Global Semantic Descriptors for Zero-Shot Action Recognition

Valter Estevam, Rayson Laroca, Helio Pedrini, *Senior Member, IEEE*, and David Menotti

***Abstract*—The success of zero-shot action recognition (ZSAR) methods is intrinsically related to the nature of semantic side information used to transfer knowledge, although this aspect has not been primarily investigated in the literature. This work introduces a new ZSAR method based on the relationships of actions-objects and actions-descriptive sentences. We demonstrate that representing all object classes using descriptive sentences generates an accurate object-action affinity estimation when a paraphrase estimation method is used as an embedder. We also show how to estimate probabilities over the set of action classes based only on a set of sentences without hard human labeling. In our method, the probabilities from these two global classifiers (*i.e.*, which use features computed over the entire video) are combined, producing an efficient transfer knowledge model for action classification. Our results are state-of-the-art in the Kinetics-400 dataset and are competitive on UCF-101 under the ZSAR evaluation. Our code is available at https://github.com/valterlej/objsentzsar.**

***Index Terms*—Zero-shot learning, sentence representation, video captioning, object recognition.**

I. Introduction

DEEP learning has been applied in human action recognition (HAR) in videos with remarkable results in the last decade [1], [2]. Deep models require many annotated samples for each class we want to classify, typically hundreds of videos. Currently, Kinetics-700 [3] is the largest (HAR) dataset, with 700 action classes and at least 700 videos per class, totaling 647,907. Even considering this large number of actions, numerous more are to be collected and annotated in the real world, demanding intensive human labor and retraining supervised models with the new data. These limitations in the supervised learning paradigm motivate the (ZSAR) problem.

A ZSAR method aims to classify samples from unknown classes, *i.e.,* classes that were unavailable in the model training phase. This goal can only be achieved by transferring knowledge from other models and adding semantic information [4]. Usually, the videos are embedded by off-the-shelf Convolutional Neural Networks (CNNs) (*e.g.*, C3D [5], i3D [1]), and the labels are encoded by attributes or word vectors (*e.g.*, Word2Vec [6], GloVe [7] or Fast Text [8]). As shown in [4], methods based on attributes frequently perform better than versions based on deep encoding. Nevertheless, annotating classes with attributes is not scalable. A strategy to overcome the limitation imposed by human annotation is to take a set of objects as attributes and pre-compute descriptors in a semantic space [9], [10], [11], [12]. Hence, we can recognize a set of objects in a video (*e.g.*, using a pre-trained CNN) and infer the most compatible human action.

For example, Jain et al. [9] introduced a method to relate objects and actions by incorporating semantic information in the form of object labels encoded with Word2Vec embeddings improved by Gaussian mixtures. In their approach, a set of objects is recognized by selecting frames from the videos and averaging the object probability estimations from a CNN pre-trained on ImageNet [13]. Posteriorly, Mettes and Snoek [14] introduced the concept of spatial-aware object embeddings in which an action signature is computed by locating objects and humans. Their label encoding was computed with Word2Vec.

Bretti and Mettes [11], on the other hand, proposed a method to improve the predictions of objects by considering object-scene compositions. They also employed Sentence-BERT (SBERT) (as used in [15] and [16]) to compute sentence embeddings over object-scene label compositions. However, unlike us, they did not observe a significant improvement compared to adopting word embeddings (using Fast Text), probably because they did not provide sufficient semantic information to the model. Finally, Mettes et al. [10] investigated some prior knowledge such as person/object location and spatial relation, expanding previous works [11], [14]. They also investigated semantic ambiguity by adopting label embeddings in languages other than English.

Estevam et al. [16] demonstrated that the automatic generation of sentences employing video captioning models [17] can be used as a significant global semantic descriptor providing information on actors, objects, scenes, and their relationships. They also demonstrated how important it is to represent actions not with a single label (*e.g.,* [9], [14]), nor with a single or a few sentences (*e.g.,* [15]), but with one or two dozens of descriptive sentences leveraging the knowledge transfer from pre-trained paraphrase estimation models [18].

