Zero-Shot Action Recognition from Diverse Object-Scene Compositions

Carlo Bretti
bretti.carlo@gmail.com

Pascal Mettes
P.S.M.Mettes@uva.nl

University of Amsterdam
Amsterdam, the Netherlands

Abstract

This paper investigates the problem of zero-shot action recognition, in the setting where no training videos with seen actions are available. For this challenging scenario, the current leading approach is to transfer knowledge from the image domain by recognizing objects in videos using pre-trained networks, followed by a semantic matching between objects and actions. Where objects provide a local view on the content in videos, in this work we also seek to include a global view of the scene in which actions occur. We find that scenes on their own are also capable of recognizing unseen actions, albeit more marginally than objects, and a direct combination of object-based and scene-based scores degrades the action recognition performance. To get the best out of objects and scenes, we propose to construct them as a Cartesian product of all possible compositions. We outline how to determine the likelihood of object-scene compositions in videos, as well as a semantic matching from object-scene compositions to actions that enforces diversity among the most relevant compositions for each action. While simple, our composition-based approach outperforms object-based approaches and even state-of-the-art zero-shot approaches that rely on large-scale video datasets with hundreds of seen actions for training and knowledge transfer.

1 Introduction

This work seeks to recognize actions in videos without the need for any video examples to train on. Akin to zero-shot learning in the image domain [21, 53], a wide range of works have shown the ability to recognize unseen actions in videos by learning a shared embedding based on seen training actions. Examples of shared action embeddings include attributes [9, 26] and word embeddings [2, 54]. Such a transfer from seen to unseen actions requires many seen actions and videos to be effective [2, 59] and can be biased towards specific unseen actions [42]. Here, we forego the need for seen actions and directly infer unseen actions in videos from other domains.

Several works have shown that unseen actions and events can be inferred by transferring visual knowledge from images and semantic knowledge from natural language. Most notably, objects have been shown to provide strong cues about which actions [16, 33, 35, 52], and events [5, 12, 34] occur in videos. In such approaches for unseen actions, objects are recognized in videos and subsequently linked semantically to actions, implicitly enabling
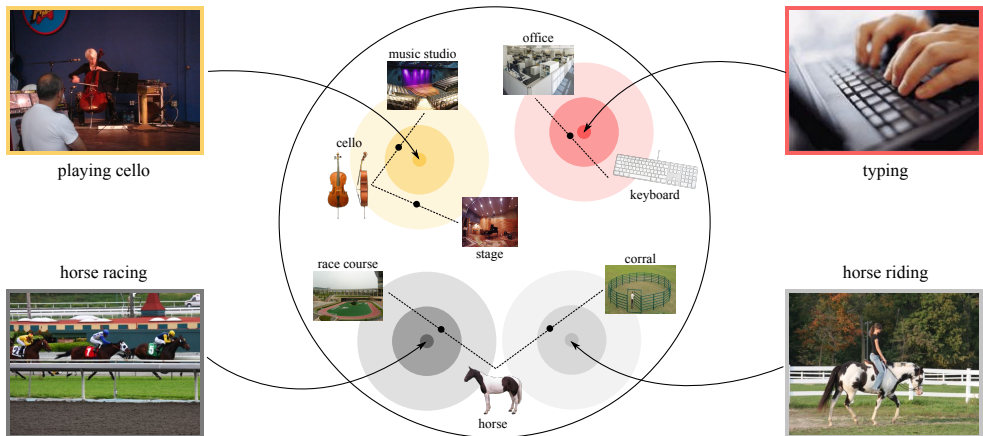
Figure 1: **Intuition behind object-scene compositions for zero-shot action recognition.**
Objects and scenes provide complementary views and modelling them as compositions can
enrich actions like in *playing cello* or *typing* actions, or even distinguish actions that are
identical when only considering objects, such as the case for *horse racing* and *horse riding*.

action inference [16]. Where objects provide a local video view, the larger context about
the scene in which actions and objects occur is ignored. In order to take both perspectives
into account, we propose a new way to recognize actions directly from objects and scenes
simultaneously by modeling them as compositions.

The main contribution of this work is the introduction of object-scene compositions for
zero-shot action recognition. We interpret objects and scenes not as independent entities,
but as a Cartesian product of all possible combinations. Figure 1 illustrates the idea behind
object-scene compositions, where compositions get the best out of both knowledge sources
to infer unseen actions. We furthermore describe a selection of most relevant object-scene
compositions that takes the diversity amongst the objects and scenes into account. Exper-
imental evaluation on UCF-101 and Kinetics shows the effectiveness of our composition-
based approach, outperforming both object-based approaches for zero-shot action recogni-
tion and approaches that rely on hundreds of seen actions during training, despite the sim-
plicity of the approach.

