Elaborative Rehearsal for Zero-shot Action Recognition

Shizhe Chen*
Inria, France
shizhe.chen@inria.fr

Dong Huang
Carnegie Mellon University, USA
donghuang@cmu.edu

Abstract

The growing number of action classes has posed a new challenge for video understanding, making Zero-Shot Action Recognition (ZSAR) a thriving direction. The ZSAR task aims to recognize target (unseen) actions without training examples by leveraging semantic representations to bridge seen and unseen actions. However, due to the complexity and diversity of actions, it remains challenging to semantically represent action classes and transfer knowledge from seen data. In this work, we propose an ER-enhanced ZSAR model inspired by an effective human memory technique Elaborative Rehearsal (ER), which involves elaborating a new concept and relating it to known concepts. Specifically, we expand each action class as an Elaborative Description (ED) sentence, which is more discriminative than a class name and less costly than manually-defined attributes. Besides directly aligning class semantics with videos, we incorporate objects from the video as Elaborative Concepts (EC) to improve video semantics and generalization from seen actions to unseen actions. Our ER-enhanced ZSAR model achieves state-of-the-art results on three existing benchmarks. Moreover, we propose a new ZSAR evaluation protocol on the Kinetics dataset to overcome limitations of current benchmarks and first compare with few-shot learning baselines on this more realistic setting. Our codes and collected EDs are released at https://github.com/DeLightCMU/ElaborativeRehearsal.

1. Introduction

Supervised video action recognition (AR) has made great progress in recent years, benefited from new models such as 3D convolutional neural networks [10, 11, 42] and large-scale video datasets [6, 16]. These supervised models require abundant training data for each action class. However, desired action classes are continuously increasing with the explosive growth of video applications on smart phones, surveillance cameras and drones. It is prohibitively expensive to collect annotated videos for each action class to fuel the training needs of existing supervised models. In order to alleviate such burden, Zero-Short Action Recognition (ZSAR) [50] has become a thriving research direction, which aims at generalizing AR models to unseen actions without using any labeled training data of unseen classes.

A common approach for ZSAR is to embed videos and action classes into a joint semantic space [12, 49], so that the associations between video and seen actions can be transferred to unseen actions. However, how to semantically represent action classes for above associations is a challenging problem due to the complexity and diversity of actions. As shown in Figure 1(a), early works employ manual-defined attributes [29] to represent actions. Despite being a natural methodology, it is hard and expensive to define a complete set of atom attributes that generalizes to arbitrary actions. To overcome difficulties in attribute definition, recent works adopt word embeddings of action names [50, 4] as class semantic representations. Though simple and effective, word embeddings can be ambiguous. Words have different meanings in different context and some actions might not even
be interpreted literally according to their names such as the “dumpster diving” action in Figure 1(b), which are confusing to relating different action classes.

In addition to class semantic representations of actions, it has been under-explored in existing ZSAR works on how to learn powerful and generalizable video semantic representations. Only until recently, deep features [19, 41] have been used to overtake traditional hand-crafted features such as fisher vectors of improved dense trajectory descriptors [43, 50]. One line of work [15, 21] utilizes objects recognized by deep image networks as video descriptors, which assumes that object recognition in image domain is prior knowledge for more advanced action recognition. The predicted objects are naturally embedded in the semantic space and thus can be well generalized to recognize actions even without any video example [21]. However, the video is more than collections of objects, but contains specific relationships among objects. Therefore, it is insufficient to represent video contents purely using object semantics. Another direction of works [4], instead, directly employs state-of-the-art video classification networks in ZSAR. Though powerful enough to capture spatio-temporal information in the video, they are prone to overfit on seen action classes and transfer poorly to unseen ones.

In this work, we take inspiration from a well-established human memory technique, namely Elaborative Rehearsal (ER) [3], for ZSAR. When we learn a new item such as “dumpster diving”, we first expand the phrase into a readily comprehensible definition, and then relate the definition to known information in our long-term memory, thereby fostering retention of the item. In a similar manner, we propose an ER-enhanced model to generalize AR models for new actions. Our approach advances ZSAR in three main aspects under the common paradigm of joint semantic space learning [12, 49]: (1) For the class semantic representation of actions, we construct Elaborative Descriptions (ED) from class names to comprehensively define action classes as shown in Figure 1(c), and embed the ED leveraging prior knowledge from pre-trained language models. (2) For the video semantic representation, we propose two encoding network streams that jointly embed spatio-temporal dynamics and objects in videos. We use a pre-trained image object classification model [24] to generate the Elaborative Concepts (EC) of objects. Since it is highly likely that some common objects involved in seen and unseen classes, incorporating EC in video semantics improves the generalization on unseen classes. (3) To further improve generalization of video semantic representations, we propose an ER objective to enforce the model to rehearse video contents with additional semantic knowledge from EC. The embedding of EC shares the same embedding function as the ED of action classes, which also implicitly makes our ZSAR model more generalizable to diverse class semantic representation. Our ER-enhanced ZSAR model achieves state-of-the-art performance on the widely used benchmarks including Olympic Sports [32], HMDB51 [25] and UCF101 [39] datasets.