Manuscript received 10 June 2022; revised 3 August 2022; accepted 7 August 2022. Date of publication 22 August 2022; date of current version 2 September 2022. This work was supported in part by the Federal Institute of Paraná, Federal University of Paraná and in part by the National Council for Scientific and Technological Development (CNPq) under Grants 309330/2018-1 and 308879/2020-1. The Titan Xp and Quadro RTX 8000 GPUs used for this work were donated by the NVIDIA Corporation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yue Deng. *(Corresponding author: Valter Estevam.)*

Valter Estevam is with the Federal Institute of Paraná, Irati 84500-000, Brazil, and also with the Federal University of Paraná, Curitiba 81531-970, Brazil (e-mail: valter.junior@ifpr.edu.br).

Rayson Laroca and David Menotti are with the Federal University of Paraná, Curitiba 81531-970, Brazil (e-mail: raysonlaroca@gmail.com; menotti@inf.ufpr.br).

Helio Pedrini is with the University of Campinas, Campinas 13083-852, Brazil (e-mail: helio@ic.unicamp.br).

Digital Object Identifier 10.1109/LSP.2022.3200605

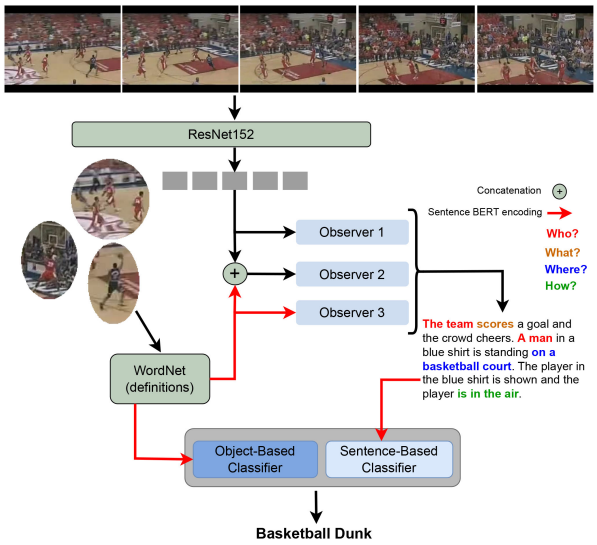

Fig. 1. Overview of the proposed method. We show the top-3 objects recognized in the video (left) and the WordNet component responsible for providing sentence definitions. We also show which features are fed to the observer models (*i.e.,* the video captioning models), and the corresponding produced sentences (right).

In this work, we improve the ZSAR performance by employing two global semantic descriptors (*i.e.,* descriptors computed over the whole video). The first is based on object-action relationships, while the second is based on sentence-actions relationships.

Fig. 1 (left) illustrates our object-based classifier, which uses a WordNet [19] encoding to provide object definitions with natural language sentences. Fig. 1 (right) shows our sentence-based classifier, a network employed to classify a set of actions using a set of soft labeled sentences (*i.e.*, annotated with a minimum human effort). In the Zero-Shot Action Recognition (ZSAR) inference step, this classifier is fed with sentences produced by video captioning methods, highlighted in Fig. 1 as *Observers 1, 2 and 3*.

In summary, our main contributions are: **(i)** we demonstrate that object definitions and paraphrase embedding can improve ZSAR models based on object-affinity. Our similarity matrices have fewer ambiguities than other methods; **(ii)** we demonstrate how textual descriptions can be used to learn a supervised action classifier based exclusively on semantic side information without hard human labeling (*i.e.,* without labeling the sentences one by one). Hence, we can generate sentences for each video (e.g., using video captioning [16]) and feed this model to predict the corresponding action class. In practice, these captioning models provide information on humans, objects, scenes, and their relationships, avoiding the need for manual definitions for affinity/prior functions on interactions while improving the performance; and, lastly, **(iii)** the predictions using objects and sentences are easily combined to reach state-of-the-art (SOTA) performance on the Kinetics-400 dataset and competitive results on UCF-101.

