KG-SP: Knowledge Guided Simple Primitives for Open World Compositional Zero-Shot Learning

Shyamgopal Karthik[1] Massimiliano Mancini[1] Zeynep Akata[1,2]
[1]University of Tübingen [2]Max Planck Institute for Intelligent Systems

Abstract

The goal of open-world compositional zero-shot learning (OW-CZSL) is to recognize compositions of state and objects in images, given only a subset of them during training and no prior on the unseen compositions. In this setting, models operate on a huge output space, containing all possible state-object compositions. While previous works tackle the problem by learning embeddings for the compositions jointly, here we revisit a simple CZSL baseline and predict the primitives, i.e. states and objects, independently. To ensure that the model develops primitive-specific features, we equip the state and object classifiers with separate, non-linear feature extractors. Moreover, we estimate the feasibility of each composition through external knowledge, using this prior to remove unfeasible compositions from the output space. Finally, we propose a new setting, i.e. CZSL under partial supervision (pCZSL), where either only objects or state labels are available during training, and we can use our prior to estimate the missing labels. Our model, Knowledge-Guided Simple Primitives (KG-SP), achieves state of the art in both OW-CZSL and pCZSL, surpassing most recent competitors even when coupled with semi-supervised learning techniques. Code available at: https://github.com/ExplainableML/KG-SP.

1. Introduction

As humans, we interact with objects depending on their state. For instance, we use ripe lemons rather than moldy ones to prepare a lemonade, and we clean dirty dishes after using them. Algorithms that can recognize objects together with their state are crucial for autonomous agents to show the same high-level interactions capabilities we have. In the literature, this problem is studied under the name of *Compositional Zero-shot Learning* (CZSL). In CZSL, we are given a training set with images of objects in a subset of their possible states and, at test time, the goal is to recognize compositions of the same set of objects and states, even unseen during training. Since an object has a different appearance depending on its state (*e.g. dry dog* vs *wet dog*) and a state modifies objects in different ways (*e.g. wet dog* vs *wet car*), the challenge of CZSL is modeling how states and objects interact with each other, extrapolating this knowledge from seen to unseen compositions. Under this perspective, multiple works modeled the interactions of objects and states, either through compositional classifiers [26, 31], or a shared embedding space [23, 28, 29].

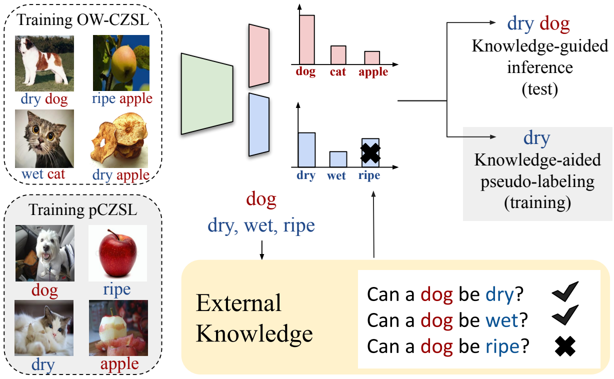

Figure 1. We consider the problems of open-world CZSL (OW-CZSL), where we lack priors on unseen compositions at test time, and CZSL under partial supervision (pCZSL) where we also lack compositional labels during training (left). We tackle them by independently predicting object (red) and state (blue) labels and by using external knowledge (bottom) to estimate the feasibility of compositions, reducing the search space during inference and improving pseudo-labeling during training in pCZSL.

Despite their effectiveness, [24] showed how the performance of CZSL methods degrade in the open-world setting (OW-CZSL). In OW-CZSL, there are no priors on the unseen compositions, and models must consider all possible compositions at test time. Due to the large cardinality of the output space, it is difficult to produce discriminative embeddings for the unseen compositions [24]. Inspired by the findings of [24], in this work we explore a completely different direction. Specifically, we design an architecture that disregards the compositional nature of the problem and produces the initial predictions independently for objects and states. The idea is that while discriminating between compositions is hard in OW-CZSL due to the large search space, recognizing primitives (*i.e.* objects and states) in isolation is easier since

1) the cardinality of the two sets is much lower and 2) the sets are fixed at both training and test time.

Inspired from [26] and [19], we design a simple method that predicts objects and states with two independent classifiers. Since recognizing states requires different features w.r.t. recognizing objects, instead of having a shared feature representation, we train our model with two different non-linear feature extractors. Furthermore, since not all compositions are equally feasible in reality (*e.g. ripe dog*) we can refine the predictions of our model by eliminating less feasible compositions from the output space. With this goal, we use external knowledge (*i.e.* ConceptNet [39]) to estimate the compatibility between a state and an object, using these estimates to remove less feasible compositions at test time. We name our model Knowledge-Guided Simple Primitives (KG-SP). As our KG-SP method does not require compositional labels during training, we explore a new challenging setting, *i.e.* CZSL under partial supervision (pCZSL). In pCZSL, training samples have either only object or state annotation, but not both. Here we use our prior on feasible compositions to aid pseudo-labeling during training. Experiments show that KG-SP is either competitive or surpasses the current state of the art in OW-CZSL and outperforms recent CZSL approaches on pCZSL setting. Figure 1 provides an overview of KG-SP and the two tasks.

Contributions. To summarize, 1) inspired by [19, 26], our model predicts state and objects independently while at the same time removing less feasible compositions from the output space based on external contextual information about the feasibility of certain compositions; 2) we explore the problem of CZSL under partial supervision, where either object or state information is missing in the ground-truth; 3) we adapt recent baselines for pCZSL showing that KG-SP outperforms them even when coupled with semi-supervised learning techniques in both OW-CZSL and pCZSL settings.

