes from the test set as the class information is already available in the training data.

Furthermore, it is also worth considering whether a given test class is completely irrelevant with respect to all the seen classes. Intuitively, even humans cannot comprehend a new activity if it involves interactions and objects that are never seen before. Hence, we also propose removing classes which are significantly different from the activities present in the training dataset. While keeping such classes does not necessarily violate the ZSL premise, it introduces a source of randomness since no reasonable algorithm can be expected to recognize such activities.

In the current literature, most approaches randomly split a single dataset and evaluate performance on it over several trials. Such an evaluation setup is not practical in a real-world scenario since most of the videos would have disjoint

Benchmark Dataset	Selected Classes
UCF-101	8
HMDB-51	3
ActivityNet-101	19

Table 1. Number of classes from each benchmark dataset in the proposed test set for Fair ZSL.

sources. Hence, to ensure a realistic ZSL setting, we extend the setup discussed in [4], which suggests using independent datasets for training and testing. However, [4] does not fully account for overlapping classes between the train and test sets, as seen in Fig. 3. We propose a new ZSL experimentation setup for action recognition, called *Fair ZSL*. For each dataset, completely removing classes that either overlap or completely irrelevant with respect to the train set (Kinetics 600/700) leaves us with a very small number of test classes per test set (as shown for UCF-101 in Fig. 3). Hence, we propose pooling the valid test classes from all benchmark datasets to form a novel test set. In the proposed Fair ZSL setup, there are 30 unique classes[1] from the UCF-101, HMDB-51, and ActivityNet datasets, as shown in Table 1. Each class was carefully handpicked such that it does not violate the ZSL premise. Additionally, the proposed test set is also more robust since it evaluates how well a single model accounts for domain shifts in addition to the ZSL performance.

3.4. Implementation Details

The proposed ViSET model is built upon the PySlowFast [8] package. For training the model, the shorter side of the input video is first resized to 256 pixels and then randomly cropped to form a 224×224 ($H \times W$) video snippet. The patch size is chosen as 16×16, resulting in $N = 196$ patches in a frame. The size of the score vectors ($z_t^p(l)$) at each encoding block for each patch is $q = 768$, and the number of self-attention heads is $A = 12$. The size of the learned semantic embedding $f(x)$ and the Sent2Vec embedding for class descriptions is 600. We train two versions of the proposed transformer, ViSET-8 and ViSET-96 for $F = 8$ and $F = 96$ frames in the input video snippet, respectively. In all our experiments, we use the ViSET-96 model unless explicitly stated otherwise. We train the models on 4 NVIDIA A40 GPUs with a batch size of 16 for ViSET-8 and 4 for ViSET-96. The loss function is minimized via synchronized SGD with a learning rate of 0.002. To extract semantic embeddings, we use the Sent2Vec algorithm proposed in [32].

3.5. Computational Efficiency

Thanks to the scalability of the proposed ViSET model, we are able to vary the length of the input video snippet (i.e., number of frames F), which also leads to an increase in the number of input tokens. In Table 3, we see a significant increase in the performance when the number of input frames are increased from 8 to 96. Increasing the number of video frames is intuitive since it allows a model to better capture the spatiotemporal activities that span several frames. However, due to the current GPU limitations, we are unable to further increase the input length. On the other hand, even after increasing our model complexity to accommodate 96 input frames, our model is still more computationally efficient as compared to the I3D model with 8 input frames, which

[1]The detailed list of all the classes in the proposed split is available in the supplementary material.

