This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Beyond Seen Primitive Concepts and Attribute-Object Compositional Learning

Nirat Saini Khoi Pham Abhinav Shrivastava
University of Maryland, College Park

Abstract

Learning from seen attribute-object pairs to generalize to unseen compositions has been studied extensively in Compositional Zero-Shot Learning (CZSL). However, CZSL setup is still limited to seen attributes and objects, and cannot generalize to unseen concepts and their compositions. To overcome this limitation, we propose a new task, Open Vocabulary-Compositional Zero-shot Learning (OV-CZSL), where unseen attributes, objects, and unseen compositions are evaluated. To show that OV-CZSL is a challenging yet solvable problem, we propose three new benchmarks based on existing datasets MIT-States [20], C-GQA [29] and VAW-CZSL [37, 43], along with new baselines and evaluation setup. We use language embeddings and external vocabulary with our novel neighborhood expansion loss to allow any method to learn semantic correlations between seen and unseen primitives. Project website: https://ov-czsl.github.io.

1. Introduction

Attributes explain the semantic properties of objects and are essential for efficient in-the-wild object recognition [1, 8, 23]. For instance, an unseen image of a `dog` can be identified for its novel attributes, "a fluffy brown dog" even if the breed is unknown. However, data annotation for object-attribute is prohibitively expensive. It is challenging to scale annotated data because: (1) labels for each attribute and object are required individually, and (2) annotated data is required for all possible object-attribute compositions. Prior works [20, 31, 42, 55] circumvent the first challenge by assuming a limited vocabulary of attributes and objects ('seen' primitives) and focus primarily on the second challenge of unseen compositions, which is referred to as Compositional Zero-shot Learning (CZSL). However, dealing with unseen attributes, unseen objects, and unseen compositions together is still an open problem. In this paper, we propose a new task, *Open Vocabulary-Compositional Zero-shot Learning (OV-CZSL)*, which attempts to address aforementioned challenges simultaneously, (1) dealing with unseen attributes and unseen objects, and (2) recognizing unseen attribute-object compositions.

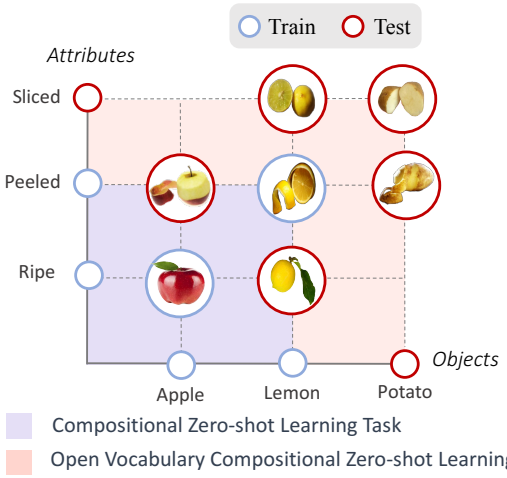

Figure 1. We propose a new task, Open Vocabulary-Compositional Zero-shot Learning (OV-CZSL), that expands upon Compositional Zero-shot Learning (CZSL). CZSL focuses on evaluating unseen compositions `ripe lemon`, `peeled apple` of seen attributes (`ripe`, `peeled`) and seen objects (`apple`, `lemon`), using training samples (`ripe apple`,`peeled lemon`). The novel OV-CZSL is trained with the same data as CZSL. However, it can be evaluated on unseen attributes (`sliced`), objects (`potato`) and their unseen compositions. These unseen compositions also include seen attribute-unseen object (`peeled potato`), unseen attribute-seen object (`sliced lemon`) and unseen attribute-unseen objects (`sliced potato`).

Recognizing unseen classes (Zero-shot learning) and composing unseen relations between primitive seen classes (compositional learning) are both well known challenges in computer vision. Towards a more generic understanding of the unseen concepts, we introduce OV-CZSL that bridges the gap between ZSl and CZSL. Drawing inspiration from neuroscience, humans prototype and learn abstract concepts (fruit) while learning concrete concepts (apple, banana) [44, 58] (also known as concept learning). For instance, we can associate visually similar concepts (apple, orange) as an abstract concept(fruit), by linking them via language semantically. By extending this idea, with a seen concept `peeled lemon`, we can relate `lemon` with another vegetable, *e.g.* `potato`. Even though an image of `peeled potato` is never seen; by understanding

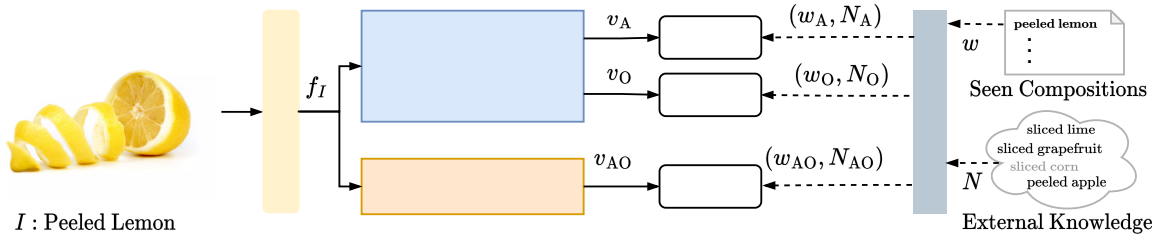

Figure 2. System Overview: ResNet18 [18] is used as visual feature extractor. Label Embedder embeds the image feature v_{AO} along with BERT [5] textual feature for the pair w_{AO}. Object-Attribute Disentanglement module separates visual features into attributes and objects as v_A and v_O. Moreover, along with target labels for attribute w_A, object w_O, and pair w_{AO}, we use neighbors for all of these from the semantic embedding space (N_A, N_O and N_{AO} respectively). We use cosine similarity loss, along with novel *Neighborhood expansion loss* to solve the task of OV-CZSL.

the concept of `peeled`, we can extrapolate what `peeled potato` would look like (Figure 1). Accordingly, we leverage seen attributes and objects, along with external knowledge, to semantically associate and learn new concepts and their compositions.