2. Related works

Compositional zero-shot learning aims to recognize compositions of states and objects in images, even unseen during training. The main challenge of this setting is modeling how states modify objects, generalizing this capability to unseen compositions. Most of the previous works focused on how to model the interactions between states and objects either at the parameter level or in a given representation space. For instance, [26, 31] proposed to generate a classifier for a given state-object composition given two classifiers (or embeddings) for specific state and object primitives, using either a compositional module [26] or a gating network [31]. Differently, [23, 29] model each state as an operator transforming object embeddings, imposing properties on the state operators (*e.g.* commutativity, symmetry). In [29] the state operators are linear, while in [23] they are coupling and decoupling networks. Recently, [28, 36] used graph convolutional networks [17] to model the interactions between state, objects and their compositions. Differently, [2] tackles CZSL from a causality perspective, learning disentangled objects and states representations. In this work, we revisit VisProd [26], predicting objects and states in isolation, showing that this strategy is effective in OW-CZSL. As in [24], we estimate the feasibility of each composition to improve the model's performance. However, we use ConceptNet to this aim, rather than compositional annotations, being the first to tackle CZSL without compositional labels during training.

Multi-task learning. Since we predict state and object independently, our work is related to Multi-Task [7, 15, 27, 35, 37] and Multi-Domain learning [4, 22, 32, 33], where the goal is to learn a unique model able to address different visual tasks. Most of the approaches in this domain either learn task-specific parameters [4, 22, 32, 33, 37] and how to combine them [27, 35], or focus on re-weighting different loss functions [15]. While we use multi-task learning to design primitive classifiers for CZSL, our final goal is different since we compose predictions from separate output spaces.

Learning from Partial Supervision. Our CZSL setting without compositional labels is related to semi-supervised learning and learning with missing labels. In semi-supervised learning, both labeled and unlabeled samples are available, and the goal is to effectively use the unlabeled samples. Popular ideas revolve around consistency regularization [5, 6, 38], and self-training [10, 20, 34]. Unlike semi-supervised learning, in pCZSL, all samples are labeled, but partially. Thus, we can also exploit the prior on how objects interact with states to estimate the missing labels.

For what concerns learning with missing labels, this is most prevalent in multi-label scenarios where it is unfeasible to annotate all labels that are present in a single image. Approaches in this field usually model the correlation among labels [8, 9, 12, 18] to impose semantic objectives on missing ones. While we are also interested in learning from partial supervision, our labels lie on two separate spaces (*i.e.* objects and states), and the missing labels (*e.g.* state) influences the appearance of the positive one (*e.g.* object). In this setting, the main challenge is to model how the two spaces influence each other without any compositional supervision.

3. Knowledge Guided Simple Primitives

Problem formulation. CZSL [24] aims to recognize compositions of a set of objects O and a set of states S. Formally, we are given a training set $\mathcal{T} = \{(x, y)\}_{i=1}^{N}$, where N is the size of the training set, $x \in X$ denotes an image in the input space X and $y \in Y_s$ is its label in the set of seen compositions Y_s. The goal is to learn a model that can recognize a set of compositions $Y_t = Y_s \cup Y_u$, where Y_u is a set of unseen compositions (*i.e.* $Y_u \cap Y_s = \emptyset$) and $Y_t \subseteq Y$, with Y being the set of all possible compositions, *i.e.* $Y = S \times O$.

OW-CZSL and pCZSL settings. In this work, we consider two different CZSL settings. Open-World CZSL (OW-CZSL) [24] assumes no prior on the set of unseen compositions at test time. This means that the model needs to operate on the full compositional space, *i.e.* $Y_t = Y$. Consequently, the number of unseen compositions is much larger than the number of seen ones *i.e.* $Y_u = Y \setminus Y_s$, thus the main challenge is operating in a very large output space where most of the compositions are unseen and thus hard to discriminate.

In this work, we also consider a new challenging task, namely CZSL under partial supervision (pCZSL), where the training set does not contain any compositional label and all training images have either object or state label, but not both. This setting is more realistic than standard CZSL since most datasets are collected with single labels (*e.g.* only object-level information) and collecting multiple labels is expensive and time-consuming. Formally, we consider the labels of our training set $\mathcal{T}$ to be of the form $y = (s, \mathtt{u}) \vee y = (\mathtt{u}, o),\ \forall (x, y) \in \mathcal{T}$, with $s \in S$, $o \in O$ and $\mathtt{u}$ denoting an unknown label. Note that, as a consequence of this formulation, the set of training compositions Y_s is not known a priori anymore. This implies that, as in OW-CZSL, we need to consider the full compositional space at test time, *i.e.* $Y_t = Y$. Moreover, since no training image contains both object and state labels, we do not have explicit supervision on how states modify objects and vice-versa.

In the following, we describe the two components of our framework, Simple Primitives (SP) where we predict the primitives, e.g. object and states, independently and Knowledge Guidance (KG) where we use external resources that guide our model on the feasibility of certain compositions.

3.1. Simple Primitives (SP) in KG-SP

Inspired by the early Visual Product (VisProd) baseline [26], our model completely disregards the compositional nature of the problem and predicts states and objects independently. This idea contrasts with recent approaches (*i.e.* [2,23,24,28,29,31]), explicitly modeling the interactions between objects and states within the model.