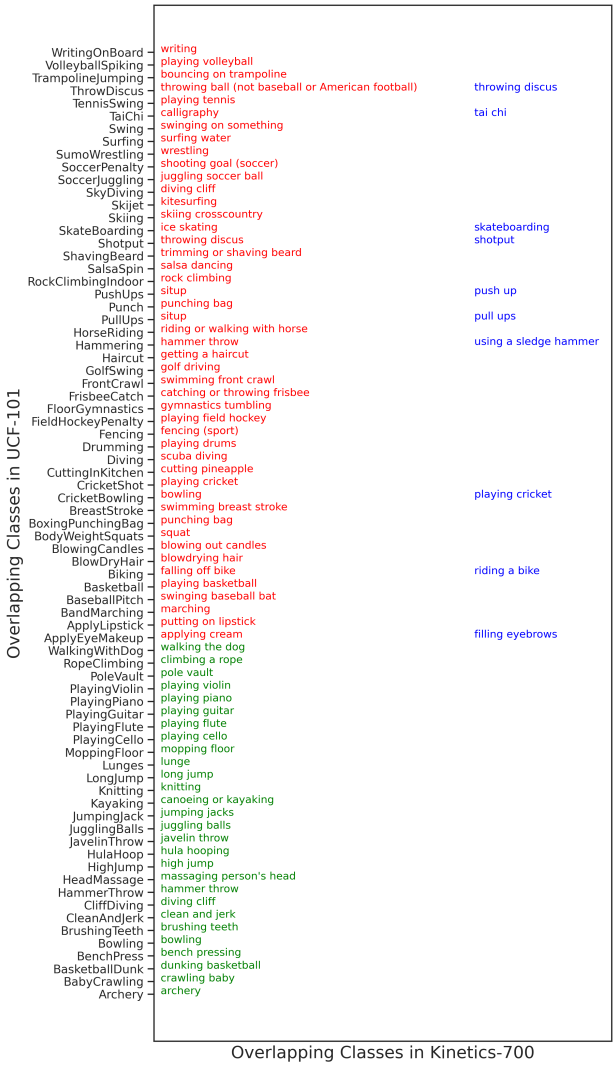

Figure 3. Visualization of the overlapping classes between UCF-101 and Kinetics-600/700. The Kinetics classes in green are considered as overlapping by the metric given by Eq. (3) and removed in [4]. The Kinetics classes in red are given by Word2Vec as the semantically nearest classes to the corresponding UCF-101 class, but not removed in [4] according to the criterion in Eq. (3). For several cases, the actual closest Kinetics classes, shown in blue, are missed by Word2Vec. They are almost identical to the corresponding UCF-101 classes, and thus violate the ZSL idea.

requires 10.8 TFLOPS for inference, in contrast to the proposed ViSET-8 model, which only requires 0.79 TFLOPS, and ViSET-96, which requires 7.57 TFLOPS.

4. Experiments

4.1. Datasets

Most of the recent works evaluate their performance on three publicly available benchmark datasets, namely the UCF-101, HMDB-51 and Olympics dataset. The UCF-101 dataset consists of 13,320 videos from 101 classes, primarily focusing on five types of actions. The HMDB-51 dataset consists of 6767 videos from 51 classes based on daily human actions. The Olympics dataset consists of 16 categories, related to an Olympic sport. Recently, Brattoli et al. [4] evaluated the performance of their approach on the ActivityNet dataset by extracting labelled frames from every video. As compared to the other benchmark datasets, ActivityNet is considerably more comprehensive, consisting of 27,801 videos from 200 classes related to daily activities. The Kinetics-700 dataset is the largest dataset available for video action recognition, with more than 500K videos in 700 categories sourced from YouTube. Since several classes from Kinetics-700 were not available or had files corrupted, we use the Kinetics-600 dataset in our experimental setup. Due to its small size and significant overlap with Kinetics, we do not consider the Olympics dataset in our evaluations.

4.2. Experimental Setup

To analyze the performance of our proposed ViSET model and provide a fair comparison with benchmark algorithms, we first follow the training and evaluation protocols used in the existing papers. The existing training protocols can be broadly classified into two categories, open-ended and restrictive. Most existing works are based on the open-ended formulation, whereas a few recent approaches [4, 20, 35] use the restrictive one.

Open-Ended: In this setup, a model is first trained on a large-scale visual dataset such as the Kinetics dataset without removing any classes, and then evaluated on a smaller application specific dataset such as UCF or HMDB. A few approaches also further fine-tune their models on the smaller dataset by randomly splitting it into train and test sets; however doing so does not necessarily improve performance. Moreover, as also suggested in [4], a model ideally should have separate video sources for training and testing to evaluate its generalization capability.