Zero-shot learning studies [10, 12, 16, 50], mostly uses external language and map unseen concepts close to semantically similar seen concepts. For OV-CZSL, we utilize pre-trained language embeddings (BERT [5]), to acquire semantically organized knowledge. Similar to CZSL, we learn a joint image and language embedding space to deal with unseen compositions. Within this embedding space, we expand the neighborhood around seen concepts to correlate different concepts and expand our vocabulary. Similar to query expansion in image retrieval [6, 19, 51], we enforce the visual feature of semantically similar linguistic concepts to be close in the embedding space. This helps in training the image encoder to recognize visually and semantically similar concepts that it has never seen before [50]. We propose a new regularizing loss function, *Neighborhood Expansion Loss*, to perform this vocabulary expansion efficiently. Henceforth, we propose new benchmarks for the OV-CZSL task, building on three standard attribute-object datasets MIT-States [20], C-GQA [29] and VAW-CZSL [37, 43]. As shown in Figure 1(b), we expand on the CZSL task, to create more challenging sets for evaluating OV-CZSL, which includes a mixture of seen and unseen compositions of seen and unseen attributes and objects. We also propose a *Neighborhood Expansion Loss* that helps train our model to generalize efficiently on unseen pairs, while maintaining the seen pair accuracy of the existing models. To summarize, our contributions are as follows:

- We propose a challenging and practical extension on CZSL, Open Vocabulary-Compositional Zero-shot Learning (OV-CZSL), for learning compositions beyond the seen attributes and objects.
- We create three new benchmarks for MIT-states [20], C-GQA [29] and VAW [37, 43] for OV-CZSL, along with an efficient evaluation setup.
- We also propose an approach for OV-CZSL, which utilizes plug-and-play *Neighborhood Expansion Loss* to regularize training and generalize to unseen concepts and compositions.

1.1. Problem Setup

OV-CZSL is motivated by in-the-wild recognition, where it is infeasible to train on all possible objects and attributes for learning their compositions. To compare it with ZSL and CZSL setups, we discuss the train and test splits for these. Zero-shot Learning (ZSL) has distinct seen labels used for training and unseen labels used for testing. For attributes, we represent the seen and unseen sets as A and A^* respectively. Similarly, ZSL for objects will have O and O^* for seen and unseen sets, respectively. Compositional Zero-Shot Learning (CZSL) deals with unseen compositions of seen attribute and seen object pairs. Training set for CZSL is seen pairs (AO) and test set is unseen compositions of seen object and attribute classes, $(AO)^*$. By definition, there is no overlap between train and test sets in both ZSL and CZSL.

The goal for Open Vocabulary-Compositional Zero-shot Learning (OV-CZSL) is to recognize attributes and objects beyond its seen vocabulary (hence it's open vocabulary). During training, only a set of compositions of seen attributes and objects AO are used. For testing, we evaluate on compositions of seen attribute-unseen object AO^*, unseen attribute-seen object A^*O, unseen attribute-unseen object A^*O^* and a set of unseen compositions of seen attributes and objects $(AO)^*$. The unseen test set for OV-CZSL is $\{AO^*, A^*O, A^*O^*, (AO)^*\}$. Hence, the formulation is OV-CZSL can be considered as a generalized combination of ZSL and CZSL tasks.

2. Related Work

Zero-shot learning. Different from ZSL, Generalized Zero-shot Learning (GZSL) has distinct seen labels used for training and evaluated on both seen and unseen labels during testing. Most works use auxiliary attribute descriptions

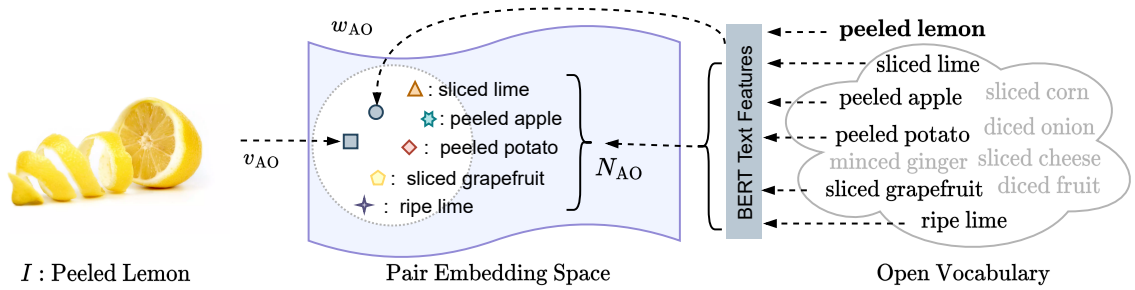

Figure 3. *Neighborhood Expansion Loss*. In the pair embedding space, we embed visual feature v_{AO} close to w_{AO}. Further, we use $k = 5$ neighbors of correct label w_{AO}, and make their word embeddings closer to the visual embedding v_{AO} as well. This gives our model the ability to generalize for unseen compositions.

of the object classes, for ZSL of objects [12, 16, 34, 50]. Other works use word embeddings and pre-computing semantic features from Wikipedia et al., for identifying attributes or objects on commonly used datasets (AwA [24], CUB [48], SUN [35], ImageNet [4]). In summary, there is no datasets and papers zero-shot attribute classification, along with zero-shot object classification and along with composition of attributes-objects.

Compositional Zero-shot learning. Most recent works rely on joint embedding space for images and labels, with linguistic losses [26–29, 33, 39, 52]. Nagarajan *et al*. [33] proposed two scenarios for evaluation: (1) Closed world, where images are classified into only unseen pairs, and (2) Open world, where both seen and unseen pairs are used for evaluation. Moreover, [28] proposed another Open world setting, where for `M` attributes and `N` objects, they consider all possible combinations of `M x N` pairs, which might or might not be part of seen and unseen pairs for evaluation. Note that, in this work, 'Open' refers to extra vocabulary it can generalize to, but the evaluation is done in a closed world setup only. None of the existing works learn to compose novel unseen compositions.