Formally, given an image x, we extract its feature representation $z = \omega(x)$ through a function ω, mapping images into a feature space Z, *i.e.* $\omega : X \rightarrow Z$. We then have an object classifier $\phi_o : Z \rightarrow \Delta_O$ that maps z to a vector in the probability simplex Δ_O, spanning all object categories. Similarly, we have another classifier that maps z to a probability over the states, *i.e.* $\phi_s : Z \rightarrow \Delta_S$. During training, we minimize the cross-entropy loss for both the object and state predictions. Specifically, we minimize:

$$\begin{aligned}\mathcal{L}_{\text{visprod}} &= \sum_{i=1}^{N} \mathbb{I}_{s_i \neq \mathtt{u}} \mathcal{L}_{\text{state}}(x_i, s_i) + \mathbb{I}_{o_i \neq \mathtt{u}} \mathcal{L}_{\text{obj}}(x_i, o_i) \qquad (1)\\ &= -\sum_{i=1}^{N} \mathbb{I}_{s_i \neq \mathtt{u}} \log \phi_s(z_i, s_i) + \mathbb{I}_{o_i \neq \mathtt{u}} \log \phi_o(z_i, o_i)\end{aligned}$$

where $z_i = \omega(x_i)$, $\phi_o(z, o)$ is the probability of the object o assigned by ϕ_o to the input z, and $\phi_s(z, s)$ is the probability of the state s assigned by ϕ_s to the input z. In Eq. (1), $\mathbb{I}$ is an indicator function used to not compute the loss in pCZSL, in the absence of primitive labels. Our prediction function is:

$$f = \underset{(s,o) \in Y}{\arg\max}\ \phi_o(w(x), o) \cdot \phi_s(w(x), s). \qquad (2)$$

Although learning simple primitives independently like this may not be effective in standard CZSL, we argue that the capability to separate state and objects predictions is crucial in OW-CZSL, where the search space is too large if predictions are made over the full compositional space.

In the original VisProd formulation, objects and state predictors are simple linear layers operating over the same feature vector. However, this choice is suboptimal and leads to low results in practice. In fact, by using separate linear layers, VisProd addresses CZSL as a multi-task learning problem [7, 15, 27, 32, 33, 35, 37], where there are two different tasks (*i.e.* states and objects prediction) that share the same feature extractor while differing only for the classification head. However, multiple works in multi-task (MTL) and multi-domain learning (MDL) discussed how fully sharing the parameters to extract the feature representation for different tasks (*i.e. hard-sharing* [7]) is suboptimal when the tasks are not strictly related [27, 32, 33] and may even lead to negative transfer [21].

In CZSL, recognizing objects is different than recognizing their states. Specifically, the former requires focusing on global features: for instance, distinguishing an animal from another requires focusing on their shapes and skins while distinguishing fruits requires detecting texture-based cues. On the contrary, recognizing states requires focusing on local patterns: for instance, the difference between *dry* and *wet* can be detected by the presence of drops in *cars* and *apples* while in animals it requires looking at the shape of the fur. With this premise, we need to overcome the limits imposed by hard-parameter sharing to ensure the objects and state classifiers have enough flexibility to learn primitive-specific features. While advanced MTL and MDL techniques can be used for this purpose, in this work we found that it is sufficient to implement the two classifiers as multi-layer perceptrons (MLP) with non-linear activations.

3.2. Knowledge Guidance (KG) in KG-SP

In the large output space of OW-CZSL and pCZSL, not all compositions are equally feasible (*e.g. ripe dog*, *hairy apple*) and taking this prior into account can help in correcting incompatible state-object predictions of our model. In the following we describe how we estimate the feasibility scores and how we use them in our model.

Estimating feasibility scores. Formally, let us associate to each composition (s, o) a compatibility score $c_s^o \in [0, 1]$.

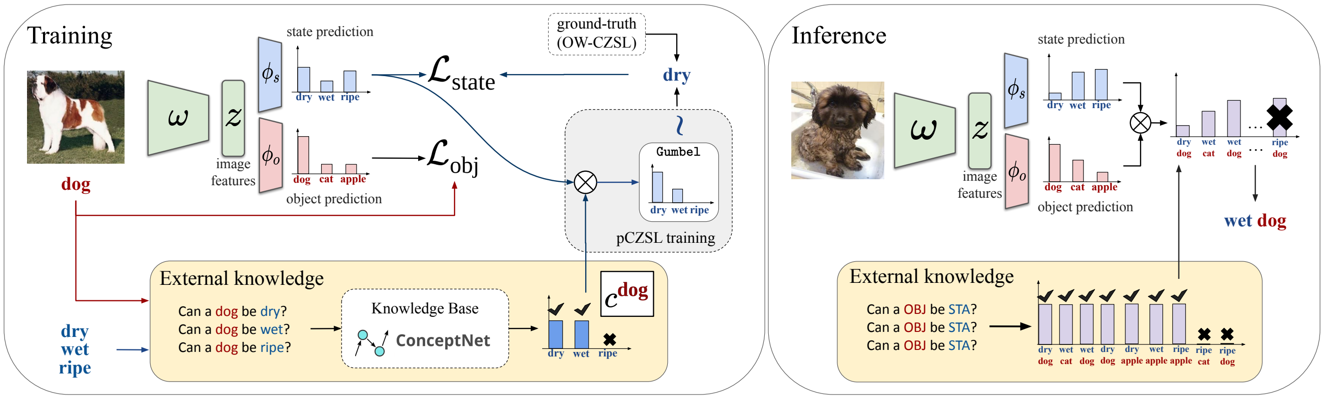

Figure 2. **Knowledge-Guided Simple Primitives (KG-SP)**. We train a separate object (red) and state (blue) predictor on top of a shared feature extractor (green) using the available state and object labels. We use external knowledge to estimate feasibility scores of compositions (yellow, bottom), using this prior during inference to directly remove unfeasible compositions from the output space. In pCZSL , we use this knowledge to re-weight the class scores and perform pseudo-labeling (grey) of missing labels, sampling them through the Gumbel-softmax.