2   Related work

Supervised action recognition. Human action recognition is a long-standing challenge
in computer vision. Foundational approaches focused on the design of invariant features
based on, for example, spatio-temporal interest points [43], cuboids [19], and dense trajec-
tories [49]. Action recognition has in recent years been accelerated with advances in deep
learning for videos, such as two-stream networks [44], 3D convolutional networks [4, 18],
and transformer-based networks [10]. Deep video networks have been fueled by new large-
scale videos datasets with action annotations [20, 58, 45, 57]. While these advances make it
possible to recognize up to hundreds of actions, each action requires many examples to train
on. Moreover, supervised approaches do not generalize to new and unseen actions. Here, we
investigate how unseen actions can still be recognized from objects and scenes.

Several supervised action approaches have previously shown the potential of knowledge about objects and scenes. Objects have shown to be effective intermediate representations for supervised action recognition [17, 23] and video event detection [28, 31, 32]. Similarly, scenes and context information have also aided actions and events [14, 25]. Other works investigating a combination of objects and scenes have proved their effectiveness [15, 27, 50]. We also seek to use objects and scenes, however we do so in the unseen setting, and we show that understanding videos as compositions of objects and scenes is fruitful.

Zero-shot action recognition. Early approaches to zero-shot action recognition followed initial success in image-based zero-shot recognition [21] by transferring knowledge from seen to unseen actions using attributes [9, 26, 56]. In attribute-based approaches, videos are mapped to an attribute space, which contains attributes such as *being outdoors* and *moving arms up and down*. This mapping is learned on training videos with seen actions and used during testing as an intermediate representation to generalize to unseen actions. Follow-up work advocated a direct mapping from seen to unseen actions using a shared semantic space based on word embeddings [1, 24, 48, 54], using knowledge about test distributions [55], feature synthesis [57], or through hierarchical knowledge [29]. The current state-of-the-art seeks to construct a universal representation of actions [59] by training a network to predict word embeddings on hundreds of seen actions [2]. Common among these approaches is the need for training videos with seen actions to enable a knowledge transfer to unseen actions. In this work, we seek to infer actions without the need for any seen action to train on.

In the setting without any videos to train on, Jain *et al*. [16] introduced an approach to classify actions based on object recognition and a semantic transfer from objects to actions. The underlying assumption of such an approach is that visually related categories are also semantically related. A similar setup has been proposed in zero-shot event detection in videos, see *e.g.* [5, 6, 12, 34]. An object-based setup has also been proposed for action localization by using localized object detections and prior knowledge about spatial relations between persons and objects [53]. Most recently, object-based recognition of unseen actions has been extended by incorporating priors about multiple languages, object discrimination, and a bias towards specific object names [35]. In this work, we follow the strict setting without any videos to train on. We build upon previous work by considering scenes in addition to objects and model these as compositions to perform action recognition. Wu *et al*. [52] have previously used objects and scenes for zero-shot action recognition. Their approach requires training a network to fuse object, scene, and generic features from a video to predict a seen action. From a seen action prediction, the most closely related unseen action is used as the final inference. Our composition-based approach requires no training on video examples and can directly predict any unseen action. A few works have recently investigated zero-shot compositional learning in images [30, 39, 40]. Where these works seek to recognize unseen object-state compositions, we propose to use object-scene compositions to recognize unseen actions in videos.

3 Unseen actions from object-scene compositions

For our problem, we are given a set of unlabelled videos V and a set of action labels A. We seek to find a transformation $f(v|O, S) \in A$ that assigns an action label to unseen video v using two sources of prior knowledge, namely objects O and scenes S. To that end, we first define objects and scenes as dense compositions. We then outline how to obtain zero-shot action

inference from object-scene compositions and how to diversify object-scene compositions for improved performance. Figure 1 highlights the main idea, where objects and scenes are paired and used to infer which actions occur in videos through a semantic transfer from object-scene compositions to actions.