II. Proposed Method

Usually, the ZSAR goal is to classify samples belonging to a set of unseen action categories $Z_u = z_1, \ldots, z_{u_n}$ (*i.e.,* never seen before by the model) given a set of seen categories $Z_s = z_1, \ldots, z_{s_n}$ as the training set. The problem is named ZSAR only if $Z_u \cap Z_s = \emptyset$.

Our work is even more restrictive because we do not use a seen set $\mathcal{Z}_s$ with actions labeled for training our model; this configuration has become popular in recent years [9], [10], [11], [14], [16]. Therefore, our goal is to classify unknown classes Z_u using two types of semantic information on the videos: a textual description s and a set of objects Y. They are independent of action labels, and the ZSAR restriction is respected. Our classifier for a video v is given by

$$\mathcal{C}(v) = \arg\max_{z \in Z_u} \left(p_{sz} + \sum_{y \in Y} p_{vy} g_{yz} \right) \tag{1}$$

where $p_{sz} \forall z \in Z_u$ is the classification score of a textual description over the set of unseen classes Z_u given by a supervised model, as described in Section II-A; $p_{vy} \forall y \in Y$ are the classification scores of objects given by an off-the-shelf classifier pre-trained in the ImageNet dataset; finally, $g_{yz}\, \forall y \in Y$ and $\forall z \in Z_u$ is an affinity score, that is, a term computed to estimate which objects are most related to which actions, inspired by Jain et al. [9], but with significant improvements as described in detail in Section II-B.

A. Sentence-Based Classifier

Unlike previous works [20], [21], where synthesized features were used for training supervised models, we project a classifier based exclusively on the semantic side information. Our classifier requires a set of descriptive sentences labeled with the corresponding action class label. We adopt the sentences from Estevam et al. [16] because they collected textual descriptions from the Internet and processed them to select a set of sentences closely related to each class name. This procedure proved beneficial for classification using the nearest neighbor rule due to the sentence embedder employed [18], and can be used to soft labeling individual sentences.

Therefore, using the sentences from [16], we create a dataset $\mathcal{D} = \{S, Z_u\}$ with sentence embedding-action label pairs and compute the probability p_{sz} as

$$p_{sz} = softmax(\text{GeLU}(sW + b)), \tag{2}$$

where s is the sentence embedding given by the SBERT model outputs [18], $softmax$ returns a probability estimation on the Z_u classes, GeLU is a usual Gaussian Error Linear Unit, W is an internal weight matrix, and b is a bias vector.

B. Object-Based Classifier

First, we encode a video v by the classification scores to the $m = |Y|$ object classes from the object recognition model [22] trained on ImageNet [13].

$$\boldsymbol{p}_v = [p(y_1|v), \ldots, p(y_m|v)], \tag{3}$$

where $p(y|v)$ is computed by averaging the logits over a set of video frames at 1 FPS. Then, we estimate the probabilities with a softmax layer.

We employ a common strategy to compute the affinity between an object class y and action class z, enabling us to identify the most meaningful objects to describe an action. Then, a translation of actions $z \in Z_u$ in terms of objects $y \in Y$ is given by

$$g_{yz} = s(w(y))^T s(z)\,, \tag{4}$$

or, in other terms, $\mathbf{g}_z = [s(w(y_1)). . .(s(w(y_m)))]^T s(z)$. In our case, $w(\cdot)$ returns the WordNet definition for the object label, and $s(\cdot)$ returns the Sentence-BERT [18] encoding. This encoding does not require the Fisher vector computation on the individual words and, combined with object and sentence descriptions, conduct us to a higher performance than other object-based methods, as our results show.