Moreover, existing ZSAR benchmarks are relative small and contain overlapped classes with video datasets for feature training. In order to benchmark ZSAR progress on a more realistic scenario, we further propose a new ZSAR evaluation protocol based on a large-scale supervised action dataset Kinetics [5, 6]. In our Kinetics ZSAR benchmark, we demonstrate the first case where ZSAR performance is comparable to a simple but strong few-shot learning baseline under clear split of seen and unseen action classes.

2. Related Work

Supervised Action Recognition. The rapid development of deep learning [18] has vigorously promoted AR research. Early deep models [23, 38, 44] adopt 2D convolutional neural networks (CNNs) in temporal domain. To more effectively encode temporal dynamics in videos, 3D CNNs [41] are proposed but are computation and parameter heavy, which require large-scale datasets to train. Therefore, different approaches have emerged to improve 3D CNNs. Carreira et al. [6] propose I3D network which inflates 2D CNN to 3D CNN to learn spatio-temporal features. Tran et al. [42] and Qiu et al. [36] decompose 3D convolution into 2D spatial and 1D temporal convolutions. Wang et al. [46] insert non-local blocks into 3D CNNs to capture long-range dependencies. Feichtenhofer et al. [11] introduce slowfast network with two pathways operating at different frame rates, and further explores expansion of 2D CNNs along space, time, width and depth in [10]. Lin et al. [28] propose temporal shift module (TSM) to achieve temporal modeling at 2D computational costs and parameters. Despite strong performance, these supervised models cannot recognize new classes without training examples. In this work, we generalize the AR models to recognize unseen actions.

Zero Shot Learning. Most ZSL works [1, 2, 12, 47, 48, 53] focus on the image domain to recognize unseen objects. A comprehensive survey can be found in [48]. Here we mainly review joint semantic space based methods. ALE [1], DEVISE [12] and SJE [2] use bilinear compatibility function to associate visual and class representations, with different objectives for training. ESZSL [37] proposes an objective function with closed form solution for linear projection. DEM [53] proposes to use visual space as embedding space to address hubness problem in ZSL. Different from above approaches, Wang et al. [47] predict classification weights based on knowledge graphs of classes. Except using different features, the ZSL methods in image domain can be applied for zero-shot action recognition.

Zero Shot Action Recognition. As the main focus of our work is to learn better video and action semantic repre-
sentations for ZSAR, we group existing works according to types of semantic representation of actions. The first type takes manual-defined attributes [29, 52] to represent an action. Gan et al. [14] improve attribute detection via multi-source domain generalization. Nevertheless, the attributes of actions are harder to define compared with the image counterparts. The second type then exploits objects as attributes. Jain et al. [21] detect objects in videos and associate videos to action category with maximum object similarity. Gan et al. [13] propose to select discriminative concepts. Gao et al. [15] utilize graph networks to learn action-object relationships and then match objects in the video with action prototypes. Though effective, the above work ignore spatio-temporal relationships in videos and actions. The third type of approaches uses word embedding of action names [4, 30, 35, 50] as semantic representation. Qin et al. [35] derive error-correcting output codes for actions via both category-level embedding and intrinsic data structures. The recent work [4] argues that end-to-end training is important for ZSAR and proposes to train a 3D CNN to predict word embedding of action names. However, word embeddings can be ambiguous and mislead knowledge transfer among action classes. The most similar work to ours is [45], which employs texts and images as alternative semantic representation for actions, but their text descriptions are rather noisy and inferior to attributes or word embeddings.