Restrictive: In the restrictive training approach proposed in [4], we remove all classes from Kinetics-600 whose distance to any class in UCF $\cup$ HMDB is smaller than τ when testing on UCF or HMDB, which results in a subset of Kinetics with 564 classes. For testing on ActivityNet, an even more restrictive approach is proposed which involves removing all classes whose distance to any class in ActivityNet $\cup$ UCF $\cup$ HMDB is smaller than τ. This setting leads to an even smaller subset of Kinetics with 505 classes.

For the open-ended formulation, we train our model on the entire Kinetics-600 dataset and choose to forgo fine-tuning on UCF or HMDB since transformer based models require significant amount of data to learn meaningful representations. To make our evaluation comparable to the existing approaches, we first randomly split the test dataset in half and evaluate our proposed approach on it over 10 trials. For the restrictive approach, in addition to the random split, we also evaluate our model on the entire UCF and HMDB datasets (Table 3) since it allows for a more robust evaluation due to lack of randomness.

Fair ZSL: Since neither of the existing training protocols provide the true essence of ZSL, we propose a new setup where a model is trained on the entire Kinetics-600/700 dataset and evaluated on the proposed test set discussed in Section 3.3. For evaluation, we compare our approach in Table 4 with the recently proposed E2E model [4] and our implementation of the CLASTER model [19]. Recently, [7] proposed an elaborate rehearsal approach which, in addition to a 3D-CNN, explicitly uses an object classifier trained on ImageNet to learn objects detected in videos. Since their approach and problem setup is significantly different from the existing approaches, we do not compare our approach with them.

4.3. Results

In the model name "ViSET-X(Y) + Z", X refers to the number of frames used in the analysis of each video, Y denotes the number of the classes from Kinetics-600 used in training, and Z represents the data type used for class semantic embedding (CD for class descriptions and CL for class labels). The impacts of frame number and class descriptions are analyzed in Section 4.4.

We first compare the proposed method with the state-of-the art approaches under the dominant random split setup in Table 2. However, [18] have presented their results using a different split (e.g., 78-23 for UCF-101), so we do not include their results in Table 2. The test datasets are randomly split into half using the seed 10, as in [4]. It is seen that the proposed spatiotemporal transformer based ZSL approach consistently outperforms all other state-of-the-art approaches under all settings. On the UCF-101 dataset, we notice a significant improvement of 21.9% and 5.2% over the next best results, for the OE and R protocols, respectively. On the HMDB-51 dataset, the improvements of the proposed method are 3.4% and 1.8% for the OE and R protocols, respectively. None of the existing approaches present their results on the ActivityNet dataset for the OE protocol, hence we could not compare our result on it. However, for the R protocol, we outperform the E2E approach

Method	Protocol	UCF	HMDB	ActivityNet
DataAug [49]	OE	18.3	19.7	-
InfDem [34]	OE	17.8	21.3	-
Bidirectional [44]	OE	21.4	18.9	-
TARN [3]	OE	19	19.5	-
Action2Vec [21]	OE	22.1	23.5	-
OD [29]	OE	26.9	30.2	-
CLASTER [19]	OE	46.4	36.8	-
DASZL [27]	OE	48.9	-	-
GGM [29]	OE	20.3	20.7	-
ViSET-96(600) + CD (Ours)	OE	**68.3**	**40.2**	**44.8**
E2E (605classes)	R	44.1	29.8	26.6
E2E (664classes)	R	48	32.7	-
PS-ZSAR (662 classes) [26]	R	49.2	33.8	-
ViSET-96(505) + CD (Ours)	R	45.6	31.3	**35.8**
ViSET-96(564) + CD (Ours)	R	**53.2**	**34.5**	-

Table 2. Comparison with the state-of-the-art methods on standard benchmark datasets using the open-ended (OE) and Restrictive (R) protocols. All the methods are evaluated by randomly splitting the dataset in half and averaging the results over 10 trials.

by a wide margin of 9.2%.