C. Sparsity

We sparsify p_{sz}, p_{yz} and g_z due to the performance improvements demonstrated in [9]. Formally, we redefine the original array as

$$\hat{\boldsymbol{p}}_{\boldsymbol{v_y}} = [p(y_1, v)\delta(y_1, T_{v_y}), \ldots, p(y_m, v)\delta(y_m, T_{v_y})] \tag{5}$$

$$\hat{\boldsymbol{p}}_{\boldsymbol{s_z}} = [p(z_1, v)\delta(z_1, T_{v_z}), \ldots, p(z_n, v)\delta(z_n, T_{v_z})] \tag{6}$$

$$\hat{\boldsymbol{g}}_{\boldsymbol{z}} = [g_{zy_1}\delta(y_1, T_z), \ldots, g_{zy_m}\delta(y_m, T_z)]\,, \tag{7}$$

where $\delta(., T_{v_y})$, $\delta(., T_{v_z})$ and $\delta(., T_{z_y})$ are indicator functions, returning, 1 if class y is among the top T_{v_y} object classes in (5); returning 1 if class z is among the top T_{v_z} action classes in (6), and returning 1 if object class y is in T_{z_y} classes in (7), and 0 otherwise. T_{v_y}, T_{v_z}, and T_{z_y} are parameters.

III. Datasets and Evaluation Protocol

Our experiments were conducted on the UCF-101 [23] and Kinetics-400 [1] datasets. UCF101 is composed of 13,320 videos from 101 action classes, sampled at 25 frames per second (FPS) and with an average duration of 7.2 s. On the other hand, Kinetics-400 comprises 306,245 videos from 400 action classes with at least 400 clips per class, collected from YouTube. Each clip has a duration of 10 s. As the videos came from YouTube, we were able to download only 242,658 clips (*i.e.,* $\approx$ 80%) of the original dataset. The videos have various frame rates and resolutions.

We encode the videos using two types of semantic information: objects and sentences. For object encoding, we use the ResNet152 model from the Big Transfer (BiT) project [22] pre-trained on ImageNet considering 21,843 object classes. For sentence encoding, we retrained the Transformer-based observers [17] from Estevam et al. [16] on the ActivityNet Captions dataset [24], without any class label from ActivityNet, replacing their i3D features with our ResNet152 features. These features are sampled at each second after standardizing the videos to 25 FPS.

We evaluate our model using accuracy and following two protocols for the UCF-101 dataset: conventional and TruZe [25]. The conventional protocol consists of splitting the dataset into seen and unseen classes. However, as explained in Section II, we do not use any class from the seen set, and the evaluated configurations are 0%/50%, 0%/20%, and 0/100%. This protocol enables a fair comparison with other methods that use objects such as [9], [10], [11], [12], [14]. Due to being more restrictive, we consider that the comparison of our method with conventional methods such as [15], [26], [27], [28], [29], [30], [31] is fair. Hence, we highlight the number of training classes each model uses in each configuration.

TABLE I
Results on the UCF-101 Dataset Under Different Numbers of Test Classes

Model	UCF-101 - Testing classes			
	Train	101	50	20
Jain *et al.* [9] (ICCV)	–	30.3	–	–
Mettes and Snoek [14] (ICCV)	–	32.8	40.4 ± 1.0	51.2 ± 5.0
Mettes *et al.* [10] (IJCV)	–	36.3	47.3	61.1
Bretti and Mettes [11] (BMVC)	–	39.3	45.4 ± 3.6	–
Mishra *et al.* [20] (WACV)	51	–	22.7 ± 1.2	–
Mishra *et al.* [21] (Neurocomputing)	51	–	23.9 ± 3.0	–
Mandal *et al.* [26] (CVPR)	51	–	38.3 ± 3.0	–
Gao *et al.* [27] (AAAI)	51	–	41.6 ± 3.7	–
Kim *et al.* [28] (AAAI)	51	–	48.9 ± 5.8	–
Chen and Huang [15] (ICCV)	51	–	51.8 ± 2.9	–
Zhu *et al.* [29] (CVPR)	200	34.2	42.5 ± 0.9	–
Brattoli *et al.* [30] (CVPR)	664	39.8	48	–
Kerrigan *et al.* [31] (NeurIPS)	664	40.1	49.2	–
Ours (objects)	–	39.8	49.4 ± 4.0	60.0 ± 8.5
Ours (sentences)	–	30.8	41.1 ± 3.3	53.4 ± 6.7
Ours (objects + sentences)	–	**40.9**	**53.1 ± 3.9**	**63.7 ± 8.3**