### 3. Our Approach

In ZSAR, we are given a source dataset $D^s = \{ (v^n, y^n) \}_{n=1}^N$ of $N$ video with labels from seen action classes $S = \{ 1, \cdots , S \}$, where $v^n$ is a video clip and $y^n \in S$ is the label. $D^f = \{ (v^m) \}_{m=1+N}^{N+M}$ is the target dataset of $M$ videos with labels from unseen action classes $T = \{ S+1, \cdots , S+T \}$. The goal of ZSAR is to classify $v^m$ in $D^f$ over unseen classes $T$ with AR models only trained on $D^s$. Following [12, 49], the main architecture of our ZSAR model embeds videos and action classes (texts) into a joint semantic space, in which classes of similar semantics are closer as nearest-neighbors. Their respective embedding functions are a video embedding function $\phi(v)$ and an action class embedding function $\psi(y)$. Both functions are only trained on $D^s$, and to be tested on $D^t$.

In the rest of the section, we present the novel components of our ER-enhanced ZSAR model: Elaborative Description (ED), action class embedding function $\phi(y)$, video embedding function $\phi(v)$, and Elaborative Rehearsal (ER) loss. The framework is illustrated in Figure 2.

#### 3.1. Elaborative Description (ED)

For each action class name or each object concept, we concatenate a name and its sentence-based definition as an ED. Examples of EDs are listed in Table 1, which are more discriminative than class names and easier to generate than attributes to semantically represent an action or an object.

**Justification for Human Involvement.** ZSL demands class-wise semantic representations, which might involve human to construct, but costs significantly less than sample-wise annotation efforts in supervised training. In fact, it is a vital step of ZSL to design a high-quality semantic representation with less class-wise annotating efforts. For ZSL on general object classification task [12, 27, 48], word embeddings of class names are gaining popularity as semantic representation, because the semantic embeddings of general object words are well-learned in pre-trained language models and can be used as prior knowledge. However, word embeddings are not applicable to other domains such as fine-grained ZSL for bird species [20] where the class name provides little information about visual appearances. Manual-defined attributes [20] or cleaned text descriptions [33] are required in such scenarios. The situation is similar in ZSAR, where action names alone are not discriminative enough to represent context of the action. For example, the action “fidgeting” in the Kinetics dataset [6] denotes “playing fidget spinner” instead of its common meaning of “making small movements”. Therefore, it is necessary for human involvement to clarify the action definitions. Compared to carefully designed and annotated attributes, a more natural way for us humans is to describe the visual process of target actions in natural language, which motivates us to collect sentence-based ED for action class representation.

**Construction of Elaborative Description.** Defining actions is more complicated than objects. In the ImageNet

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Elaborative Description (ED)</th>
<th>ED Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action Class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cleaning gutters</td>
<td>cleaning gutters: make clean; remove dirt, marks, or stains from; a shallow trough fixed beneath the edge of a roof for carrying off rainwater.</td>
<td>Wikipedia + Dictionary + Modification</td>
</tr>
<tr>
<td>clean and jerk</td>
<td>clean and jerk: a two - movement weightlifting exercise in which a weight is raised above the head following an initial lift to shoulder level.</td>
<td></td>
</tr>
<tr>
<td>Object Concept</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chipboard</td>
<td>chipboard: a cheap hard material made from wood chips that are pressed together and bound with synthetic resin</td>
<td>WordNet</td>
</tr>
</tbody>
</table>

Table 1: Examples of Elaborative Descriptions (ED) for action classes and object concepts.
Therefore, we did not observe performance gains using content words are more strengthened than other stopwords. Since there are multi-layers of self-attention in BERT, the action class embedding function (left) extends action names to EDs towards the video. The ER loss utilizes recognized object semantics $z_o$ (right) to match $[x_{vo}, x_{ov}]$ which improves the action recognition loss.

Figure 2: Architecture of our ER-enhanced ZSAR model. The action class embedding function (left) extends action names to EDs towards action class embedding $z$. The multimodal video embedding function generates spatio-temporal and object features $[x_{vo}, x_{ov}]$ (middle) for the video. The ER loss utilizes recognized object semantics $z_o$ (right) to match $[x_{vo}, x_{ov}]$ which improves the action recognition loss.

dataset [8], object classes are directly linked to concepts in WordNet [31], and thus EDs of objects are straightforward to obtain. However, currently there are no such resources to define actions. To reduce manual efforts of writing EDs from scratch, we first automatically crawl candidate sentences from Wikipedia and dictionaries using action names as queries. Then we ask annotators to select and modify a minimum set of sentences from the candidates to describe the target action given few video exemplars. More details are presented in the supplementary material. It takes less than 20 seconds on average to generate the ED per action class, which is very efficient. The average length of EDs for actions in the Kinetics dataset [6] is 36 words.

### 3.2. Action Class Embedding $\psi(y)$

Denote $d = \{w_1, \ldots, w_{N_d}\}$ the ED for action $y$, where $w_i$ is the composed word. The goal of action class embedding is to encode $d$ into semantic feature $z \in \mathbb{R}^K$ with dimension of $K$.