A disjoint training and testing split allows us to evaluate our model on the entire UCF-101, HMDB-51 and ActivityNet datasets without any random split. As shown in Table 3, the performances drop compared to the random split case in Table 2. However, we still considerably outperform the E2E based approach under the R protocol while using the class descriptions and class labels. Since the rest of the existing state-of-the-art approaches use some part of UCF-101 or HMDB-51 for fine tuning their models, we could not compare our results with them in this setup.

Finally, we also consider the Fair ZSL protocol proposed in Section 4.2. Here, we compare our model with the E2E [4] method and our implementation of the CLASTER [19] method. We again notice a significant improvement of 5.9% over E2E, which shows the robustness and generalization capability of the proposed method. It should be noted that we use the E2E model trained on 664 classes for comparison since the model trained on all 700 classes is not available. However, since we train on Kinetics-600, which has even less number of classes, we believe it is a fair comparison.

4.4. Ablation Study

In this section, we analyze the contributions of different components of the proposed approach by performing empirical studies.

Impact of the Spatiotemporal Transformer: Here, we study how well the proposed ViSET model is able to learn the spatiotemporal visual features. For comparison, we consider the a state-of-the-art 3D-CNN model called I3D [6], which has been a popular choice for video action recognition, and has been used by several existing approaches [19, 29, 46] for zero-shot action recognition. For ViSET, we extract the output of the last block $z_0^0(L)$, which serves as a visual embedding for the entire video. To extract I3D features, we follow the approach employed by [19, 29, 46] and average the output from the *Mixed 5c* layer across the temporal dimension followed by pooling by four in the spatial dimension and finally flattening it to a vector of size 4096. We use both RGB and flow features and concatenate them to form a vector of size 8192. In Fig. 5, we present the t-SNE visualization for the I3D and ViSET features on UCF-101, where each point represents a video in the UCF dataset. It is clearly seen that, as compared to the I3D features, ViSET learns more semantically separable features with more well-defined clusters. We also quantitatively compare them in Table 5 by using various statistical metrics. The average silhouette score measures how tightly grouped all the points in the cluster are. The adjusted rand index computes a similarity measure between the clusterings and the ground truth. The homogeneity score checks if a cluster contains samples belonging to a single class. Finally, we also apply a simple k-NN classification algorithm on the extracted features to compute the accuracy for traditional video classification. In all metrics, the ViSET features provide a better separation.

Impact of Class Descriptions: In Table 3, we compare the performance of our approach when class labels (CL) are used instead of class descriptions (CD). Here, we train the ViSET model on the semantic embeddings extracted using the Sent2Vec model on class labels, and test on the entire

Method	Protocol	UCF	HMDB	ActivityNet
E2E (605classes)	R	35.3	24.8	20.0
E2E (664classes)	R	37.6	26.9	-
ViSET-96(505) + CD	R	36.1	25.7	**26.3**
ViSET-96(564) + CD	R	**40.2**	**30.5**	-
ViSET-8(564) + CD	R	37.6	27.9	-
ViSET-96(505) + CL	R	33.9	25.2	24.7
ViSET-96(564) + CL	R	38.3	27.6	-

Table 3. Comparison with the E2E [4] approach under the Restrictive (R) protocol using the entire datasets for testing without any random split. E2E and ViSET are trained on Kinetics-700 and Kinetics-600, respectively.

Method	Protocol	FZSL Split
CLASTER	FZSL	24.3
E2E (664 classes)	FZSL	30.8
ViSET-96(600) + CD	FZSL	**36.7**

Table 4. Comparison with E2E [4] and our implementation of CLASTER [19] using the proposed Fair ZSL protocol. The models are tested on the test set proposed in Section 3.3.