3.1 Composition construction

We main idea of the paper is that when using both objects and scenes to infer unseen actions, we need to consider them as compositions. We start by formally defining the compositions:

Definition 1. (Object-scene compositions). *Given two sets of recognition functions with corresponding semantic labels, O for objects and S for scenes, the object-scene compositions are defined as the Carthesian product over both sets $C = O \times S$. Each composition $c \in C$ is accompanied with a semantic representation $\phi(c)$ and a likelihood estimate $p(c|\cdot)$.*

By defining objects and scenes as a Cartesian product of its pairs, we arrive at a large set of entities for which to derive semantic representations and likelihood estimates. To keep the inference tractable, we define both as linear combinations of the corresponding object and scene origin. For a composition $c = (c_o, c_s) \in C$ consisting of one object and one scene, the semantic embedding vector is given as the summed representation over the object and scene,

$$\phi(c) = \phi(c_o) + \phi(c_s), \tag{1}$$

where $\phi(\cdot) \in \mathbb{R}^{300}$ denotes the semantic embedding function, *e.g.* a word2vec representation [36] or sentence transformers [51]. To be able to perform unseen action recognition, we additionally need to be able to estimate the likelihood of each composition in a video and a semantic similarity from composition to action. For the likelihood estimation, let $\mathcal{F}_O$ and $\mathcal{F}_S$ denote deep networks pre-trained on objects O and scenes S. Then the likelihood for video v is given as:

$$p(c|v) = p(c_o|v; \mathcal{F}_O) \cdot p(c_s|v; \mathcal{F}_S). \tag{2}$$

The semantic similarity between an object-scene composition c and an action a is given by a direct cosine similarity from their embedding [36]:

$$s(c, a) = \cos(\phi(c), \phi(a)). \tag{3}$$

3.2 Unseen action inference

To assign an action label to an unseen video v, we follow object-based action literature [16, 53] and first determine the top k most relevant compositions for each action based on Equation 3. Let C_a^k denote the set with the k most semantically similar compositions for action a. Then the score for action a in video v is given as:

$$\ell(a, v) = \sum_{c' \in C_a^k} s(c', a) \cdot p(c'|v). \tag{4}$$

Finally, the action prediction is given as:

$$f(v) = \arg\max_{a \in A} \ell(a, v). \tag{5}$$

Similar to object-based approaches, the action inference is a relatively straightforward combination of likelihood estimation and semantic transfer. The key to our approach is that we use both objects and scenes and we consider them as compositions, which provides empirically stronger results than performing likelihood estimation and semantic transfer for objects and scenes separately.

3.3 Diverse selection of top compositions

For unseen action recognition, a standard approach to determine the top compositions C_a for action a is to select the individual compositions with the highest semantic similarity. An issue for compositions specifically is that this leads to low diversity in either the objects or scenes in the top compositions. Most notably, an object or scene that is highly semantically similar to an action will likely be over-represented in the top compositions, which leads to redundancy. We therefore seek to obtain a selection of top compositions that incorporates joint object and scene diversity.

To obtain diverse top compositions, we take inspiration from information retrieval literature [7], specifically Maximum Marginal Relevance [3]. For action a, we start by selecting the most semantically relevant composition, $i.e.$ $C_a^\star = \{\arg\max_{c \in C} s(c,a)\}$. We then iteratively add the composition that is most semantically similar to action a and least semantically similar to the compositions in $C_a^\star$. Hyperparameter λ controls the relative weight of semantic similarity and diversity, where $\lambda = 1$ denotes the baseline setting where only semantic similarity to the action is taken into account. The update rule for Maximum Marginal Relevance in this context is given as:

$$C_a^\star := C_a^\star \cup \underset{c' \in C \backslash C_a^\star}{\arg\max} \left[\lambda \cdot s(c',a) - (1-\lambda) \cdot \max_{c'' \in C_a^\star} s(c',c'') \right]. \qquad (6)$$

We perform the above iterative update rule to populate $C_a^\star$ until it contains k compositions.

4 Experimental setup

4.1 Datasets

UCF-101 is composed of 13,320 video clips with 101 actions labels such as *band marching*, *ice dancing*, or *pizza tossing*. The labels cover a wide range of actions, from sports to playing musical instruments and human-object interaction [46].

Kinetics-400 consists of 104,000 YouTube videos clips covering 400 action labels. The action labels range from individual actions such as *drawing* to human-human actions such as *shaking hands* and human-object interactions such as *washing dishes* [20].