In order to capture the sequential order in $d$ and transfer prior knowledge from large-scale text models, unlike previous works that use tf-idf [34], average word embedding [50] or RNNs trained from scratch [53], we propose to employ a pre-trained BERT model [9] for ED encoding. The BERT model has demonstrated excellent capability in implicitly encoding commonsense knowledge [7], which is beneficial to embed global semantics of the sentences.

Denote $h_i \in \mathbb{R}^{768}$ as the hidden state from the last layer of BERT for word $w_i$, we apply average pooling to obtain a sentence-level feature $h$:

$$h = \frac{1}{N_d} \sum_{i=1}^{N_d} h_i. \quad (1)$$

Since there are multi-layers of self-attention in BERT, the content words are more strengthened than other stopwords. Therefore, we did not observe performance gains using more complicated methods to aggregate $h_i$ than our average pooling. Then we use a linear transformation layer to translate $h$ into the joint semantic embedding space:

$$\hat{z} = W_c h + b_c, \quad (2)$$

where $W_c \in \mathbb{R}^{K \times 768}, b_c \in \mathbb{R}^K$ are parameters to learn. Finally, the action class embedding is normalized as $z = \hat{z} / ||\hat{z}||_2$.

### 3.3. Multimodal Video Embedding $\phi(v)$

Unseen action classes may involve both novel spatio-temporal features and objects. For better generalization, we propose to encode videos in two modality streams to capture both the spatio-temporal dynamics and object semantics.

**Spatio-Temporal Stream in Visual Modality.** Encouraged by the recent success of 3D CNNs in supervised AR, we employ 3D CNNs, specifically TSM [28], to extract Spatio-Temporal (ST) features. Denote $\bar{x}_v \in \mathbb{R}^{2048}$ as the output from the last pooling layer of TSM, we map $\bar{x}_v$ into the joint embedding space through linear transformation:

$$\hat{x}_v = W_v \bar{x}_v + b_v, \quad (3)$$

where $W_v \in \mathbb{R}^{K \times 2048}, b_v \in \mathbb{R}^K$ are parameters to learn. We also normalize the embedding as $x_v = \hat{x}_v / ||\hat{x}_v||_2$.

**Object Stream in Text Modality.** It is a widely acknowledged assumption that the objects associated with actions are prior knowledge available to the ZSAR model [15, 21]. We leverage objects automatically recognized from frames to construct a video representation in text modality. Specifically, we use the BiT model [24] pretrained on ImageNet21k dataset to predict object probabilities from evenly sampled 8 frames from video $v$. The object probabilities over frames are averaged and only the top $N_o$ objects $O = \{o_1, \ldots, o_{N_o}\}$ are preserved and embedded in a concatenated sequence:

$$x_o = \psi([\text{ED}(o_1); \ldots; \text{ED}(o_{N_o})]), \quad (4)$$
where ED($o_i$) denotes ED of object $o_i$ as in Table 1. Here we use the same $\psi(\cdot)$ function as in the action class embeddings, which explicitly encourages that the object embedding $x_o$ from videos and action class embedding $z$ from action names to lie in the same semantic space.

**Multimodal Channel Attention.** The awareness of object semantics can focus the spatio-temporal stream to object-highlighted video features, while object semantics can be enriched with motion features. We thus further propose to dynamically fuse the two embeddings, $x_v$ and $x_o$, to enhance each other. The formula of injecting $x_o$ to improve $x_v$ is as follows:

$$g_{vo} = \sigma(W_{vo}^2 \text{RELU}(W_{vo}^1[x_v; x_o])), \quad (5)$$

$$x_{vo} = x_v g_{vo} / ||x_v g_{vo}||_2, \quad (6)$$

where $W_{vo}^1 \in \mathbb{R}^{2K \times K}$, $W_{vo}^2 \in \mathbb{R}^{K \times K}$ are parameters, $\sigma$ is the sigmoid function. Similarly, we obtain $x_{ov}$ from object embedding $x_v$ with guidance of $x_o$. Therefore, our video encoder $\phi(v)$ produces two video embeddings $x_{vo}$ and $x_{ov}$ to comprehensively represent the videos.