Open Vocabulary. Sergio *et al*. [15] first proposed the task of open vocabulary object retrieval using descriptive natural language phrases by combining category and instance-level recognition. Further, Hang *et al*. [59] proposed an open vocabulary framework for scene parsing, which built an image-text common embedding space using hierarchical semantic relations. Open vocabulary setup has evolved over time, for scene parsing [15], object detection [7, 14, 46, 53, 57], and object segmentation [11]. Moreover, we want to emphasize that most Open Vocabulary works use CLIP [40] in the pipeline, however using CLIP [40] makes the setup not-ZSL, since CLIP has already seen most attributes and objects. Hence, we avoid using CLIP for our setup. To the best of our knowledge, this is the first work exploring Open Vocabulary for Compositional Zero-shot Learning (OV-CZSL), and proposing a new challenging benchmark. We compare with baselines for both tasks ZSL and CZSL for OV-CZSL.

3. Approach

In this section, we describe our method to tackle OV-CZSL. We emphasize that similar to ZSL, for solving OV-CZSL, the approach must balance the tradeoff between seen and unseen class accuracy. Existing works in CZSL consider only the composition of seen attributes and objects [28, 29, 33, 39, 43], and fail to generalize on compositions of unseen classes. We recommend new baselines for future work in OV-CZSL.

3.1. Task Formulation

OV-CZSL focuses on learning to compose seen and unseen attributes and objects. Each image I, has an attribute label A and an object label O. We use $*$ to represent the unseen concepts, *i.e*. unseen attributes are A^* and unseen objects are O^*. Training labels are set of seen attribute-object pairs, represented by Y^s, where $Y^s = AO$. Similar to CZSL, we evaluate on seen attribute-objects yet unseen compositions $(AO)^*$. Further, with unseen attributes and objects, there are three combinations of labels: 1) seen attribute and unseen object AO^*, 2) unseen attribute and seen object A^*O, 3) unseen attribute and unseen object A^*O^*. The overall test set is denoted as Y^u, where $Y^u = (AO)^* \cup AO^* \cup A^*O \cup A^*O^*$. Note that train and test set compositions are mutually exclusive, *i.e*., $Y^s \cap Y^u = \emptyset$.

3.2. Methodology

Our architecture is based upon common state-of-the-art baselines in CZSL (mostly OADis [43]). We extract image and textual features using pre-trained networks (ResNet18 [18] and BERT [5] respectively). The pair embedder is LabelEmbedder [33] (LE) and the Object-Attribute Disentanglement module is from OADis [43], as shown in Figure 2. The Disentanglement module separates the visual features for attribute and object.

System Architecture. Similar to OADis [43], we use the second last layer before AveragePool of pre-trained ResNet18 [18], for spatial features $f_I \in \mathbb{R}^{512\times49}$. The Label Embedder module (MLP) extracts final feature v_{AO} for

Table 1. **Dataset Splits.** We denote $*$ for unseen concept, such that A and O are seen attribute and objects, whereas A^* and O^* denote unseen attributes and objects. AO and $(AO)^*$ are seen and unseen compositions or seen attributes and seen objects respectively. AO^* are seen attribute-unseen object pairs, A^*O are unseen attribute-seen object set and A^*O^* are the unseen attribute-unseen object pairs. We propose new benchmark splits for OV-CZSL on datasets MIT-states [20], C-GQA [29] and VAW-CZSL [37, 43].

	Attributes		Objects		Training Set	Validation Set	Test Set
Datasets	A	A^*	O	O^*	AO	$AO/(AO)^*/A^*O/AO^*/A^*O^*$	$AO/(AO)^*/A^*O/AO^*/A^*O^*$
MIT-states [20]	84	31	182	63	955	236 / 105 / 126 / 177 / 44	289 / 130 / 157 / 218 / 50
C-GQA [29]	311	102	504	170	4094	1012 / 447 / 525 / 517 / 147	1239 / 542 / 664 / 655 / 176
VAW-CZSL [37]	330	135	406	110	7142	1767 / 803 / 1420 / 1253 / 412	2161 / 982 / 1737 / 1532 / 504

pair embedding, with same dimension as the word embedding final feature w_{AO}, extracted from BERT [5] Text Features (Figure 2). The disentanglement module separates the attribute and object visual features using backbone features for I, I_{attr} and I_{obj}, where I_{attr} is image with same attribute as I, and I_{obj} is an image with same object as I. I and I_{attr} are used to extract visual feature for attribute v_{A}. Similarly, visual feature for object v_{O} is extracted from I and I_{obj}. Using textual features (w_{A} and w_{O}), these visual embeddings for attributes and objects are regularized, *i.e.* given a seen image with label `peeled apple`, we push the visual embedding of this image closer to text embedding of the label. More details can be found in [43]. We use cross-entropy along with cosine similarity to get the final classification score for each attribute, object and pair, same as [29]. Let visual feature is v and text feature is w, y is the correct seen label, then the main classifier logits can be defined as (where δ is the temperature factor):

$$C(v,w) = \frac{e^{h(v,w)}}{\sum_{y \in Y^s} e^{h(v,y)}}$$
$$h(v,w) = \cos(v,w) = \delta \cdot \frac{v^T w}{\|v\| \, \|w\|} \quad (1)$$

3.3. Neighborhood Expansion Loss

Previous CZSL methods perform fairly well for OV-CZSL, however fail to generalize on totally unseen compositions set, *i.e.* A^*O^*. By using common techniques of data augmentation and learning rate decay leads to improved generalization on unseen compositions, but deteriorates performance on seen compositions. Hence the goal of novel *Neighborhood Expansion Loss* (NEL) is to balance the generalizability and learnability of the model. To achieve this, we leverage ideas from *label smoothing* and *label propagation*. Label smoothing is a regularization technique used to reduce overfitting. Although, as explained in [32, 56], it is not always helpful and often introduces noise in the system. We use label smoothing to make the model less confident with training classes, such that negative bias for unseen classes decreases, and the model does not overfit on seen classes. Further, to transfer knowledge from seen to unseen concepts, we use label propagation techniques [60]. It is a graph-based method for semi-supervised learning [9, 47]. In a transductive setting, given labeled and unlabeled examples, label propagation defines a graph between samples, to connect unlabeled samples to potential labels. In our paper, we use neighbors from external knowledge sources (open vocabulary) to learn unseen pairs from seen pairs.