Since there exist no database contain such information, previous works exploited the set of seen compositions Y_s to estimate c_s^o [24]. Here we explore an alternative direction by using external knowledge. In this way, our estimation is independent of the actual availability of the set Y_s and can also be applied in pCZSL, where Y_s is unknown. While we explored different strategies (see supplementary), we found ConceptNet [39] to give reliable feasibility estimates.

ConceptNet is a knowledge graph connecting words and phrases with labeled edges, extracted from various sources [39]. We can use ConceptNet in two ways. The first is querying for the existence of a composition and the second is querying for the relatedness between two entries (*i.e.* object and state). Since direct queries are very sparse, we follow the second approach, defining the scores as:

$$c_s^o = \rho_{\text{KB}}(s, o) \tag{3}$$

where $\rho_{\text{KB}}(s, o)$ returns the relational score between s and o. In ConceptNet, these scores are computed from the cosine-similarity of *ConceptNet Numberbatch* embeddings [40]. The latter are built from ConceptNet adjacency matrix and existing word embeddings (e.g. word2vec [25], GloVe [30]).

Using the feasibility scores during inference. Similarly to [24], the most straightforward-way to use the feasibility sores is to remove from the output space less feasible compositions during inference. Thus, our prediction function becomes:

$$f_{\text{KG}} = \underset{(s,o)\in Y, c_s^o > 0}{\arg\max} \quad \phi_o(z, o) \cdot \phi_s(z, s) \tag{4}$$

where we consider feasible all compositions with $c_o^s > 0$.

Using the feasibility scores during training for pCZSL. In pCZSL, we may obtain additional supervision by estimating missing labels. One straightforward way to achieve this is through pseudo-labeling [20], a semi-supervised learning technique that takes the model predictions as ground-truth for unlabeled samples. In pCZSL, this means that, when the state (object) label is missing, pseudo-labeling will impute as label the object (state) predicted with the highest score. To avoid that the pseudo-labels form unfeasible compositions, we can use our prior to aid the pseudo-labeling process.

Given either an object label o or a state label s, we estimate their respective state and object pseudo-labels as:

$$\hat{s} \sim \texttt{Gumbel}\left(\phi_s(z) \odot c^o\right), \quad \hat{o} \sim \texttt{Gumbel}\left(\phi_o(z) \odot c_s\right) \tag{5}$$

where c^o is the vector containing the compatibility scores for all states given the object o, *i.e.* $c^o = [c_s^o]_{s\in S}$, and c_s is its counterpart for all objects given the state s, *i.e.* $c_s = [c_s^o]_{o\in O}$[1]. Note that in both equations we sample the pseudo-labels using Gumbel-softmax (`Gumbel`) [14]. We found this choice helpful to make the model more robust to noisy predictions and less biased toward the training set latent label distribution. Our objective function becomes:

$$\begin{aligned} \mathcal{L}_{\text{visprod}}^{\text{pCZSL}} = &\sum_{(x_s, s)\in\mathcal{T}_s} \mathcal{L}_{\text{state}}(x_s, s) + \mathcal{L}_{\text{obj}}(x_s, \hat{o}) \\ &+ \sum_{(x_o, o)\in\mathcal{T}_o} \mathcal{L}_{\text{obj}}(x_o, o) + \mathcal{L}_{\text{state}}(x_o, \hat{s}). \end{aligned} \tag{6}$$

We use this objective in pCZSL during training and Eq.(1) for OW-CZSL. In both cases we perform inference through Eq.(4). Since we couple independent primitive prediction with external knowlede to refine them, we name the method **K**nowledge-guided **S**imple **P**rimitives (KG-SP). Figure 2 illustrates our approach during training and inference.

4. Experiments

Datasets. We use the three standard datasets for Compositional Zero-Shot Learning, namely UT-Zappos [43,44], MIT-States [13] and the recently proposed C-GQA [28] dataset.

[1] We assume S and O to be alphabetically ordered.

Method	MIT-States				UT Zappos				C-GQA			
	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
TMN [31]	12.6	0.9	1.2	0.1	55.9	18.1	21.7	8.4	NA	NA	NA	NA
AoP [29]	16.6	5.7	4.7	0.7	50.9	34.2	29.4	13.7	NA	NA	NA	NA
LE+ [26]	14.2	2.5	2.7	0.3	60.4	36.5	30.5	16.3	19.2	0.7	1.0	0.08
VisProd [26]	20.9	5.8	5.6	0.7	54.6	42.8	36.9	19.7	24.8	1.7	2.8	0.33
SymNet [23]	21.4	7.0	5.8	0.8	53.3	44.6	34.5	18.5	26.7	2.2	3.3	0.43
CompCos$^{\text{CW}}$ [24]	25.3	5.5	5.9	0.9	59.8	45.6	36.3	20.8	28.0	1.0	1.6	0.20
CGE$_{\text{ff}}$ [28]	29.6	4.0	4.9	0.7	58.8	46.5	38.0	21.5	28.3	1.3	2.2	0.30
CompCos [24]	25.4	**10.0**	**8.9**	**1.6**	59.3	46.8	36.9	21.3	28.4	1.8	2.8	0.39
CGE [28]	**32.4**	5.1	6.0	1.0	61.7	47.7	39.0	23.1	**32.7**	1.8	2.9	0.47
KG-SP$_{\text{ff}}$	23.4	7.0	6.7	1.0	58.0	47.2	39.1	22.9	26.6	2.1	3.4	0.44
KG-SP	28.4	7.5	7.4	1.3	**61.8**	**52.1**	**42.3**	**26.5**	31.5	**2.9**	**4.7**	**0.78**

Table 1. **Open World CZSL results** on MIT-States, UT Zappos and C-GQA. We measure best seen (S) and unseen accuracy (U), best harmonic mean (HM), and area under the curve (AUC) on the compositions. KG-SP$_{\text{ff}}$ refers to our proposed method with a frozen backbone.