### 3.4. Elaborative Rehearsal (ER) enhanced Training

Following the standard ZSAR training, given video $v^m$ of seen classes $S$, we first generate the video embeddings $[x_{vo}^m, x_{ov}^m] = \phi(v^m)$ and action class embedding matrix $Z \in \mathbb{R}^{K \times S}$ where each column $z_i = \psi(i)$. Then, we compute video-action similarity scores:

$$p_c^m = x_{vo} \cdot Z, \quad p_o^m = x_{ov} \cdot Z, \quad (7)$$

where $\cdot$ denote vector-matrix multiplication, $p_c^m, p_o^m \in \mathbb{R}^S$.

As the negative score between object and action class embeddings mainly indicates that the recognized objects are irrelevant to the action, the magnitude is less important. We thus can fuse the two similarity scores as:

$$p^m = p_c^m + \max(p_o^m, 0). \quad (8)$$

We use a standard contrastive loss to train the action recognition model. To be generalizable, $p \in \mathbb{R}^C$ denotes the predicted score, $q \in \mathbb{R}^C$ is the ground-truth label where $q_i = 1$ if the $i$-th label is true otherwise $q_i = 0$, and $C$ is the number of classes. The contrastive loss is:

$$L(p, q) = -\frac{1}{C} \sum_{j=1}^{C} q_j \log \frac{\exp(p_j / \tau)}{\sum_{i=1}^{C} \exp(p_i / \tau)}, \quad (9)$$

where $\tau$ is a temperature hyper-parameter. To conduct action recognition on seen data $D^s$, we convert label $y^m$ into one-hot vector $q^m$ and the loss is:

$$L_{ar} = \frac{1}{N} \sum_{n=1}^{N} L(p^m, q^m) + L(p_o^m, q^m) + L(p_c^m, q^m). \quad (10)$$

Because $x_{vo}$ is more powerful than $x_{ov}$, the model trained on $D^s$ tends to overweight $x_{vo}$ in Eq. 8 and results in over-fitting. Therefore, we average the three losses in Eq. 10.

Moreover, as there are only $S$ semantic labels in $L_{ar}$ as semantic supervision, the learned video and text representations might be less generalizable to more diverse semantics. To address this problem, we further propose an Elaborative Rehearsal (ER) loss, which rehearses the video representation with semantics from ECs obtained from frame-wise object classification. Denote $O^n = \{o_1^n, \cdots, o_{N_t}^n\}$ the top recognized objects in video $v^n$, we generate semantic representation $\psi(ED(o^n))$ for each $o^n$. Since the total number of all objects are large, we sample top few object classes during training for efficiency. Let $Z^o$ be the object class embeddings in a mini-batch of training, and the ER loss is computed as:

$$L_{er} = \frac{1}{N} \sum_{n=1}^{N} L(p_c^n, q_c^n) + L(p_o^n, q_o^n) + L(p_{c,o}^n, q_{c,o}^n), \quad (11)$$

where $p_c^n = x_{vo} \cdot Z^o$, $p_o^n = x_{ov} \cdot Z^o$, $p_{c,o}^n = p_c^n + p_o^n$, and $q_{c,o}^n$ is the ground-truth object labels for $v^n$.

Finally, we combine $L_{ar}$ and $L_{er}$ in our ZSAR model training with a balance factor $\lambda$:

$$L = L_{ar} + \lambda L_{er}. \quad (12)$$

Comparing to Eq. 10, our model trained by Eq. 12 learns a shared $\psi(\cdot)$ from ECs (i.e. $\psi(o_i)$), and ED (i.e., $\psi(y_i)$). The sharing advocates to learn more comprehensive associations between videos and classes in the common semantic space defined by ($\phi(\cdot)$, $\psi(\cdot)$), and thus leads to better generalization to unseen classes.

In inference, the action class of $v^m \in D^t$ is recognized with the highest similarity score:

$$\hat{y}^m = \arg\max_{y \in T} (x_{vo} \cdot \psi(y) + \max(x_{ov} \cdot \psi(y), 0)) \quad (13)$$

where $x_{vo}^m, x_{ov}^m = \phi(v^m)$.

### 4. Experiments

### 4.1. Datasets and Splits

**Existing ZSAR Benchmarks.** Olympic Sports [32], HMDB51 [25] and UCF101 [39] are the three most popular datasets used in existing ZSAR papers [22], which contain 783, 6766 and 13320 videos of 16, 51, 101 action categories respectively. For robust evaluation, Xu et al. [50] proposed to evaluate on 50 independent data splits and report the average accuracy and standard deviation. In each split, videos of 50% randomly selected classes are used for training and the remaining 50% classes are held unseen for testing. We adopt the same data splits as [50] for fair comparison.
There are two major limitations in the above ZSAR protocols. Firstly, it is problematic to use deep features pretrained on other large-scale supervised video datasets because there exist overlapped action classes between pretraining classes and testing classes. Secondly, the size of training and testing data is small which leads to large variations among different data splits, so that abundant numbers of experiments are necessary to evaluate a model. To address these limitations, Brattoli et al. [4] proposed another setting which excludes classes overlapped with the above testing dataset in pre-training dataset Kinetics. Nevertheless, their overlapped class selection algorithm is too tender, leaving the testing classes still seen in the training. Moreover, new end-to-end training of video backbones is needed because this setting does not follow the official Kinetics data split. Therefore, in this work, we propose a more realistic, convenient and clean ZSAR protocol.