4.2 Implementation details

Object and scene likelihoods. For the object network $\mathcal{F}_o$, we follow [35] and employ a GoogLeNet [47], pre-trained on ImageNet [8] using the 12,988 object labels as per [54]. For the scene network $\mathcal{F}_s$, we use a ResNet50 [13], pre-trained on the Places365 dataset [58]. For each video, we sample two frames per second and feed them to both networks. For both objects and scenes, we average their respective softmax output over all frames in the video.
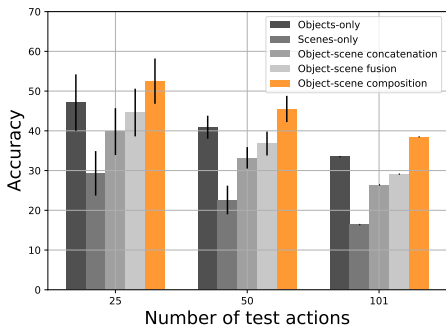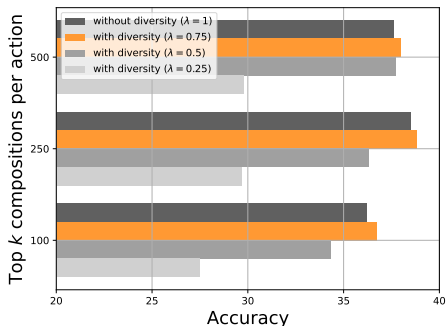
(a) Ablation I.   (b) Ablation II.

Figure 2: **Ablation studies I and II** on UCF-101. In (a), we evaluate the importance of modeling objects and scenes as compositions for zero-shot action recognition. Individually, objects are better able to differentiate actions than scenes, and standard combinations results in lower performance. Modeling objects and scenes as compositions on the other hand improves the performance, highlighting their potential. In (b) we show the effect of diverse composition selection when using all 101 actions for testing. Across all thresholds for top compositions per action, adding diversity provides a modest but consistent improvement.

Semantic Embeddings. We investigate two semantic embeddings for zero-shot action recognition. The first is a FastText embedding [11] to allow for a direct comparison to the recent object-based approach of Mettes *et al*. [35]. We additionally investigate the potential of sentence embeddings based on the SentenceTransformers framework [41]. Specifically, we use the pre-trained MiniLM [51] to generate the label embeddings.
**Code.** The code is available here.

5 Experimental results

For the experimental evaluation, we first perform a series of ablation studies and qualitative analyses on the proposed object-scene compositions. We then perform a comparative evaluation to the current state-of-the-art for zero-shot action recognition.

5.1 Ablation studies on object-scene compositions

We perform the following ablation studies, to investigate: (I) the importance of compositionality when combining objects and scenes, (II) the effect of diversity in composition selection per action, (III) the effect of sentence embeddings in the semantic matching between actions and object-scene compositions, and (IV) a qualitative analysis.

Ablation I: Importance of object-scene compositionality. As a starting point, we investigate the composition-based approach for combining objects and scenes. We draw a comparison to four baselines, namely one where only objects are considered, one where only scenes are considered, one where objects and scene probabilities are concatenated in a single feature, and one using a late fusion of objects and scenes by averaging their respective action scores. For the baselines, we use the top 100 objects per action as advocated
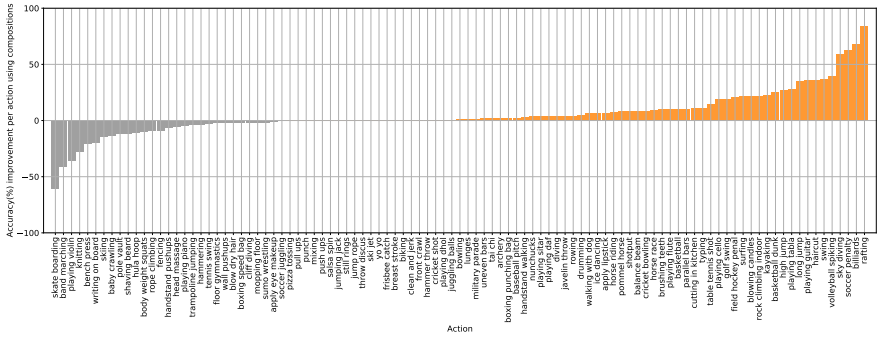
(a) Improvements compared to the object-only baseline.
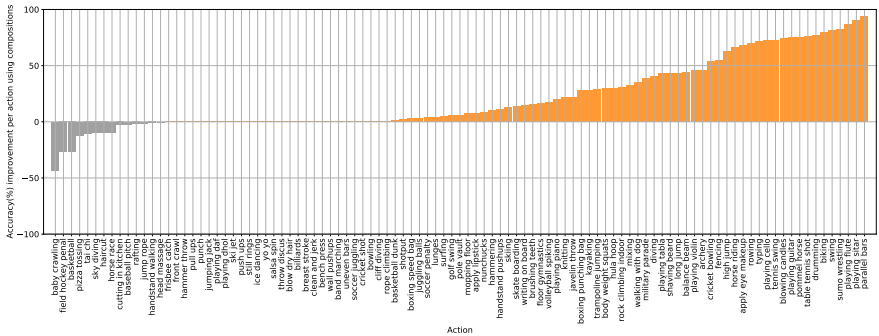
(b) Improvements compared to the scene-only baseline.