With traditional cross-entropy loss, the model only learns to minimize the distance between the the visual embeddings and text embeddings for each image and correct label (`peeled lemon`). However, for limited seen compositions, this strategy causes a negative bias against unseen pairs. In a way, it forces the model to never learn unseen pairs, since it is too confident for the correct label. To overcome this, we apply label smoothing, which reduces negative bias for unseen classes during training with NEL. Another problem is extending knowledge from seen to unseen compositions. For that, we use label propagation. For each seen pair, we find k nearest-neighbors N_k using word embeddings for labels (details in next section). These neighbors make the 'open vocabulary' aspect of OV-CZSL. As shown in Figure 3, if seen pair is `peeled lemon`, we also learn visual embeddings for unseen compositions `sliced lime`, `peeled apple`, `peeled potato`, `sliced grapefruit` and `ripe lime` as well by minimizing the distance between visual feature for given image of `peeled lemon` with the textual embeddings of the 5 mentioned neighbors. This intuition is from human learning, that we can correlate similar looking objects using language, such as oranges are similar to lemons. With this new loss, we expand our vocabulary beyond seen classes, by correlating similar concepts using language priors. Note that these neighbors are weighted, and the distance between visual feature with its original label (`peeled lemon`) is smaller than distance with its neighbor embeddings. so that the correct label is still learnt with higher confidence than it's neighbors.

Let N represent the neighbor set of seen pair text embedding w. For M training labels, cross-entropy is defined as H where, y_m is 1 for the correct class and 0 for the rest. Label smoothing [32] makes the model less confident for the

Table 2. **Results on MIT-states [20] and CGQA [28].** We report Top 1 AUC, which balances % between seen and unseen compositions with different bias terms. HM is Harmonic Mean where maximum AUC is computed. Following [39], best accuracy values are reported for Seen AO and Unseen pairs $\{AO^*, A^*O, A^*O^*, (AO)^*\}$. All other accuracies for individual splits are computed with bias where HM is maximum. Our method with NEL loss outperforms previous methods on most unseen compositions.

	MIT-States									C-GQA								
Model	Test@1	HM	Seen	Unseen	AO	$(AO)^*$	A^*O	AO^*	A^*O^*	Test@1	HM	Seen	Unseen	AO	$(AO)^*$	A^*O	AO^*	A^*O^*
LE [33]	1.01	7.64	16.29	9.46	10.24	11.38	5.98	4.15	2.87	1.17	8.39	19.37	8.36	10.76	6.51	9.53	2.67	1.08
CompCos [28]	1.97	10.22	26.53	10.29	14.32	21.09	5.86	2.89	0.63	2.35	9.64	40.19	7.25	21.19	20.24	4.47	1.95	0.26
OADis [43]	1.83	9.55	25.35	10.79	12.18	16.06	6.40	5.41	1.34	2.33	9.74	**42.88**	7.12	20.86	15.19	6.17	3.47	0.61
SCEN [45]	1.73	9.72	22.08	8.25	11.85	**30.02**	3.82	0.33	0.08	1.97	9.03	41.65	7.83	20.65	**21.42**	3.61	1.08	0.05
CANet [49]	2.40	10.52	26.42	9.54	**16.56**	23.08	6.15	4.08	0.58	3.04	11.96	40.52	9.21	**22.43**	20.87	4.95	2.03	0.64
Ours	**2.41**	**10.94**	**29.02**	**11.13**	14.11	18.87	**8.24**	**5.49**	**3.54**	**3.18**	**12.11**	42.38	**9.77**	19.78	16.07	**12.86**	**2.87**	**3.04**

Table 3. **Results on VAW-CZSL [37, 43].** All measures are shown in Top 3 AUC. HM is Harmonic Mean where maximum AUC is computed. All other accuracies for individual splits are computed with bias where HM is maximum. Our approach outperforms previous baselines for unseen compositions.

Model	Test@3	HM	AO	$(AO)^*$	A^*O	AO^*	A^*O^*
LE [33]	1.49	8.27	15.62	10.48	5.79	2.78	0.98
CompCos [28]	2.69	10.68	20.21	20.58	5.04	2.48	0.5
OADis [43]	2.68	10.91	21.19	15.65	6.75	3.16	0.76
SCEN [45]	2.53	10.64	19.06	**20.76**	4.52	2.05	0.42
CANet [49]	2.89	11.21	**24.56**	18.42	5.74	2.86	0.95
Ours	**2.91**	**11.35**	23.02	16.18	**7.86**	**3.37**	**1.36**

correct label, by weighting the loss for correct label with a smoothing factor $\alpha < 1$. Thus cross-entropy between the modified targets y_m^{LS} and and the network's outputs C_m is minimized.

$$H(y, C) = \sum_{m=1}^{M} -y_m \log(C_m)$$
$$y_m^{LS} = y_m(1-\alpha) + \alpha/M \tag{2}$$

The *Neighborhood Expansion* makes the target for actual label is weighted highest, and rest weights are distributed among neighbors, with least weights are assigned to other labels.

$$y_m^{NE} = y_m(1-\alpha)T + y_k(1-\alpha)(1-T) + \alpha/(M+k)$$

where, k is the number of neighbors, y_k are labels for neighbors from open vocabulary, T is smoothing term for label propagation (weighting neighbors) and α is smoothing term for label smoothing. Thus, *Neighborhood Expansion Loss* is a combination of label propagation and label smoothing. We use Cosine Similarity-based cross-entropy H classification loss and *Neighborhood Expansion Loss* for attribute, object, and pair embeddings, represented by Attr Cls, Obj Cls, and Pair Cls shown in Figure 2. Here, we elaborate on the Pair Cls loss:

$$\mathcal{L}_{\text{AO}} = H(y_{AO}, C(v_{\text{AO}}, w_{\text{AO}}))$$
$$\mathcal{L}_{\text{AO}}^{\text{NE}} = H(y_{AO}^{\text{NE}}, C(v_{\text{AO}}, w_{\text{AO}})) \tag{3}$$