Our Proposed Kinetics ZSAR Benchmark. The evolution of the Kinetics dataset [5, 6] naturally involves increment of new action classes: Kinetics-400 and Kinetics-600 datasets contains 400 and 600 action classes, respectively. Due to some renamed, removed or split classes in Kinetics-600, we obtain 220 new action classes outside of Kinetics-400 after cleaning. Therefore, we use 400 action classes in Kinetics-400 as seen classes for training. We randomly split the 220 new classes in Kinetics-600 into 60 validation classes and 160 testing classes respectively. We independently split the classes for three times for robustness evaluation. As shown in our experiments, due to the large-size training and testing sets, the variations of different splits are significantly smaller than previous ZSAR benchmarks. In summary, our benchmark contains 212,577 training videos from Kinetics-400 training set, 2,682 validation videos from Kinetics-600 validation set and 14,125 testing videos from Kinetics-600 testing set on average of the three splits. More details of our evaluation protocol are in the supplementary material.

4.2. Implementation Details

For action class embedding, we use a pretrained 12-layer BERT model [9], and finetune the last two layers if not specified. For video embedding, we use TSM [28] pretrained on Kinetics-400 in the spatio-temporal stream for Kinetics benchmark, and BiT image model [24] pretrained on ImageNet for the other three benchmarks to avoid overlapped action classes in Kinetics; the object stream uses BiT image model [24] pretrained on ImageNet21k [8] and top-5 objects are selected for each video. The above backbones are fixed for fast training. We use one Nvidia TITAN RTX GPU for the experiments. More details are presented in the supplementary material. We set the dimensionality $K = 512$ of the common semantic space, $\tau = 0.1, \lambda = 1$ in the loss and use top-5 objects in the ER loss. We use ADAM algorithm to train the model with weight decay of 1e-4. The base learning rate is 1e-4 with warm-up and cosine annealing. The model was trained for 10 epochs except on the Olympic Sports dataset where we train 100 epochs due to its small training size. The best epoch is selected according to performance on the validation set. Top-1 and top-5 accuracies (%) are used to evaluate all models.

4.3. Evaluation on Existing ZSAR Benchmarks

We compare our model with: (1) Direct/Indirect Attribute Prediction (DAP, IAP) [26]; (2) Human Actions by Attribute (HAA) [29]; (3) Self-training method with SVM and semantic Embedding (SVE) [49]; (4) Embarrassingly Simple Zero-Shot Learning (ESZSL) [37]; (5) Structured Joint Embedding (SJE) [2]; (6) Multi-Task Embedding (MTE) [51]; (7) Zero-Shot with Error-Correcting Output Codes (ZSECOC) [35]; (8) Universal Representation (UR) model [54]; (9) Objects2Action (O2A) [21]; (10) Alternative Semantic Representation (ASR) [45], which uses text descriptions and images as alternative class embedding; (11) TS-GCN [15] which builds graphs among action and object classes with ConceptNet for better action class embedding; (12) End-to-End Training (E2E) [4] which uses a reduced Kinetics training set by excluding part of action classes overlapped with testset. All above methods are evaluated on the inductive ZSL setting, where the videos of unseen action classes are unavailable during training. The unseen action classes are not used in training except [15].

Table 2 presents the comparison. To avoid leaking information from features pretrained on Kinetics video dataset, we only use image features and predicted objects from a