UT-Zappos contains 12 object categories (shoe types) and 16 state categories (material types), with 83 seen compositions and a compositonal space of 192 compositons. **MIT States** is a larger dataset that contains 245 object categories in 115 possible states. In total, it contains $1,262$ seen compositions and an output space of $28,175$ compositions in OW-CZSL. Finally, **C-GQA** is a recently proposed dataset[2] with 674 object categories and 413 state categories. It contains $5,592$ training compositions and a full compositional space with $278,362$ compositions.

Baselines. In OW-CZSL, we compare KG-SP against standard CZSL approaches; namely Attributes as Operators (AoP) [29], Label Embed+ (LE+) [26], Task Modular Networks (TMN) [31], SymNet [23], Compositional Graph Embeddings (CGE) [28] and Compositional Cosine Logits (CompCos) [24]. In the tables, we refer to the closed world version of CompCos as CompCos$^{\text{CW}}$ and the variant of CGE with a frozen feature extractor as CGE$_{\text{ff}}$.

In pCZSL, we compare KG-SP with CGE [28] and CompCos [24], the state-of-the-art models in the closed and open-world settings respectively. These methods are adapted to pCZSL by marginalizing their predictions over states/objects, when the state/object information is available, minimizing the cross-entropy loss on the ground-truth annotation. We also experiment with popular semi-supervised learning techniques such as entropy minimization [10] and pseudo-labeling [20], adding them to both CompCos and CGE.

Evaluation Protocol. For the OW-CZSL setting, we follow the standard splits of [28, 31], evaluate all the methods on the generalized setting, where the model recognizes samples from both seen and unseen compositions. Following the protocol in [31], we apply a bias on the seen compositions at test time, measuring the performance as best seen (S) and best unseen (U) accuracy, best harmonic mean (HM) as well as the area under the curve (AUC) by varying the bias.

For the pCZSL setting, we propose a new split of the training split set, separating samples with object and state labels. This is done by keeping only the object label for half the samples, while for the remaining half, only the state label, ensuring that every object and state is seen in the training set.Furthermore, for pCZSL setting, the model has no access to seen compositons Y_s. Thus, we evaluate the model in the full compositional space, without subtracting any bias on Y_s. Therefore, we use as metric the seen (S) and unseen (U) accuracy, and their harmonic mean (HM), as it is standard in Zero-Shot Learning [42].

Implementation Details. We follow the standard practices in the CZSL literature [24, 28], by using a ResNet18 [11] feature extractor. For the state and object classifiers, we follow [28] and use Multi-Layer Perceptrons with three layers, comprising Layer Normalization [3] and Dropout [41]. The model is optimized using Adam [16] with the default hyperparameters, a learning rate and weight-decay of $5e\text{-}5$.

4.1. Open-World CZSL

The results on the challenging OW-CZSL setting are reported in Table 1. In this setting KG-SP either outperforms or it is competitive with the state of the art. Specifically, on UT-Zappos KG-SP outperforms the best competitor (CGE) in all metrics, with an improvement of 3.4% in AUC (26.5 vs 23.1), 3.3% in best HM (42.3 vs 39.0) and 4.4% best unseen (52.1 vs 47.7). Similarly, without end-to-end training, KG-SP$_{\text{ff}}$ surpasses the best baseline (CGE$_{\text{ff}}$) by 1.3 in AUC (22.9 vs 21.5) and by 1.3 in best HM (39.1 vs 38.0).

These results are confirmed in the challenging C-GQA dataset. Despite an output space of almost 280k compositions, KG-SP obtains 0.78 AUC vs 0.47 of the best competitor (CGE), 4.7 HM (vs 3.3 of SymNet) and 2.9 on best unseen (vs 2.2 of SymNet). When non-finetuned, the method achieves competitive results w.r.t. SymNet (*e.g.* 3.4 HM) while being much easier to optimize, since KG-SP does not impose any constraint on the compositional space. Our re-

[2] We refer to the updated split in https://github.com/ExplainableML/czsl.

	Marginaliz.	Seen	Unseen	HM	AUC
CGE$_{\text{ff}}$		**25.5**	5.7	6.5	1.0
	✓	24.0	**7.8**	**8.1**	**1.3**
CGE		**27.2**	6.6	7.0	1.3
	✓	25.1	**8.1**	**8.1**	**1.4**

Table 2. OW-CZSL results in the validation set of MIT-States when using marginalization. CGE$_{\text{ff}}$ is the approach of [28] with frozen backbone whereas CGE performs end-to-end training.

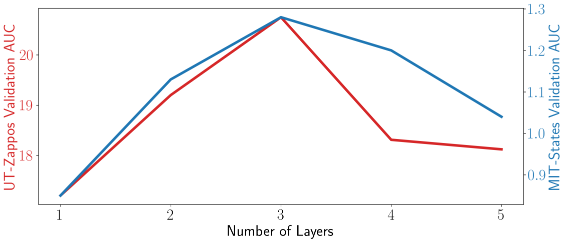

Figure 3. Ablation study on the importance of the depth of the object and state-classifiers for KG-SP$_{\text{ff}}$ on UT-Zappos (red curve) and MIT-States (blue curve) validation set for the OW-CZSL setting. Performance is measured in AUC.