We can define similar loss functions for Attr Cls and Obj Cls as well. For each embedding space, we use both losses, and overall objective function $\mathcal{L}$ is:

$$\mathcal{L}_{\text{pair}} = \beta_1 \mathcal{L}_{\text{AO}} + (1-\beta_1)\mathcal{L}_{\text{AO}}^{\text{NE}}$$
$$\mathcal{L}_{\text{attr}} = \beta_2 \mathcal{L}_{\text{A}} + (1-\beta_2)\mathcal{L}_{\text{A}}^{\text{NE}}$$
$$\mathcal{L}_{\text{obj}} = \beta_3 \mathcal{L}_{\text{O}} + (1-\beta_3)\mathcal{L}_{\text{O}}^{\text{NE}}$$
$$\mathcal{L} = \mathcal{L}_{\text{pair}} + \gamma_1 \mathcal{L}_{\text{attr}} + \gamma_2 \mathcal{L}_{\text{obj}}. \tag{4}$$

4. Experimental Setup

Following CZSL works, we propose new splits for OV-CZSL on existing dataset MIT-States [20], C-GQA [29] and VAW-CZSL [37, 43]. MIT-states [20] is relatively small, is a popular choice for CZSL. It has 115 attributes, 245 objects, 1962 compositional pairs, and 53k images. C-GQA [29] a larger dataset with 413 attributes, 674 objects and a total of $\sim 7k$ pairs and 39k images. VAW-CZSL [37, 43] is the largest of all, with 533 attributes, 543 objects, $\sim 15k$ (15785) pairs and 92k images. The scale for C-GQA and VAW-CZSL is similar, however, VAW-CZSL has more shared attributes among objects. C-GQA has more one-to-one attribute object pairing, without much sharing of attributes among objects. We do not use UT-Zappos [55] since it only has 16 attributes and 12 objects.

Dataset Splits. Following the ZSL works [16, 50], we split attributes and objects into 75-25% split for training and testing. Since we are using ResNet18 [18] trained on Imagenet [4] as backbone for visual features, we make sure that attributes and objects common with Imagenet labels are part of training set (seen attributes and seen objects). This ensures that the unseen attributes and objects are truly zero-shot, and are never seen. A set of valid compositions of seen attribute and object pairs becomes the training set Y^s. Test set, denoted as Y^u has unseen compositions of seen attributes and objects $(AO)^*$, as well as, other sets of seen and unseen attributes with unseen objects (AO^*, A^*O,

Table 4. **Comparison with ZSL baselines.** We show results on two ZSL baselines for seen and unseen attributes and objects. Using NEL significantly improves outperforms existing ZSL baselines.

Model	Emb.	A	A^*	O	O^*
SEKG [50]	GloVe	**5.04**	**3.10**	5.19	**3.17**
	BERT	3.37	2.33	**5.72**	2.29
TF-VAEGAN [34]	GloVe	**5.86**	**5.76**	**5.93**	**4.52**
	BERT	4.54	3.05	5.82	3.93
Ours	GloVe	**20.75**	**13.67**	**32.19**	7.86
	BERT	20.37	11.42	29.3	**10.07**

A^*O^*). We further split Y^u 40-60%, to make validation and test splits, similar to CZSL. To evaluate learnability on seen compositions, test and validation split also have subset of Y^s. All split creations are random and are selected from 10 random splits, based on the balance between seen and unseen accuracies (refer supplementary). Table 1 shows statistics of the benchmark split.

Evaluation. We follow Generalized CZSL [26, 28, 39, 43] evaluation protocol, to evaluate on both seen and unseen pairs (Y^s,Y^u) with a scalar term used to overcome negative bias for unseen pairs. Since, the OV-CZSL task is already challenging, we use Closed world evaluation setup (mentioned in related work), where AUC is reported over only "valid" unseen pairs while ignoring the "invalid" pairs. Area Under the Curve (AUC) is computed between the accuracy on seen and unseen compositions with different bias and Harmonic Mean (HM), to balance the bias. We also report separately accuracy for each set (AO, $(AO)^*$, AO^*, A^*O, A^*O^*) at the bias-term where HM is maximum. The seen pairs Y^s consist of AO whereas the unseen pairs Y^u are $\{AO^*, A^*O, A^*O^*, (AO)^*\}$. Following [39], best accuracy values are reported for seen and unseen pairs.

Neighborhood list. Since some datasets are small, and the testing pair labels might not be semantically similar to training pairs. To expand the vocabulary, we use external knowledge sources. In total, we use 2294 attributes/adjectives and 4090 objects aggregated from Visual Genome [22], Flick30k [54], COCO-captions [3], and LocalizedNarratives [38]. These attributes and object pairs make up ∼118650 compositions, as valid extra compositions we use for neighborhood search. We find 10 neighbors for each seen attribute and objects from the external source, using GloVe [36] embeddings cosine similarity score. Using the neighbors for attribute and object, we find all possible compositions (∼100) for each pair. The compositions are then analyzed for validity using the external pair list described above. 10 valid compositions are chosen as neighbors for the seen pair. If the external pair list does not have any compositions, we only choose compositions of first 6 closest attributes and objects from the neighbor list. Hence, neighbor search is based on individual attribute and object, instead of pairs directly. More details on for neighbor search and hyperparameter sensitivity of NEL are explained in suppl.

Table 5. **Using NEL with other baselines**. We show effect of NEL for different baselines. All methods using NEL perform better for OV-CZSL splits.