Additionally, we evaluate our model under the TruZe protocol to provide a fair comparison with Estevam et al. [16], which is the only method using sentence descriptions generated with video captioning techniques in the ZSAR literature. In the TruZe protocol, overlapping classes between UCF-101 and Kinetics-400 are removed, enabling comparisons with methods that use 3DCNNs pre-trained on Kinetics-400.

Finally, we evaluate the performance on the Kinetics-400 dataset. We adopt the same configurations from [10] (*i.e.,* 0/25, 0/100 and 0/400 classes). When a random subset of classes is used, we perform the evaluations with 50 runs in all the protocols and datasets and report the average results.

IV. Experiments and Discussion

As shown in Table I, our complete method presented a higher performance in the UCF-101 dataset than other approaches in the literature under three split configurations. Our results are impressive compared to highly sophisticated object-based methods that explore intra-frame information such as scenes, actors, and interactions using manual defined affinity/relationship functions [10], [11]. Even our object-based classifier evaluated separately showed competitive results against 51/50 and 664/50 approaches. These results demonstrate the effectiveness of our approach and the need to include more semantic information in ZSAR methods.

Table II shows the results obtained in the UCF-101 datasets under the TruZe protocol. To enable a fair comparison, we show the results from Estevam et al. [16] and include their pre-computed sentences in our model. As expected, our sentence-based classifier, using sentences generated with ResNet152, produced results with lower accuracy than the version using

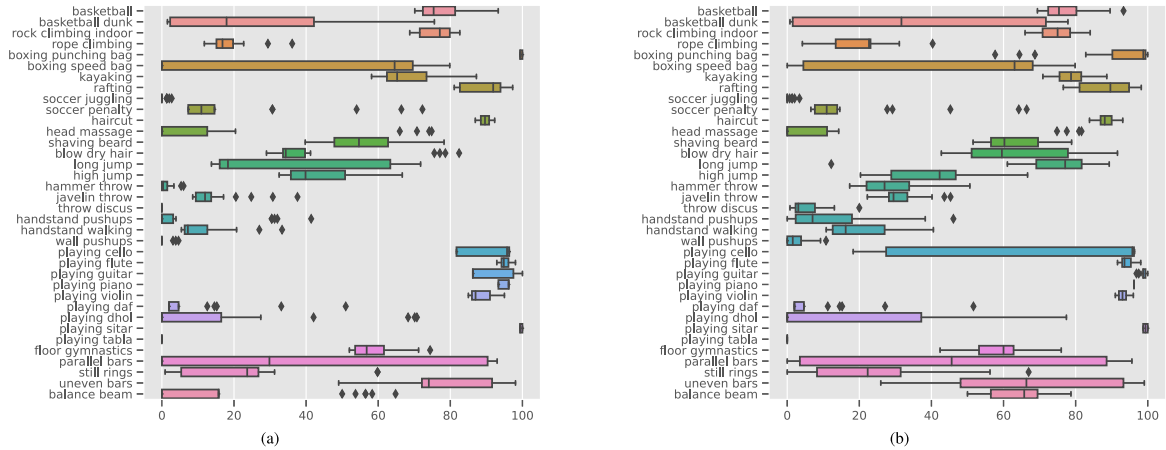

Fig. 2. *Per-class accuracy* computed over 50 random runs on the UCF101 dataset for a subset of similar semantic classes. In (a) the results are shown for the object-based model and in (b) for the complete model.