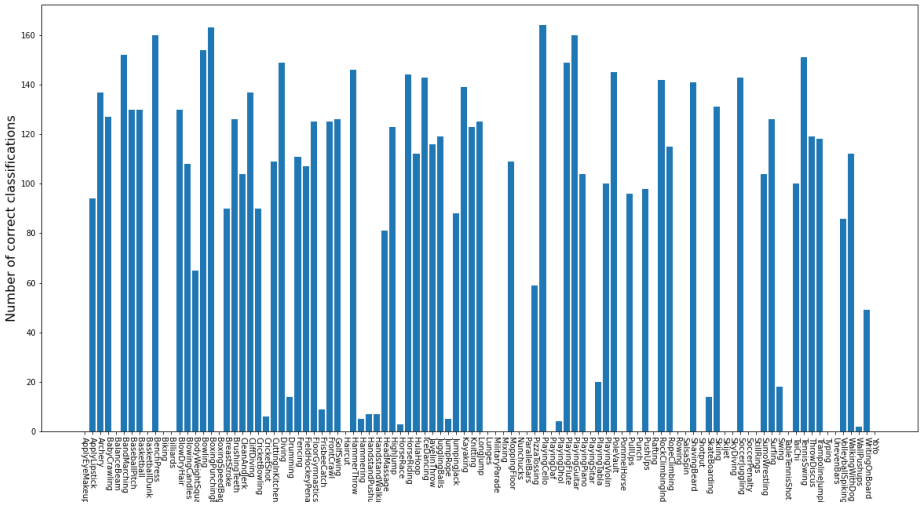

Figure 4. Performance of the proposed approach on all classes of the UCF-101 dataset.

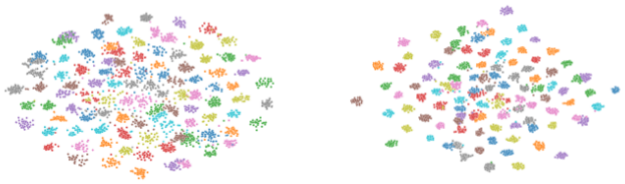

Figure 5. t-SNE visualization of the I3D (left) and ViSET (right) features extracted from the UCF-101 dataset. Each point represents a video and various classes are represented with different colors. We see that the features learned by the proposed ViSET model (right) are semantically more separable than the I3D features (left). Best viewed in color.

Method	I3D	ViSET
Average Silhouette	0.119	**0.196**
Adjusted Rand Index	0.80	**0.88**
Homogeneity Score	0.92	**0.96**
Classification	0.93	**0.96**

Table 5. Comparison between the ViSET and I3D features in terms of clustering and classification performance using different metrics. The classification accuracy is for the k-NN classifier.

UCF, HMDB and ActivityNet datasets using the restrictive (R) setting. We see that there is a noticeable improvement in the performance when class descriptions are used, which ascertains our conjecture that descriptive semantic embeddings are crucial in improving the learned class representations. Moreover, our performance using class label embeddings is still better than existing state-of-the-art, which shows the efficacy of our proposed ViSET model. While generating class descriptions require some human involvement, we argue that the cost is still significantly less as compared to video level annotations required for supervised learning. Furthermore, manually defined attributes are already being used for zero-shot image classification [23] since class labels are not always discriminative enough to distinguish the context.

Class wise performance: In Fig. 4, we breakdown the performance of the proposed model over all the classes in the UCF dataset (i.e., not only the ones included in the proposed test set). We observe that for several classes such as *nunchucks*, *YoYo*, *unevenbars*, the proposed approach is unable to classify even a single video correctly. This problem is not due to the proposed method, but due to the sheer dissimilarity of these classes with respect to the training classes in the Kinetics dataset. Since any practical algorithm will miss such classes, this emphasizes the need for removing classes that are completely irrelevant with respect to the training set from the test set.

Number of Input Frames: Finally, we analyze the impact of the number of input frames on the ViSET performance. As shown in Table 3, there is a 2.6% gain in both UCF and HMDB datasets when 96 frames are used instead of 8.

5. Conclusion

In this work, we introduce a spatiotemporal transformer architecture for zero-shot video action recognition, called ViSET. Moreover, we highlight several areas where the existing approaches either violate the zero-shot learning (ZSL) premise or are unable to perform well due to the limited capabilities of 3D-CNN based visual extractors. We propose a new evaluation protocol, Fair ZSL, that strictly adheres to the ZSL premise. Through several experiments, we show that the proposed approach consistently outperforms the existing approaches under various experimental setups, including the existing ones in the literature and the

proposed Fair ZSL setup.