Model	AO	$(AO)^*$	A^*O	AO^*	A^*O^*
LE [33]	10.24	11.38	5.98	4.15	2.87
LE + NEL	10.65	6.11	5.61	4.09	7.71
	+0.4	-5.2	-0.3	-	+5.5
CompCos [29]	14.32	21.09	5.86	2.89	0.63
CompCos + NEL	15.73	19.72	8.03	4.13	1.77
	+1.4	-1.3	+2.1	+1.24	+1.1
OADis [43]	12.18	16.06	6.40	5.41	1.34
Ours (OADis+NEL)	14.11	18.87	8.24	5.49	3.54
	+2.9	+2.8	+1.8	-	+2.2
CANet [49]	16.56	23.08	6.15	4.08	0.58
CANet [49] + NEL	19.24	25.94	6.78	5.12	1.53
	+2.7	+2.8	-	+1.0	+1.0
CLIP [40] (ZSL)	20.08	22.03	23.38	23.21	34.93
CLIP+ FT	25.89	22.03	24.87	25.60	34.50
CLIP+FT+NEL	25.97	23.06	25.90	25.70	36.45
	-	+1.0	+1.3	-	+2.0

Training Details. Following standard practice in CZSL [28, 29, 33, 39], we use Frozen ResNet18 [18], pretrained on ImageNet [4] for image features (without finetuning) and BERT [5] text embeddings for labels. A linear layer on top of BERT [5] features is used for pair embeddings. We use image augmentations (random crop, horizontal flip) for all baselines and our method, like OADis [43]. For MIT-States [20], the network is trained with Adam optimizer, with weight decay 1e-6, learning rate 3e-5 and decay at epoch 120 and 130. Smoothing factor $\alpha = 0.8$, temperature for cosine similarity $\delta = 0.05$, temperature for weights of neighbors $T = 0.5$, number of neighbors $k = 5$, weights for losses are $\beta_1 = 0.8, \beta_2 = \beta_3 = 0.95$, and $\gamma_1 = \gamma_2 = 0.05$. The model is trained for 150 epochs and best performance based on validation AO performance. More details are mentioned in suppl.

5. Results

5.1. CZSL Baselines

The main task for OV-CZSL is compositional learning for seen and unseen attribute-object pairs. Our architecture is most similar to OADis [43], with embedding space and

Table 6. **Ablation for varying number of neighbors (k)**. We show how changing the number of neighbors can introduce noisy labels in the setup. For generalization to unseen classes, we need fewer neighbors (5 as shown here).

k	Val@1	Test@1	HM	AO	$(AO)*$	A^*O	AO^*	A^*O^*
1	2.67	2.21	10.46	15.16	17.46	8.02	4.41	2.21
3	2.69	2.18	10.39	**16.37**	16.71	7.67	4.47	2.21
5	**2.69**	**2.41**	**10.94**	14.11	**18.87**	**8.24**	**5.49**	**3.54**
7	2.65	2.24	10.44	15.16	17.67	8.09	4.50	2.37
10	2.60	2.16	10.35	13.79	18.33	8.01	4.61	2.14

losses from LE [33] and CompCos [28]. We include some recent models such as SCEN [45] and CANet [49], and comapre our method with 5 baselines. Other baselines for CZSL, (*e.g*. TMN [39], AttrOpr [33], GraphEmb [29]) are either outperformed by CompCos [28] or perform poorly on unseen pairs (*e.g*. KG-SP [21], Symnet [26]). For fair comparison, all baselines use ResNet18 [18] visual features and BERT [5] word embeddings.

Results on MIT-States and C-GQA. We report the AUC and harmonic mean (HM) for Top1 predictions on both datasets in Table 2. Seen and Unseen accuracy is the overall best accuracy without bias calibration. Interestingly, LE [33] generalizes well on unseen compositions, but does not perform as well for seen compositions overall. Moreover, CompCos [28] works the best for CZSL task, as its performance for AO and $(AO)^*$ remains unbeatable. However, it fails to generalize on compositions of unseen attributes or objects. OADis [43] is somewhere in the middle but does not generalize as well on unseen compositions A^*O^*. SCEN [45] is slightly better than OADis [43] for unseen compositions of seen concepts [$(AO)^*$], but yields very poor results for the unseen categories [A^*O^*]. CANet [49] also yields better performance in AO and $(AO)^*$, but fails to generalize on A^*O^*. Our proposed approach gets best of both world, as it performs close to other baselines for CZSL task and beats most of those for unseen compositions A^*O, AO^* and A^*O^*.

For C-GQA [29], there are less attributes that share a common object (and vice-versa), which makes object-attribute disentanglement inefficient on this dataset. It relies more on backbone visual features (biased for objects), and performs well on A^*O than on AO^*. Our method with NEL surpasses most unseen sets with significant margin. Similar to MIT-states [20], SCEN [45] does better on unseen compositions of seen concepts, while CANet [49] does better on seen compositions AO. Overall, in comparing SCEN, CANet and our method, for CZSL compositions AO and $(AO)^*$, the drop is accuracy is $< 5\%$ while the improvement in A^*O^* is almost 5 times, which is the most challenging split of unseen attribute-unseen object compositions.

Results on VAW-CZSL. This dataset created from multi-label VAW [37], and VAW-CZSL [43] uses the least frequent labels for each image across the dataset. Hence top 1 predictions latch on to different attributes which are present in the image, but are not labeled. We show Top 3 AUC, HM and accuracy for this dataset. Similar variations in performance of baselines are obeserved for this dataset as well Table 3, *i.e*. SCEN performs better in $(AO)^*$, CANet performs slightly better for AO while our method outperforms all baselines in every other set, majorly the unseen compositions splits A^*O, AO^* and A^*O^*.

5.2. ZSL Baselines

Zero-shot Learning (ZSL) for unseen attributes and objects is a by-product of OV-CZSL. Hence, we also compare with some baselines from ZSL. There are various ZSL methods, but most of them use pre-defined semantics for objects [12, 13, 50], AwA [24], CUB [48], Imagenet [4]. Our work only relies on text label embeddings and visual features, therefore we compare with two recent works TF-VAEGAN [34] and CE-GZSL [16]. Note that these works do not have any information about unseen test classes during training, which might or might not be part of external knowledge but are not explicitly added to the open vocabulary. These methods use semantic features, which we replace with GloVe [36] and BERT [5] text embeddings. We show results for MIT-States [20] as we observe similar pattern for other datasets as well. All the values are bias calibrated, to balance seen and unseen accuracy, for Top-1 predictions. Our proposed method outperforms common baselines in ZSL as well, with a significant margin. More details on why GloVe [36] embeddings perform the best are explained in the next section.