TABLE II
RESULTS ON THE UCF-101 DATASET UNDER THE TRUZE PROTOCOL (34 CLASSES FOR TESTING). TOP-2 RESULTS ARE HIGHLIGHTED

Model	UCF-101	
	Train	Accuracy (%)
Wang and Chen [32] reported by [33]	67	16.0
Mandal *et al.* [26] reported by [33]	67	23.4
Brattoli *et al.* [30] reported by [33]	664	45.2
Gowda *et al.* [33]	67	45.3
Estevam *et al.* [16]	–	49.1
Ours (objects)	–	55.3
Ours (sentences as in [16])	–	42.7
Ours (objects + sentences as in [16])	–	**60.5**
Ours (objects)	–	55.3
Ours (sentences)	–	40.1
Ours (objects + sentences)	–	**57.0**

TABLE III
RESULTS ON THE KINETICS-400 DATASET UNDER DIFFERENT NUMBERS OF TEST CLASSES. NO CLASSES WERE USED FOR TRAINING. THE BEST RESULTS ARE HIGHLIGHTED

Model	Kinetics-400 - Testing classes		
	400	100	25
Mettes and Snoek [14] (ICCV)	6.0	10.8 ± 1.0	21.8 ± 3.5
Mettes *et al.* [10] [10] (IJCV)	6.4	11.1 ± 0.8	21.9 ± 3.8
Bretti and Mettes [11] (BMVC)	9.8	18.0 ± 1.1	29.7 ± 5.0
Ours (objects)	**20.4**	32.4 ± 2.4	49.3 ± 6.8
Ours (sentences)	13.3	25.1 ± 2.2	44.2 ± 5.5
Ours (objects + sentences)	19.4	**35.1** ± **2.4**	**54.6** ± **6.1**

sentences generated with i3D. Surprisingly, this difference is only 2.7% (42.7% against 40.1%). When compared to [16], the difference to our ResNet152 version is remarkable. However, the complete model achieves considerably better results.

The Kinetics-400 dataset is very challenging for ZSAR. There are several classes semantically similar to each other (*e.g.,* eating [burger, cake, carrots, chips, doughnuts, hotdog, ice cream, spaghetti, watermelon] and juggling [balls, fire, soccer ball]). Moreover, as several methods are trained with features pre-computed in this dataset, there is not a sufficiently large list of methods with which they can be compared. In Table III, we present our results compared to [10], [11], and [14], which are object-based. As can be observed, the inclusion of semantic information in the form of natural language embedded with SBERT improves the accuracy by around 40% to 50% in all configurations. Surprisingly, the 0/400 performance for the complete model was lower than that of the object-based classifier, contrary to the results obtained in all the other experiments. We believe this occurred because the sentences produced with video captioning techniques were not sufficiently discriminative for similar actions.

Fig. 2 illustrates a similar effect in the UCF101 dataset. We compute the *per-class accuracy* for each action in each random run. Then, we produce the boxplot shown in the figure by grouping semantic similar classes. For instance, considering the classes "*basketball*" and "*basketball dunk*," they are not necessarily unknown in all runs. We observe that "*basketball dunk*" varies from 0 in some cases to around 70% in others. At the same time, "*basketball*" shows lower variation in their per-class accuracy. Hence, we conclude that the model is prone to predict "*basketball*" when both classes are unknown. The same behavior occurs between "*boxing punching bag*" and "*boxing speed bag*," and, also in other cases, as shown in the figure. For some classes (*e.g.*, "*handstand pushups*," "*handstand walking*," and "*playing dhol*"), we observe an increase in the performance shown by the increase in the bar length and a shift of the median. At the same time, "*playing cello*" presents the worst performance.

V. CONCLUSION

In this work, we introduced a new ZSAR model based on two global semantic descriptors. We demonstrated the effectiveness of adopting semantic information with sentences in natural language for both descriptors. Our supervised sentence classifier is considerably more straightforward than other supervised approaches in the literature (*e.g.,* [20], [21]) and presents a higher performance compared to them. Additionally, our object-based classifier also benefits from sentences, thus reaching remarkable results compared to other object-based methods. In future work, we intend to investigate different semantic descriptors with a focus on improving semantically similar classes, a problem that we still observe in our method.