Figure 3: **Per-action improvements for our composition-based approach on UCF-101.**
In (a) we show improvements compared to the object-based standard. The overall improve-
ments are a result of large gains for scene-oriented actions (*e.g. soccer penalty*), while scene-
independent actions do not benefit from object-scene compositions (*e.g. skateboarding*). In
(b) we show improvements compared to the scene-only baseline. The overall improvements
are a result of large gains for object-oriented actions (*e.g. playing sitar* or *playing flute*).

in [16, 33] and the top 5 scenes per action and top 100 objects/scenes for the concatenation
baseline, which work well empirically. For our approach, we use the top 250 object-scene
compositions without diversity. All use FastText embeddings for semantic matching.

The results for this ablation study are shown in Figure 2a on UCF-101 for 25, 50, and 101
test actions. For 25 and 50 actions, we perform 10 trials where we randomly select actions
for each trial and we report the mean and standard deviation over all trials. Independently,
objects perform better than scenes in all three settings. The results for scenes are however
far from random, indicating that there is a lot of signal in scenes for actions. Both the
concatenation and late fusion of objects and scenes fail to get the best of objects and scenes.
The scene-based scoring in fact has a negative effect on the object-based approach. When
considering objects and scenes as compositions, however, we observe improvements across
all three settings. With 101 actions for testing, the performance increases from 33.5% for
the object-based baseline to 38.5% for our composition-based approach. We conclude that
viewing objects and scenes as compositions is preferred for zero-shot action recognition.

To better understand the importance of objects and scenes as compositions for zero-shot
action recognition, we have performed two additional analyses. First, we considered objects

	UCF-101			
	50 classes		101 classes	
	FastText	MiniLM	FastText	MiniLM
Objects-only	40.9 ± 2.9	43.6 ± 3.5	33.5	36.5
Scenes-only	22.6 ± 3.6	20.9 ± 2.8	16.4	15.6
Object-scene concatenation	33.2 ± 2.7	37.8 ± 2.2	26.4	30.4
Object-scene fusion	36.8 ± 3.0	38.2 ± 3.0	29.1	31.1
Object-scene compositions	$\mathbf{45.4 \pm 3.6}$	$\mathbf{45.2 \pm 4.6}$	$\mathbf{38.8}$	$\mathbf{39.3}$

Table 1: **Ablation III: Effect of sentence embeddings** on UCF-101. For all but the scene-based baseline, sentence embeddings have a positive effect on the overall performance.

with other objects as compositions. We made two random subsets with respectively 12,623 and 365 objects from ImageNet. On UCF-101 with 101 test actions, this approach obtains an accuracy of 32.2% compared to 33.5% for the object-only baseline, highlighting the importance of making compositions from diverse sources, here objects and scenes. Second, we consider an extension of our model in which a weight for each composition is included as an additional guide for selection or scoring, similar in spirit to an object prior [55]. The weight is defined as the cosine similarity between the object and scene embeddings. When including the weights before computing the top selection of compositions we obtain an accuracy of 30.3%. When the weights are instead included during the scoring (*i.e.* in Eq.4), we obtain an accuracy of 37.7% compared to 38.5% for the composition-based approach. We conclude that add a composition-based prior in the form of a weight is ineffective. Compositions themselves already contain the desired information from their textual sources.

In Figure 3a, we show per-action improvement for all the 101 actions in UCF-101 compared to the object-only baseline. The largest improvement is obtained on *rafting*, where we see an 83.7% improvement from 4.5% accuracy using the object-only baseline to 88.2% using compositions. This is not surprising, since rafting occurs in a specific scene only, namely rivers with currents. On the other hand, using compositions degrades performance on *knitting* from 47% for the object-only baseline to 22%. This is likely due to the scene-independent nature of the action considered, as the action can take place in many scenes. Hence the stronger the link to both scenes and objects, the better the action is recognized.

Figure 3b shows the improvement brought upon by using compositions rather than scenes alone. Two of the largest improvements registered are with regard to *playing sitar* and *playing flute*, with a 90.4% and 86.5% improvement respectively. This is likely due to the fact that for such actions, objects are particularly relevant and convey important information that scenes alone cannot capture. Although large improvements across most of the actions can be obtained by using compositions, for actions such as *baby crawling*, *field hockey penalty*, and *basketball* it appears that incorporating knowledge from objects degrades performance. This is likely to be explained by a noisier prediction of object likelihoods.