5.3. Ablation Studies

We experimentally motivate the design choices as well as novelty aspect of our approach, mostly based on performance on A^*O^* split. All the experiments in this section are shown for MIT-States [20] dataset. More ablation can be found in Appendix.

Neighborhood Expansion Loss. As the main contribution of this work is OV-CZSL and *Neighborhood Expansion Loss*, we explore if this loss can be used as plug-and-play for other baselines. In Table 5, third row for each model, shows the change caused by using NEL with respect to originally without NEL. red denotes negative change in value, green denotes the positive change and '-' represents no or change is less than ± 0.2. All methods with NEL improve significantly for A^*O^* split, whereas OADis [43] and CANet [49] improves for all splits. This is because NEL is applied for not just pair embedding space, but also for attribute and object spaces, which are not present for CompCos [29] and LE [33]. Although, as mentioned earlier,

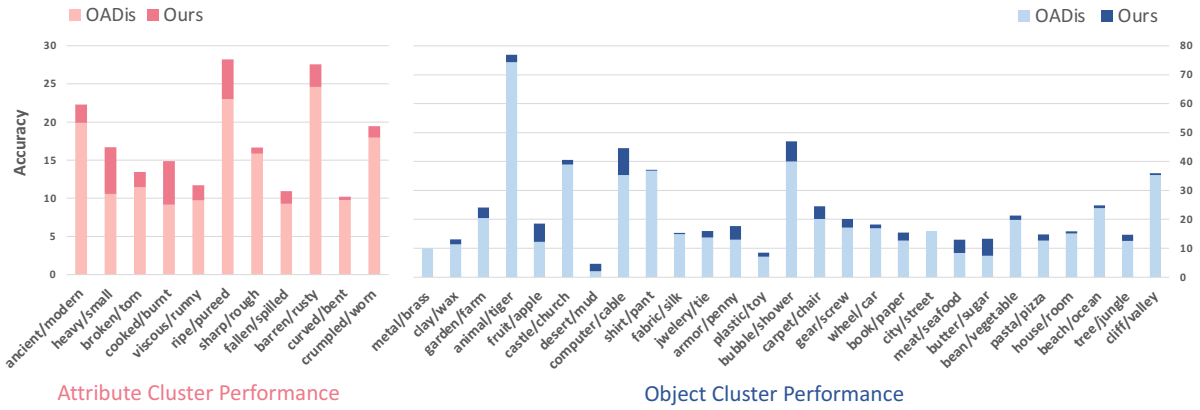

Figure 4. To show the impact of NEL, we cluster all attribute and object class labels for MIT-States [20], and evaluate the accuracy for each cluster. These clusters have both seen (ancient, new, young) and unseen (modern, old) concrete concepts, such that each cluster has one abstract concept (time/age). Our model (OADis+NEL) shows either same or improved performance for attributes and object clusters, backing our hypothesis of NEL generalizes to unseen concrete concepts for the similar abstract concepts.

CLIP [40] makes the setup not-ZSL, we still acknowledge the presence of such larger models and show the importance of NEL loss. The loss can be applied along with CLIP [40] loss function, and slightly improve generalization for the unseen compositions. Despite NEL being a small modification and the CLIP loss has more stronger impact, we show the potential to harness bigger models to generalize beyond seen classes. We expect similar boost for other Vision+Language models such as LLaVA [17] and BLIP [25] along with NEL, while showing results with CLIP [40] as a representative sample.

Number of Neighbors. Using too many noisy labels in label smoothing can affect the performance adversely [32, 56]. In our case, more neighbors can act as noisy labels. We experiment to find the ideal number of neighbors, as shown in Table 6. Among all numbers, our model achieves best performance using 5 neighbors on MIT-States [20] and C-GQA [29]. For VAW-CZSL [37, 43], we use 10 neighbors for best performance.

5.4. Can NEL help in learning abstract concepts?

Our inspiration for NEL is to learn to generalize similar concrete concepts (apple, banana) to all other abstract concepts (any fruits), such that if 1-2 fruits are seen in the training set, the model generalizes to other unseen fruits in the test set. To quantitatively verify this, we cluster all attributes and objects classes for MIT-states [20], using k-nearest neighbors, along with a threshold cosine similarity of 0.4 across GLoVe [36] features. We manually check the similarity between these label clusters and ignore the attributes and objects across clusters: (1) cluster size is less than 3 and (2) clusters which have classes that are only part of test set, without having any class in training set. Each cluster has some seen and unseen concepts. We compare OADis [43] and Our (OADis+NEL) model's accuracy for these clusters. Attributes are spread across 11 clusters, and objects are spit into 26 clusters. As shown in Figure 5, each cluster is represented with two labels from the cluster to disclose the abstract concept of the cluster. We observe that most clusters either show same or improved performance while using NET (for our model), across all attributes and objects. This shows that similar concepts are learnt together to generalize to unseen abstract concepts. However, this setup is still limited, such that the model cannot discriminate similar concepts `peel` from `slice` and `chop`, but can only learn these together closer in the visual+textual embedding space as `styles of cutting`.

6. Conclusion and Discussion

In the era of CLIP [40] and DALL-E [41], we emphasize that labeled data is still a bottleneck, which these bigger models dodge by extensively using all available labeled data. We present a novel task OV-CZSL, which not only focuses on learning unseen compositions of seen attributes and objects, but also generalizes to unseen attributes, unseen objects and their compositions. We propose new benchmark splits, backed by scientifically stable methods on existing datasets, MIT-states [20], C-GQA [29] and VAW-CZSL [37, 43]. Any CZSL model (or even LLM) can be extended with Neighborhood Expansion Loss, for solving OV-CZSL, through semantic transfer from pretrained language embeddings. Open vocabulary compositional learning is still a challenging problem, with tremendous scope in improving benchmarks (datasets and evaluation). This work is an attempt towards exploring generalized compositional learning for attributes and objects.