	Mask	Seen	Unseen	HM	AUC
VisProd		**24.8**	6.8	7.3	1.1
	✓	24.7	**7.2**	**7.6**	**1.2**
KG-SP$_{\text{ff}}$		26.3	7.4	7.9	1.3
	✓	**26.5**	**7.7**	**8.2**	**1.4**

Table 3. OW-CZSL results on MIT states validation set while applying our feasibility-based masks (f_{KG}) on different models.

sults indicate that modeling states and objects independently may be an effective approach to deal with the very large output space of OW-CZSL. This independence assumption ensures that each predictor learns a discriminative classifier over a few hundred classes rather than a single classifier over thousands of compositions, which does not scale well even with powerful architectures (*e.g*. graph-convolutional neural network of [28]) and initialization through side-information (*e.g*. word-embeddings [25]).

Finally, the table also highlights the gap between KG-SP and the VisProd baseline of [26]. In particular, our revised model (without fine-tuning) consistently surpasses VisProd in AUC (*i.e*. 1.0 vs 0.7 on MIT-States, 22.8 vs 19.7 in Zappos, 0.44 vs 0.33 in C-GQA) and best harmonic mean (6.7 vs 5.6 on MIT-States, 39.3 vs 36.9 on Zappos and 3.4 vs 2.8 on C-GQA). These results confirm the importance of our modifications to the original VisProd model, as we will ablate in the following.

4.1.1 Why KG-SP works in OW-CZSL?

Separately predicting objects and states. We argue that an important reason for the success of KG-SP is the separate treatment of states and objects, As stated in previous discussion, predicting state and objects independently makes the OW-CZSL problem easier and more scalable w.r.t. predicting directly over thousands of compositions. To verify this hypothesis, we take the state-of-the-art method in standard CZSL, CGE [28] (with and without end-to-end training) and we modify its classifier in such a way that it can output separate objects and states. Specifically, we take the state predictions by marginalizing their scores across all possible objects and, similarly, we marginalize object predictions over the set of possible states. Results are reported in Table 2. We see a consistent increase in best unseen accuracy (5.7 vs 7.8 for CGE$_{\text{ff}}$, 6.6 vs 8.1 for CGE) and in the best HM (6.5 vs 8.1 for CGE$_{\text{ff}}$, 7.0 vs 8.1 for CGE) when we separate the two predictions. As a consequence, the methods also improve in AUC (1.1 to 1.2 CGE$_{\text{ff}}$, 1.3 to 1.4 for CGE). This indicates how providing a separate ground for objects and states predictions is a useful strategy in the open-world setting. Operating in the primitives rather than the compositional space, provides a simplification of the problem that can improve the performance even of existing state-of-the-art CZSL models.

Effect of depth of the classifier on KG-SP. We ablate the impact of the depth in Fig 3 for KG-SP$_{\text{ff}}$ on both MIT-States and UT-Zappos validation sets. The validation AUC on both UT-Zappos (red curve) and MIT-States (blue curve) rapidly increases with the depth of the classifier. This shows the importance of taking into account the multi-task nature of the problem and instantiate objects and state classifiers that have enough capacity to extract primitive-specific features. While deeper predictors help, after 3 layers the performance degrades (*i.e*., going from 26.9 to 24.3 on UT-Zappos when depth is increased from 3 to 5 layers). The reason behind this drop is mainly linked to overfitting on seen compositions.

Effect of the knowledge-based masking. We ablate the impact of masking out unfeasible compositions from the output space in Table 3 for MIT-states validation set. We experiment with both VisProd and KG-SP$_{\text{ff}}$. As the table shows, KG-SP$_{\text{ff}}$ benefits from removing unfeasible compositions, with the unseen accuracy going from 7.3 to 7.5 and the best HM from 7.6 to 8.1. Similarly, also for VisProd we see a consistent improvement for both the unseen accuracy (6.8 to 7.2) and the best HM (7.3 to 7.6) when its output space is filtered. These results show how removing unfeasible compositions from the output space benefit the performance of OW-CZSL models and that ConceptNet is a reliable source for estimating feasibility scores of the compositions.

4.2. CZSL under Partial supervision

The results on our proposed pCZSL setting are reported in Table 4. In addition to recognizing unseen compositions in an extremely large compositional space, in pCZSL the model has to cope with the lack of compositional labels.

In this scenario KG-SP achieves state-of-the-art results on all the three datasets. On UT-Zappos, KG-SP achieves a HM of 13.1 vs 10.7 of the best competitor (CGE). Similarly, on

Method		MIT-States			UT Zappos			C-GQA		
		S	U	HM	S	U	HM	S	U	HM
CGE$_{\text{ff}}$ [28]		19.6	1.3	2.4	50.3	3.4	5.0	17.4	0.4	0.9
	+Pseudo-Lab.	**19.7**	0.9	1.8	48.5	1.1	2.2	19.8	0.2	0.4
	+Entropy Min.	15.1	1.7	3.1	51.9	4.2	6.4	22.1	0.4	0.9
CompCos [24]		10.8	2.0	3.6	52.4	4.1	7.6	24.3	0.4	0.7
	+Pseudo-Lab.	9.2	1.9	3.2	52.9	3.7	6.8	23.6	0.3	0.5
	+Entropy Min.	13.2	2.1	3.7	55.0	4.2	7.9	23.1	0.6	1.1
CGE [28]		17.9	1.6	3.0	55.8	5.9	10.7	25.6	0.7	1.4
	+Pseudo-Lab.	10.6	2.3	3.8	56.1	3.9	7.3	21.3	0.6	1.2
	+Entropy Min.	17.8	1.6	3.0	60.1	4.7	8.7	24.8	1.0	1.8
KG-SP$_{\text{ff}}$		13.5	**2.6**	**4.4**	53.8	6.9	12.3	22.3	0.9	1.7
KG-SP		18.4	2.2	4.0	**57.9**	**7.4**	**13.1**	**26.9**	**1.2**	**2.3**

Table 4. **Partial Open-World CZSL results** on MIT-States, UT Zappos and C-GQA. We measure seen (S) and unseen accuracy (U) on the compositions and their harmonic mean (HM). KG-SP refers to our full model with our knowledge-guided pseudo-labeling and inference. CGE$_{\text{ff}}$ and KG-SP$_{\text{ff}}$ denotes the non-finetuned version of the methods. For each CZSL baseline, we show the results of the original methods and the same coupled with Entropy-Minimization (Entropy Min.) [10] or Pseudo-labeling (Pseudo-Lab.) [20].