REFERENCES

[1] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4724–4733.

[2] H. Basak, R. Kundu, P. K. Singh, M. F. Ijaz, M. Woz'niak, and R. Sarkar, "A union of deep learning and swarm-based optimization for 3D human action recognition," *Sci. Rep.*, vol. 12, 2022, Art. no. 5494.

[3] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," 2019, *arXiv:1907.06987*.

[4] V. Estevam, H. Pedrini, and D. Menotti, "Zero-shot action recognition in videos: A survey," *Neurocomputing*, vol. 439, pp. 159–175, 2021.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, vol. 2, pp. 3111–3119.

[7] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[8] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2018, pp. 3483–3487.

[9] M. Jain, J. C. van Gemert, T. Mensink, and C. G. M. Snoek, "Objects2Action: Classifying and localizing actions without any video example," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4588–4596.

[10] P. Mettes, W. Thong, and C. Snoek, "Object priors for classifying and localizing unseen actions," *Int. J. Comput. Vis.*, vol. 129, pp. 1954–1971, 2021.

[11] C. Bretti and P. Mettes, "Zero-shot action recognition from diverse object-scene compositions," in *Proc. Brit. Mach. Vis. Conf.*, 2021, pp. 1–14.

[12] P. Mettes, "Universal prototype transport for zero-shot action recognition and localization," 2022, *arXiv:2203.03971*.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[14] P. Mettes and C. G. M. Snoek, "Spatial-aware object embeddings for zero-shot localization and classification of actions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4453–4462.

[15] S. Chen and D. Huang, "Elaborative rehearsal for zero-shot action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13638–13647.

[16] V. Estevam, R. Laroca, H. Pedrini, and D. Menotti, "Tell me what you see: A zero-shot action recognition method based on natural language descriptions," 2021, *arXiv:2112.09976*.

[17] V. Estevam, R. Laroca, H. Pedrini, and D. Menotti, "Dense video captioning using unsupervised semantic information," 2021, *arXiv:2112.08455*.

[18] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 3982–3992.

[19] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.

[20] A. Mishra, V. K. Verma, M. S. K. Reddy, A. Subramaniam, P. Rai, and A. Mittal, "A generative approach to zero-shot and few-shot action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 372–380.

[21] A. Mishra, A. Pandey, and H. A. Murthy, "Zero-shot learning for action recognition using synthesized features," *Neurocomputing*, vol. 390, pp. 117–130, 2020.

[22] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient and consistent," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10925–10934.

[23] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[24] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 706–715.

[25] S. N. Gowda, L. Sevilla-Lara, K. Kim, F. Keller, and M. Rohrbach, "A new split for evaluating true zero-shot action recognition," in *Proc. German Conf. Pattern Recognit.*, 2021, pp. 191–205.

[26] D. Mandal et al., "Out-of-distribution detection for generalized zero-shot action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9985–9993.

[27] J. Gao, T. Zhang, and C. Xu, "I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 8303–8311.

[28] T. S. Kim et al., "DASZL: Dynamic action signatures for zero-shot learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1–10.

[29] Y. Zhu, Y. Long, Y. Guan, S. D. Newsam, and L. Shao, "Towards universal representation for unseen action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9436–9445.

[30] B. Brattoli, J. Tighe, F. Zhdanov, P. Perona, and K. Chalupka, "Rethinking zero-shot video classification: End-to-end training for realistic applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4613–4623.

[31] A. Kerrigan, K. Duarte, Y. Rawat, and M. Shah, "Reformulating zero-shot action recognition for multi-label actions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 25566–25577.

[32] Q. Wang and K. Chen, "Zero-shot visual recognition via bidirectional latent embedding," *Int. J. Comput. Vis.*, vol. 124, no. 3, pp. 356–383, 2017.

[33] S. N. Gowda, L. Sevilla-Lara, F. Keller, and M. Rohrbach, "CLASTER: Clustering with reinforcement learning for zero-shot action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–22.