Acknowledgements. This work was supported by NSF CAREER Award (#2238769).

References

[1] Gedas Bertasius and Lorenzo Torresani. Cobe: Contextualized object embeddings from narrated instructional video. *Advances in Neural Information Processing Systems*, 33: 15133–15145, 2020. 1

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 2

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, 2015. 6

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3, 5, 6, 7, 1

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2, 3, 4, 6, 7, 1

[6] Victor C. Dibia. Neuralqa: A usable library for question answering (contextual query expansion + bert) on large datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.*, 2020. 2

[7] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 3

[8] Ali Farhadi, Ian Endres, Derek Hoiem, and David Alexander Forsyth. Describing objects by their attributes. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, 2009. 1

[9] Rob Fergus, Yair Weiss, and Antonio Torralba. Semi-supervised learning in gigantic image collections. *Advances in neural information processing systems*, 22, 2009. 4

[10] Junyuan Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI Conference on Artificial Intelligence*, 2019. 2

[11] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. 3

[12] Pallabi Ghosh, Nirat Saini, Larry S. Davis, and Abhinav Shrivastava. All about knowledge graphs for actions. *ArXiv*, abs/2008.12432, 2020. 2, 3, 7

[13] Pallabi Ghosh, Nirat Saini, Larry S. Davis, and Abhinav Shrivastava. Learning graphs for knowledge transfer with limited labels. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11146–11156, 2021. 7

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2021. 3

[15] Sergio Guadarrama, Erik Rodner, Kate Saenko, Ning Zhang, Ryan Farrell, Jeff Donahue, and Trevor Darrell. Open-vocabulary object retrieval. In *Robotics: science and systems*, page 6, 2014. 3

[16] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2371–2381, 2021. 2, 3, 5, 7

[17] Qingyang Wu Yong Jae Lee Haotian Liu, Chunyuan Li. Visual instruction tuning. In *Neural Information Processing Systems*, 2023. 8

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 3, 5, 6, 7, 1

[19] Yuanfeng He, Yuanxi Li, Jiajia Lei, and Clement H. C. Leung. A framework of query expansion for image retrieval based on knowledge base and concept similarity. *Neurocomputing*, 204:26–32, 2016. 2

[20] Phillip Isola, Joseph J. Lim, and Edward Adelson. Discovering states and transformations in image collections. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1383–1391, 2015. 1, 2, 4, 5, 6, 7, 8, 3

[21] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9326–9335, 2022. 7, 3

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 6

[23] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009. 1

[24] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:453–465, 2014. 3, 7

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 8

[26] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11313–11322, 2020. 3, 6, 7

[27] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016.

[28] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 5, 6, 7, 1

[29] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP: 1–1, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[30] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013. 2

[31] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1160–1169, 2017. 1

[32] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. *When does label smoothing help?* Curran Associates Inc., Red Hook, NY, USA, 2019. 4, 8

[33] Tushar Nagarajan and Kristen Grauman. Attributes as operators: Factorizing unseen attribute-object compositions. In *European Conference on Computer Vision (ECCV)*, 2018. 3, 5, 6, 7, 1, 2

[34] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *European Conference on Computer Vision*, pages 479–495. Springer, 2020. 3, 6, 7

[35] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758, 2012. 3

[36] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 6, 7, 8, 2

[37] Khoi Pham, Kushal Kafle, Zhe Lin, Zhi Ding, Scott D. Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13013–13023, 2021. 1, 2, 4, 5, 7, 8

[38] Jordi Pont-Tuset, Jasper R. R. Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, 2019. 6

[39] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3592–3601, 2019. 3, 5, 6, 7, 1, 4

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 6, 8, 5

[41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 8

[42] Nirat Saini, Bo He, Gaurav Shrivastava, Sai Saketh Rambhatla, and Abhinav Shrivastava. Recognizing actions using object states. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022. 1

[43] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13648–13657, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[44] Nirat Saini, Hanyu Wang, Archana Swaminathan, Vinoj Jayasundara, Bo He, Kamal Gupta, and Abhinav Shrivastava. Chop & learn: Recognizing and generating object-state compositions. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20190–20201, 2023. 1

[45] Kwan-Yee K. Wong Shaozhe Hao, Kai Han. Learning attention as disentangler for compositional zero-shot learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5, 7

[46] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9611–9620, 2022. 3

[47] Abhinav Shrivastava, Saurabh Singh, and Abhinav Kumar Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *European Conference on Computer Vision*, 2012. 4

[48] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 3, 7

[49] Qingsheng Wang, Lingqiao Liu, Chenchen Jing, Hao Chen, Guoqiang Liang, Peng Wang, and Chunhua Shen. Learning conditional attributes for compositional zero-shot learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11197–11206, 2023. 5, 6, 7

[50] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6857–6866, 2018. 2, 3, 5, 6, 7

[51] Hongtao Xie, Yongdong Zhang, Jianlong Tan, Li Guo, and Jintao Li. Contextual query expansion for image retrieval. *IEEE Transactions on Multimedia*, 16:1104–1114, 2014. 2

[52] Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10245–10253, 2020. 3

[53] Keren Ye, Mingda Zhang, Wei Li, Danfeng Qin, Adriana Kovashka, and Jesse Berent. Learning to discover and localize

visual objects with open vocabulary. *ArXiv*, abs/1811.10080, 2018. 3

[54] Peter Young, Alice Lai, Micah Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[55] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014. 1, 5

[56] Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3902–3910, 2020. 4, 8

[57] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14388–14397, 2021. 3

[58] Dagmar Zeithamova, Michael L. Mack, Kurt Braunlich, Tyler Davis, Carol A. Seger, Marlieke T.R. van Kesteren, and Andreas Wutz. Brain mechanisms of concept learning. *Journal of Neuroscience*, 39(42):8259–8266, 2019. 1

[59] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2002–2010, 2017. 3

[60] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Scholkopf. Learning with local and global consistency. In *Neural Information Processing Systems*, 2003. 4