References

[1] Ioannis Alexiou, Tao Xiang, and Shaogang Gong. Exploring synonyms as context in zero-shot action recognition. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4190–4194. IEEE, 2016. 2

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 1, 2, 3

[3] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019. 2, 7

[4] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4613–4623, 2020. 2, 3, 4, 5, 6, 7

[5] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 1

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 7

[7] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13638–13647, 2021. 6

[8] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3435–3444, 2019. 5

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 1

[12] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4768–4777, 2017. 2

[13] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 2

[14] Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5378–5387, 2015. 2

[15] Chuang Gan, Ming Lin, Yi Yang, Gerard De Melo, and Alexander G Hauptmann. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In *Thirtieth AAAI conference on artificial intelligence*, 2016. 2

[16] Chuang Gan, Ming Lin, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. 2

[17] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 87–97, 2016. 2

[18] Pallabi Ghosh, Nirat Saini, Larry S Davis, and Abhinav Shrivastava. Learning graphs for knowledge transfer with limited labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11151–11161, 2021. 6

[19] Shreyank N Gowda, Laura Sevilla-Lara, Frank Keller, and Marcus Rohrbach. Claster: Clustering with reinforcement learning for zero-shot action recognition. *arXiv preprint arXiv:2101.07042*, 2021. 2, 4, 6, 7

[20] Shreyank N Gowda, Laura Sevilla-Lara, Kiyoon Kim, Frank Keller, and Marcus Rohrbach. A new split for evaluating true zero-shot action recognition. *arXiv preprint arXiv:2107.13029*, 2021. 2, 4, 6

[21] Meera Hahn, Andrew Silva, and James M Rehg. Action2vec: A crossmodal embedding approach to action learning. *arXiv preprint arXiv:1901.00484*, 2019. 2, 7

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[23] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4483–4493, 2020. 8

[24] 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". *Computer Vision and Image Understanding*, 155:1–23, 2017. 2

[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

[26] Alec Kerrigan, Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Reformulating zero-shot action recognition for multi-label actions. *Advances in Neural Information Processing Systems*, 34:25566–25577, 2021. 7

[27] Tae Soo Kim, Jonathan Jones, Michael Peven, Zihao Xiao, Jin Bai, Yi Zhang, Weichao Qiu, Alan Yuille, and Gregory D Hager. Daszl: Dynamic action signatures for zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1817–1826, 2021. 7

[28] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 1, 2

[29] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9985–9993, 2019. 2, 7

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[31] Ashish Mishra, Vinay Kumar Verma, M Shiva Krishna Reddy, S Arulkumar, Piyush Rai, and Anurag Mittal. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 372–380. IEEE, 2018. 2

[32] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. 4, 5

[33] AJ Piergiovanni and Michael S Ryoo. Learning shared multimodal embeddings with unpaired data. *CoRR*, 2018. 2

[34] Alina Roitberg, Ziad Al-Halah, and Rainer Stiefelhagen. Informed democracy: voting-based novelty detection for action recognition. *arXiv preprint arXiv:1810.12819*, 2018. 4, 7

[35] Alina Roitberg, Manuel Martinez, Monica Haurilet, and Rainer Stiefelhagen. Towards a fair evaluation of zero-shot action recognition using external data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2, 6

[36] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 2

[37] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020. 1

[38] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 2

[39] Kevin Tang, Li Fei-Fei, and Daphne Koller. Learning latent temporal structure for complex event detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1250–1257. IEEE, 2012. 1

[40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[41] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[43] Qian Wang and Ke Chen. Alternative semantic representations for zero-shot human action recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 87–102. Springer, 2017. 2

[44] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision*, 124(3):356–383, 2017. 2, 7

[45] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S Yu. Spatiotemporal pyramid network for video action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1529–1538, 2017. 2

[46] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018. 7

[47] Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 63–67. IEEE, 2015. 2

[48] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, 123(3):309–333, 2017. 2

[49] Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *European Conference on Computer Vision*, pages 343–359. Springer, 2016. 2, 7

[50] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 374–390, 2018. 2

[51] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13577–13587, 2021. 2

[52] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9436–9445, 2018. 2