Ablation II: Effect of diverse composition selection. In the second ablation, we investigate the potential effect of diversifying the selection of most relevant object-scene compositions per action. We perform this ablation on UCF-101 using all 101 actions for testing with FastText for the semantic embeddings. For the diverse composition selection, we test different values of λ as shown in Figure 2b. We find that the best performance is obtained by including some notion of diversity (*i.e.* with $\lambda = 0.75$) compared to including no diversity at all ($\lambda = 1$) or to preferring diversity over relevance ($\lambda = 0.25$). For different selections of top k compositions per action, diversity provides small but consistent improvements. Through-

Top objects	Top scenes	Top objects	Top scenes	Top objects	Top scenes
climbing frame	playground	wing	sky	skateboard	desert vegetation
swing	lawn	dive	ocean	roller skates	home theater
trampoline	yard	snowboard	aquarium	skater	golf course

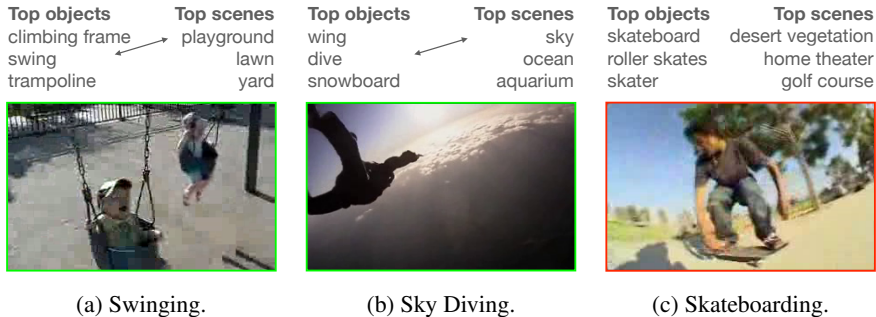(a) Swinging.   (b) Sky Diving.   (c) Skateboarding.

Figure 4: **Qualitative results for object-scene compositions** on UCF-101. *Swinging* and *sky diving* have ambiguous relations to objects and scenes by themselves, while their top composition (arrows) provides a strong and disambiguated relation to their action. A composition-based approach is not beneficial when the object or scene are independent of the action, as is the case for the *skateboarding* example.

out the rest of the paper, we will therefore incorporate diverse composition selection.

Ablation III: Effect of sentence embeddings for semantic matching. Sentence embeddings have recently been shown to be beneficial for zero-shot recognition in the image domain [22]. Here, we investigate their potential in the video domain. We perform a comparison between the word embeddings from FastText [11] and the sentence embeddings from MiniLM [51] on UCF-101 using 50 and 101 test actions. The results are shown in Table 1. For all approaches except the scene-only baseline, using sentence embeddings provides a direct improvement. On 101 actions, the object-only baseline improves from 33.5% to 36.5%, while our composition-based approach improves from 38.8% to 39.3%. We conclude that sentence embeddings are also beneficial for matching objects and compositions to actions.

Ablation IV: Qualitative analysis. Figure 4 shows three example videos from UCF-101. For Figure 4a, we can see that compositions are an effective tool, as *swinging* can effectively be understood as a composition of the object *swing* used in a certain context, *i.e.* a *playground*. Indeed, *swinging* directly benefits from compositions, as per Figure 3. In Figure 4b, *dive* and *sky* are deemed a top object-scene composition for the action *sky diving*. For this action, objects and scenes in isolation are ambiguous, while their composition removes this ambiguity. In Figure 4c, we also show a failure case for *skateboarding*, which is object-dependent but scene-independent.

5.2 Comparative evaluation

We perform a comparison to the current state-of-the-art in zero-shot action recognition on UCF-101 in the most challenging setting, namely using all 101 actions for testing. We also perform a comparison on Kinetics, which has recently been performed in [35].