Method	Seen	Unseen	HM
KG-SP	**16.6**	2.8	4.8
+ Pseudo-Labeling	15.9	2.7	4.6
+ Gumbel Softmax	16.1	2.6	4.5
+ ConceptNet-scores	**16.6**	**3.1**	**5.3**

Table 5. **Partial Open-World CZSL results** on MIT-States validation set for different methods in terms of Seen, and Unseen accuracy and their Harmonic Mean (HM). Pseudo-Labeling, sampling the pseudo-label using Gumbel-Softmax, and using ConceptNet to filter unfeasible labels are added one after another.

MIT-States and C-GQA, the results obtained by KG-SP exceed the performance of the competitors, with KG-SP achieving 4.4 HM vs 3.7 of CompCos with entropy minimization on MIT-States, and 2.3 (1.7 when non end-to-end trained) HM on C-GQA vs 1.8 of the best competitor (CGE). It is interesting to note how, even incorporating semi-supervised learning techniques in CGE and CompCos does not bridge the gap between KG-SP and these methods. Interestingly, semi-supervised learning techniques do not bring consistent improvements across CZSL models and tasks. Entropy minimization, by pushing the method toward confident predictions, improves CGE (1.8 vs 1.4 AUC) and CompCos (1.1 vs 0.7) on C-GQA, and improves CGE$_{\text{ff}}$ on MIT-States (3.1 vs 2.4 AUC). However, in the other settings either achieves the same performances of the baseline (*e.g*. CGE on MIT-States) or degrades them (*e.g*. CGE on UT-Zappos, 8.7 vs 10.7 AUC). On the other hand, pseudo-labeling degrades the performance of the CZSL models in all settings (*e.g*. from 3.6 to 3.2 AUC of CompCos on MIT-States), providing advantages only for CGE on MIT-States (3.8 vs 3.0 AUC). The low efficacy of standard semi-supervised learning techniques, is due to the fact that the top predicted object/state may form an unfeasible composition with the ground-truth state/object. Without modeling the feasibility of state-object compositions, pseudo-labeling may assign incorrect (and unfeasible) labels while entropy minimization may push the model toward incorrect predictions. In the following section, we discuss why this happens and why it is important to restore to our knowledge-aided pseudo-labeling in pCZSL.

4.2.1 Ablation Study

We ablate the two important components of our proposed approach: the pseudo-labeling strategy and the quality of the ConceptNet feasibility scores.

Effect of the pseudo-labeling strategy. In Table 5, we consider a few alternatives to our proposed pseudo-labeling strategy for KG-SP. The first option is to just introduce vanilla pseudo-labeling. This is done by replacing the missing labels with the top predicted class, when the confidence of the model is greater than a threshold. This strategy has a negative effect, reducing the accuracy of the model on both seen and unseen compositions (*i.e*. 16.6 vs 15.9 on the seen, 2.8 vs 2.7 on the unseen). This is because pseudo-labeling alone is prone to confirmation bias [1] since the model can become increasingly more confident about the predictions that, without modeling the feasibility of compositions, may be not only incorrect, but also unfeasible.

The second alternative is to use Gumbel-softmax [14], sampling a label instead of performing direct pseudo-labeling. This approach may allow the model to overcome the confirmation bias, especially in the initial training stages by sampling different labels w.r.t. the top prediction. However, this strategy alone does not achieve good results, degrading the performance of the base model (*e.g*. 4.5 vs 4.8 HM) and performing slightly worse than standard pseudo-labeling (4.5 vs 4.6 HM). This performance degradation is due in both cases by not exploiting the external knowledge to correct the pseudo-labeling process, avoiding using

	Compositions	
	Most Feasible (Top-5)	Least Feasible (Bottom-5)
	painted paint	grimy balloon
	muddy mud	steaming bracelet
	frozen ice	blunt clock
	mossy moss	thin garage
	cloudy cloud	unpainted belt
Objects	**States**	
	Most Feasible (Top-3)	Least Feasible (Bottom-3)
chains	frayed, broken, loose	pureed, unripe, steaming
sugar	melted, whipped, caramelized	scratched, ancient, coiled
sword	blunt, sharp, splintered	filled, runny, closed
laptop	small, shattered, modern	cloudy, sunny, dull
tulip	bright, wilted, ruffled	grimy, raw, damp

Table 6. Examples of ConceptNet feasibility scores. Top: Top-5 (left) and Least-5 (right) compositions per feasibility; Bottom: Top-3 highest and Bottom-3 lowest feasible states per random objects.

incompatible compositions as supervision.

Our pseudo-labeling strategy, on the other hand, provides a clear improvement (5.3 vs 4.8 HM) over our method which already attains state-of-the-art results on open-world CZSL benchmarks. Its benefits come from using external knowledge to assign a feasibility score to each composition. These feasibility scores are used to refine the pseudo-labeling process, avoiding incompatible compositional labels. In addition to this, using the model scores to sample the label in a probabilistic manner ensures that confirmation bias and model drift can be avoided.