<table>
<thead>
<tr>
<th>Method</th>
<th>Video</th>
<th>Class</th>
<th>Olympics</th>
<th>HMDB51</th>
<th>UCF101</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAP [26]</td>
<td>FV</td>
<td>A</td>
<td>45.4 ± 12.8</td>
<td>N/A</td>
<td>15.9 ± 1.2</td>
</tr>
<tr>
<td>IAP [26]</td>
<td>FV</td>
<td>A</td>
<td>42.3 ± 12.5</td>
<td>N/A</td>
<td>16.7 ± 1.1</td>
</tr>
<tr>
<td>HAA [29]</td>
<td>FV</td>
<td>A</td>
<td>46.1 ± 12.4</td>
<td>N/A</td>
<td>14.9 ± 0.8</td>
</tr>
<tr>
<td>SVE [49]</td>
<td>BoW</td>
<td>W$_N$</td>
<td>13.0 ± 2.7</td>
<td>10.9 ± 1.5</td>
<td></td>
</tr>
<tr>
<td>ESZSL [37]</td>
<td>FV</td>
<td>W$_N$</td>
<td>39.6 ± 9.6</td>
<td>18.5 ± 2.0</td>
<td>15.0 ± 1.3</td>
</tr>
<tr>
<td>SJE [2]</td>
<td>FV</td>
<td>W$_N$</td>
<td>28.6 ± 4.9</td>
<td>13.3 ± 2.4</td>
<td>9.9 ± 1.4</td>
</tr>
<tr>
<td>SJE [2]</td>
<td>FV</td>
<td>A</td>
<td>47.5 ± 14.8</td>
<td>N/A</td>
<td>12.0 ± 1.2</td>
</tr>
<tr>
<td>MTE [51]</td>
<td>FV</td>
<td>W$_N$</td>
<td>44.3 ± 8.1</td>
<td>19.7 ± 1.6</td>
<td>15.8 ± 1.3</td>
</tr>
<tr>
<td>ZSECOC [35]</td>
<td>FV</td>
<td>W$_N$</td>
<td>59.8 ± 5.6</td>
<td>22.6 ± 1.2</td>
<td>15.1 ± 1.7</td>
</tr>
<tr>
<td>UR [54]</td>
<td>FV</td>
<td>W$_N$</td>
<td>24.4 ± 1.6</td>
<td>17.5 ± 1.6</td>
<td></td>
</tr>
<tr>
<td>ASR [45]</td>
<td>C3D$^*$</td>
<td>W$_T$</td>
<td>21.8 ± 0.9</td>
<td>24.4 ± 1.0</td>
<td></td>
</tr>
<tr>
<td>TS-GCN [15]</td>
<td>Obj$^*$</td>
<td>W$_N$</td>
<td>56.5 ± 6.6</td>
<td>23.2 ± 3.0</td>
<td>34.2 ± 3.1</td>
</tr>
<tr>
<td>E2E [4]</td>
<td>r(2+1)d$^*$</td>
<td>W$_N$</td>
<td>32.7</td>
<td>48</td>
<td></td>
</tr>
</tbody>
</table>

Ours (S+Obj)$^*$ | ED     | 60.2 ± 8.9 | 35.3 ± 4.6 | 51.8 ± 2.9 |
Moreover, compared with the results in Table 2, the performance variations on different splits are much lower than those in previous benchmarks, which further proves the superiority of our benchmark for future ZSAR research.

### 4.4.2 Ablation Studies

We present the following Q&As to prove the effectiveness of our proposed semantic representations and the ER training objectives. More hyper-parameter ablation and analysis are in the supplementary material. All the ablation studies below are carried out on the Kinetics ZSAR benchmark.

**Is human involvement necessary for action class representation?** In Table 4a, we compare different action class representations including action class names ($W_N$), Wikipedia entries (Wiki), Dictionary definitions (Dict) and the manually modified EDs. All the models use TSM video features and the $L_{ar}$ objective in Eq. 10 for training. The $W_N$ is encoded with pre-trained Glove word embedding while others are encoded by BERT because we observe that BERT is not suitable to encode short text such as the class names. We can see that the automatically crawled texts of the action class are very noisy which are even inferior to the ambiguous class names. However, with a minimal manual clean of crawled descriptions, we achieve significant improvements such as 8.5% absolute gains on top-5 accuracy compared to $W_N$. This proves that even such easy human involvement is beneficial to the class representation quality as justified in Section 3.1, and ED is more discriminative action class prototype than word embedding.

**How much improvements are from the pre-trained BERT model?** In Table 4b, we compare different action class encoding modules for EDs. AvgPool, AttnPool and RNN all transfer knowledge from a pre-trained Glove word embedding, and apply average pooling, attentive weighted pooling and bi-directional GRU respectively to encode the ED sentence. Similar to Table 4a, all the models use TSM video features and are trained with $L_{ar}$. The pretrained BERT significantly boosts the performance over the other three encoding modules, demonstrating its effectiveness to understand action descriptions.