Comparison on UCF-101. In Table 2, we compare our approach to the current state-of-the-art in zero-shot action recognition. The most direct comparison is to object-based action recognition approaches [16, 33, 34], which do not use any training videos with seen actions. Where Mettes *et al.* [35] obtain an accuracy of 36.3%, we improve the results to 38.8% with the same word embeddings and further boost the accuracy to 39.3% with sentence embeddings. We also include a comparison to Zhu *et al.* [59] and Brattoli *et al.* [2], which perform zero-shot action recognition by training on hundreds of seen actions and performing

	UCF-101		
	nr. actions		accuracy
	train	test	
Zhu *et al.* [59] CVPR'18	200	101	34.2
Brattoli *et al.* [2] CVPR'20	664	101	37.6
Jain *et al.* [16] ICCV'15	-	101	30.3
Mettes and Snoek [33] ICCV'17	-	101	32.8
Mettes *et al.* [35] IJCV'21	-	101	36.3
This paper w/ word embeddings	-	101	**38.8**
This paper w/ sentence embeddings	-	101	**39.3**

Table 2: **Comparative evaluation on UCF-101.** Both compared to the object-based state-of-the-art [35] and compared to the overall best performing approach that uses 664 seen Kinetics actions for training [2], we obtain higher accuracies. We obtain state-of-the-art results with word embeddings and obtain further improvements with sentence embeddings.

	Kinetics		
	Number of test actions		
	25	100	400
Mettes *et al.* [33] as in [35]	21.8 ± 3.5	10.8 ± 1.0	6.0
Mettes *et al.* [35]	22.0 ± 3.7	11.2 ± 1.0	6.4
Objects-only	18.3 ± 2.5	10.5 ± 0.9	5.8
Scenes-only	28.2 ± 6.0	14.2 ± 1.2	7.3
This paper	$\mathbf{29.7 \pm 5.0}$	$\mathbf{18.0 \pm 1.1}$	**9.4**

Table 3: **Comparative evaluation on Kinetics.** In the Kinetics dataset, scenes are more informative for actions than objects, as indicated by the higher classification performance. The composition between objects and scenes provides the best overall results.

a semantic transfer from seen to unseen actions. Compare to the state-of-the-art approach by Brattoli *et al.* [2], we improve the overall accuracy from 37.6% to 39.3% highlighting the effectiveness of object-scene compositions for zero-shot action recognition.

   Comparison on Kinetics. We also perform a comparative evaluation on Kinetics, based on the zero-shot protocol recently outlined in [35]. The object-only baseline follows an approach comparable to that of Jain *et al.* [16]. We find scenes to be slightly more informative than objects. Modeling objects and scenes as compositions remains directly beneficial. Where Mettes *et al.* [35] obtain a mean accuracy of 11.2% with 100 actions, our mean accuracy is 18.0%. On the full 400-way classification, our approach improves the scores from 6.4% to 9.4%, restating the effectiveness of our composition-based approach.

6   Conclusions

This work advocates objects and scenes as compositions for zero-shot action recognition. Where the current standard is to transfer knowledge from objects when no training videos are available, we propose a Cartesian product of all objects and scenes as basis for inferring unseen actions. Due to the homogeneous nature of such a set, we find that actions benefit from diversity during composition selection. Experimentally, we show that object-scene compositions provide a simple yet effective approach for zero-shot action recognition, especially in settings where relying on objects or scenes is ambiguous. Comparisons on UCF-101 and Kinetics show that our composition-based approach outperforms the current object-based standard and even recent works that rely on training on hundreds of seen actions.

References

[1] Ioannis Alexiou, Tao Xiang, and Shaogang Gong. Exploring synonyms as context in zero-shot action recognition. In *ICIP*, 2016.

[2] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *CVPR*, 2020.

[3] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

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

[5] Xiaojun Chang, Yi Yang, Alexander Hauptmann, Eric P Xing, and Yao-Liang Yu. Semantic concept discovery for large-scale zero-shot event detection. In *IJCAI*, 2015.

[6] Xiaojun Chang, Yi Yang, Guodong Long, Chengqi Zhang, and Alexander Hauptmann. Dynamic concept composition for zero-example event detection. In *AAAI*, 2016.

[7] Gobinda G Chowdhury. *Introduction to modern information retrieval*. Facet publishing, 2010.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning Attributes Equals Multi-Source Domain Generalization. In *CVPR*, 2016.

[10] Rohit Girdhar, Joao Joao Carreira, Carl Doersch, and Andrew Zisserman. Video Action Transformer Network. In *CVPR*, 2019.

[11] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning Word Vectors for 157 Languages. In *LREC*, 2018.

[12] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, 2014.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[14] Fabian Caba Heilbron, Ali Thabet, Juan Carlos Niebles, and Bernard Ghanem. Camera Motion and Surrounding Scene Appearance as Context for Action Recognition. In *ACCV*, 2015.

[15] Nazli Ikizler-Cinbis and Stan Sclaroff. Object, Scene and Actions: Combining Multiple Features for Human Action Recognition. In *ECCV*, 2010.

[16] Mihir Jain, Jan C. van Gemert, Thomas Mensink, and Cees G. M. Snoek. Objects2action: Classifying and Localizing Actions without Any Video Example. In *ICCV*, 2015.