Quality of the ConceptNet feasibility scores. An important aspect of KG-SP is the quality of the estimated feasibility scores. In Table 6 we show some qualitative results of our strategy, reporting detailed quantitative analyses on the supplementary material. ConceptNet assigns the highest feasibility scores to compositions where the state and the objects share the same root, such as *painted paint*, *muddy mud*, and *mossy moss*. This is a consequence of the very similar contexts in which such states and objects appear. On the other hand, among the least feasible compositions we find objects with incompatible physical transformations (*i.e. steaming bracelet*), or unclear states (*e.g. unpainted belt*, *blunt clock*). These scores reflect the reliance of ConceptNet relatedness scores to the context in which words appear. This is also a limitation, since rare co-occurrences can be deemed as unfeasible compositions (*e.g. grimy balloon*, *thin garage*).

When inspecting the most and least feasible compositions for random objects (Table 6, bottom), we find that the compositions ranked as most feasible such as *bright tulip*, *caramelized sugar*, and *blunt sword* are indeed feasible in the reality, while the compositions marked as least feasible are not. Interestingly, the unfeasible compositions merge objects and states from different categories. Examples are food vs tools (*e.g. pureed chains*, *scratched sugar*) and weather vs objects (*e.g. cloudy laptop*). This suggests that the ConcepNet-based feasibility scores encode a high-level notion of grouping, where objects and states can be separated depending on the contexts where they commonly occur.

5. Conclusion

In this work, we addressed the problem of Open-World Compositional Zero-Shot learning (OW-CZSL), where the goal is to recognize compositions of objects and states in images given only a subset of them during training and without any prior on unseen compositions at test time. We address the problem by revisiting a simple CZSL method, VisProd, that independently predicts state and object labels. The idea is to simplify the problem by exploiting the much smaller cardinality of object and state sets w.r.t. the compositional labels. Our model, KG-SP, uses two different feature extractors to account for the dissimilarity between the two tasks, and uses external information to remove less feasible compositions from the output space at test time. As KG-SP does not require compositional labels, we explore a new challenging setting, CZSL under partial supervision (pCZSL) where training images have either only object or state annotation. In pCZSL, we use the feasibility scores to aid the estimation of the missing labels. Experiments show that KG-SP achieves the state of the art in OW-CZSL and pCZSL, outperforming recent CZSL approaches coupled with standard semi-supervised learning techniques.

Limitations and Broader Impact. One weakness of our approach, shared with CZSL methods, is that the absolute performance on all OW-CZSL benchmarks is quite low (*e.g.* 2.9 unseen accuracy on C-GQA). This can significantly hamper the real-world deployment of these models. However, we believe that this research topic is crucial to bridge the gap between machine and human visual compositional recognition abilities: our work contributes to the field by making another step toward this direction. Another limitation of our work is that automatic retrieving feasible state-object compositions is a process without expert supervision, thus vulnerable to inaccuracies in the knowledge base. For instance, if a valid composition is marked as unfeasible, it will not be predicted during inference and removed from candidate pseudo-labels during pCZSL training. This may lead to the model producing erroneous outcomes, even reflecting potential biases in the knowledge base. Modeling issues in the knowledge base and/or merging multiple external sources is an important topic for future research in OW-CZSL and pCZSL.

Acknowledgments This work has been partially funded by the ERC (853489 - DEXIM), by the DFG (2064/1 – Project number 390727645), and as part of the Excellence Strategy of the German Federal and State Governments. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Shyamgopal Karthik.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, 2020. 7

[2] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In *NeurIPS*, 2020. 2, 3

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[4] Rodrigo Berriel, Stephane Lathuillere, Moin Nabi, Tassilo Klein, Thiago Oliveira-Santos, Nicu Sebe, and Elisa Ricci. Budget-aware adapters for multi-domain learning. In *CVPR*, 2019. 2

[5] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *ICLR*, 2020. 2

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 2019. 2

[7] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 2, 3

[8] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *CVPR*, 2021. 2

[9] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *CVPR*, 2019. 2

[10] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. *NeurIPS*, 2004. 2, 5, 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[12] Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *CVPR*, 2020. 2

[13] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 4

[14] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2018. 4, 7

[15] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 2, 3

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

[17] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 2

[18] Kaustav Kundu and Joseph Tighe. Exploiting weakly supervised visual patterns to learn from partial annotations. In *NeurIPS*, 2020. 2

[19] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2

[20] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML-WS*, 2013. 2, 4, 5, 7

[21] Hae Beom Lee, Eunho Yang, and Sung Ju Hwang. Deep asymmetric multi-task feature learning. In *ICML*, 2018. 3

[22] Yunsheng Li and Nuno Vasconcelos. Efficient multi-domain learning by covariance normalization. In *CVPR*, 2019. 2

[23] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *CVPR*, 2020. 1, 2, 3, 5

[24] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, 2021. 1, 2, 3, 4, 5, 7

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013. 4, 6

[26] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, 2017. 1, 2, 3, 5, 6

[27] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016. 2, 3

[28] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7

[29] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 2018. 1, 2, 3, 5

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 4

[31] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *ICCV*, 2019. 1, 2, 3, 5

[32] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *NeurIPS*, 2017. 2, 3

[33] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *CVPR*, 2018. 2, 3

[34] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2021. 2

[35] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In *AAAI*, 2019. 2, 3

[36] Frank Ruis, Gertjan J Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. In *NeurIPS*, 2021. 2

[37] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 2, 3

[38] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 2020. 2

[39] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017. 2, 4

[40] Robyn Speer and Joanna Lowry-Duda. Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *International Workshop on Semantic Evaluation (SemEval-2017)*, 2017. 4

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1), 2014. 5

[42] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*, 41(9), 2018. 5

[43] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 4

[44] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *ICCV*, 2017. 4