**Is the ER loss beneficial?** Table 4c compares models trained with or without ER loss. The generalization ability on unseen actions is boosted by a large margin through the ER-enhanced training for both ST and object features. The ER loss augments the semantic labels for videos from automatic elaborative concepts, making the features more generalizable to unseen classes.

**Whether ST features and object features are complementary?** The object features alone “Obj” in Table 4d are comparable with ST features on top-1 accuracy, but are worse than ST on top-5 accuracy. Their combination “ST+Obj” via the proposed multimodal channel attention

---

**Table 3: ZSAR performance on the proposed Kinetics benchmark. Notations are the same as Table 2; ST: spatio-temporal feature.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Video</th>
<th>Class</th>
<th>top-1</th>
<th>top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEVISE [12]</td>
<td></td>
<td></td>
<td>23.8 ± 0.3</td>
<td>51.0 ± 0.6</td>
</tr>
<tr>
<td>ALE [1]</td>
<td></td>
<td></td>
<td>23.4 ± 0.8</td>
<td>50.3 ± 1.4</td>
</tr>
<tr>
<td>SJE [2]</td>
<td>ST</td>
<td></td>
<td>22.3 ± 0.6</td>
<td>48.2 ± 0.4</td>
</tr>
<tr>
<td>DEM [53]</td>
<td></td>
<td>$W_N$</td>
<td>23.6 ± 0.7</td>
<td>49.5 ± 0.4</td>
</tr>
<tr>
<td>ESZSL [37]</td>
<td></td>
<td></td>
<td>22.9 ± 1.2</td>
<td>48.3 ± 0.8</td>
</tr>
<tr>
<td>GCN [17]</td>
<td></td>
<td></td>
<td>22.3 ± 0.6</td>
<td>49.7 ± 0.6</td>
</tr>
<tr>
<td>Ours</td>
<td>ST</td>
<td>ED</td>
<td>37.1 ± 1.7</td>
<td>69.3 ± 0.8</td>
</tr>
<tr>
<td>Ours ST+Obj</td>
<td>ED</td>
<td></td>
<td><strong>42.1 ± 1.4</strong></td>
<td><strong>73.1 ± 0.3</strong></td>
</tr>
</tbody>
</table>

1 We observe large performance variations with different random weight initialization, which mainly results from the small training set.
achieves the best performance on the Kinetics ZSAR setting. This shows that object features alone are not discriminative enough, compared to ST features, to differentiate actions. But adding object features enriches ST with the shared semantic embedding among the action classes.

### Whether EDs are universal representations for both actions and objects?

Though we show that ED is beneficial to represent action classes, it remains a question whether ED also improves semantic representation for objects. To be noted, the EDs for objects are automatically extracted from WordNet thanks to the good correspondence between ImageNet classes and WordNet concepts. Therefore, we replace the ED with the class name of the object in Eq. 4 for video object embedding, and in Eq. 11 for the ER training objective. From Table 4e, we see that even though objects are less ambiguous than actions, it is still beneficial to use its ED instead of its class name.

We provide more hyper-parameter ablations and analysis in the supplementary material.

#### 4.5. Comparison with Supervised Learning

Previous ZSAR works mainly benchmark the progress with respect to zero-shot methods. However, it is interesting to know how well the state-of-the-arts ZSAR methods really work from a practical prospective of video action recognition. We present one of the first attempts for this purpose.

In Table 5, we compare our ZSAR model with supervised models trained with different numbers of labeled videos of unseen classes in our Kinetics ZSAR benchmark. To avoid overfitting on few training samples, we use the same fixed ST features from TSM and only train a linear classification layer. It serves as a simple but strong baseline for few-shot learning as suggested in [40]. Our ER-enhanced ZSAR model improves over the 1-shot baseline by a large margin, but is still inferior to the model using 2 labeled videos per classes. Although our work is the new state-of-the-arts in Table 2 and 3, it only establishes a start-

### 5. Conclusion

We present an Elaborative Rehearsal (ER) enhanced model to advance video understanding under the zero-shot setting. Our ER-enhanced ZSAR model leverages Elaborative Descriptions (EDs) to learn discriminative semantic representation for action classes, and generates Elaborative Concepts (ECs) from prior knowledge of image-based classification to learn generalizable video semantic representations. Our model achieves state-of-the-art performances on existing ZSAR benchmarks as well as our newly proposed more realistic ZSAR setting based on the Kinetics dataset. We demonstrated the potential that our new state-of-the-art on ZSAR benchmarks start to catch up with the supervised AR baselines. In the future, we will explore the unification of zero-shot and few-shot for action recognition.

#### Acknowledgments

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00340.
References