[17] Mihir Jain, Jan C. van Gemert, and Cees G. M. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, 2015.

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D Convolutional Neural Networks for Human Action Recognition. *TPAMI*, 2013.

[19] Jingen Liu, Saad Ali, and Mubarak Shah. Recognizing human actions using multiple features. In *CVPR*, 2008.

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. arXiv:1705.06950.

[21] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *TPAMI*, 2014.

[22] Yannick Le Cacheux, Hervé Le Borgne, and Michel Crucianu. Using sentences as semantic representations in large scale zero-shot learning. In *ECCV*, 2020.

[23] Li-Jia Li, Hao Su, Yongwhan Lim, and Li Fei-Fei. Object bank: An object-level image representation for high-level visual recognition. *IJCV*, 2014.

[24] Yikang Li, Sheng-hung Hu, and Baoxin Li. Recognizing unseen actions in a domain-adapted embedding space. In *ICIP*, 2016.

[25] Cheng-Chang Lien, Chiu-Lung Chiang, and Chang-Hsing Lee. Scene-based event detection for baseball videos. *JVCIR*, 2007.

[26] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[27] Jingen Liu, Qian Yu, Omar Javed, Saad Ali, Amir Tamrakar, Ajay Divakaran, Hui Cheng, and Harpreet Sawhney. Video event recognition using concept attributes. In *WACV*, 2013.

[28] Jingen Liu, Qian Yu, Omar Javed, Saad Ali, Amir Tamrakar, Ajay Divakaran, Hui Cheng, and Harpreet Sawhney. Video event recognition using concept attributes. In *WACV*, 2013.

[29] Teng Long, Pascal Mettes, Heng Tao Shen, and Cees G M Snoek. Searching for actions on the hyperbole. In *CVPR*, 2020.

[30] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, 2021.

[31] Masoud Mazloom, Efstratios Gavves, and Cees GM Snoek. Conceptlets: Selective semantics for classifying video events. *TMM*, 2014.

[32] Michele Merler, Bert Huang, Lexing Xie, Gang Hua, and Apostol Natsev. Semantic model vectors for complex video event recognition. *TMM*, 2011.

[33] Pascal Mettes and Cees G. M. Snoek. Spatial-Aware Object Embeddings for Zero-Shot Localization and Classification of Actions. In *ICCV*, 2017.

[34] Pascal Mettes, Dennis C. Koelma, and Cees G. M. Snoek. Shuffled ImageNet Banks for Video Event Detection and Search. *TOMM*, 2020.

[35] Pascal Mettes, William Thong, and Cees G. M. Snoek. Object Priors for Classifying and Localizing Unseen Actions. *IJCV*, 2021.

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013.

[37] Ashish Mishra, Anubha Pandey, and Hema A Murthy. Zero-shot learning for action recognition using synthesized features. *Neurocomputing*, 2020.

[38] Mathew Monfort, Carl Vondrick, Aude Oliva, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, and Dan Gutfreund. Moments in Time Dataset: One Million Videos for Event Understanding. *TPAMI*, 2020.

[39] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *CVPR*, 2021.

[40] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *ICCV*, 2019.

[41] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019.

[42] Alina Roitberg, Manuel Martinez, Monica Haurilet, and Rainer Stiefelhagen. Towards a Fair Evaluation of Zero-Shot Action Recognition Using External Data. In *ECCV*, 2019.

[43] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004.

[44] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *NIPS*, 2014.

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

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

[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[48] Yi Tian, Qiuqi Ruan, and Gaoyun An. Zero-shot action recognition via empirical maximum mean discrepancy. In *ICSP*, 2018.

[49] Heng Wang and Cordelia Schmid. Action Recognition with Improved Trajectories. In *ICCV*, 2013.

[50] Ruiqi Wang and Xinxiao Wu. Combining multiple deep cues for action recognition. *MTAP*, 2019.

[51] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *NeurIPS*, 2020.

[52] Zuxuan Wu, Yanwei Fu, Yu-Gang Jiang, and Leonid Sigal. Harnessing Object and Scene Semantics for Large-Scale Video Understanding. In *CVPR*, 2016.

[53] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018.

[54] Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *ICIP*, 2015.

[55] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *IJCV*, 2017.

[56] Zhong Zhang, Chunheng Wang, Baihua Xiao, Wen Zhou, and Shuang Liu. Robust relative attributes for human action recognition. *PAA*, 2015.

[57] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. In *ICCV*, 2019.

[58] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *TPAMI*, 2018.

[59] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *CVPR*, 2018.