LOCL: Learning Object-Attribute Composition using Localization

Satish Kumar
satishkumar@ucsb.edu
ASM Iftekhar
iftekhar@ucsb.edu
Ekta Prashnani
eprashnani@ucsb.edu
B.S. Manjunath
manj@ucsb.edu

ECE Department,
University of California
Santa Barbara

Abstract

This paper describes LOCL: Learning Object-Attribute (O-A) Composition using Localization – that generalizes composition zero shot learning to objects in cluttered/more realistic settings. The problem of unseen O-A associations has been well studied in the field, however, the performance of existing methods is limited in challenging scenes. In this context, our key contribution is a modular approach to localizing objects and attributes of interest in a weakly supervised context that generalizes robustly to unseen configurations. Localization coupled with a composition classifier significantly outperforms state-of-the-art (SOTA) methods, with an improvement of about 12% on currently available challenging datasets. Further, the modularity enables the use of localized feature extractor to be used with existing O-A compositional learning methods to improve their overall performance.

1 Introduction

Human visual reasoning allows us to leverage prior visual experience to recognize previously unseen Object-Attribute (O-A) relationships. Predicting such complex relationships of novel O-A compositions – referred to as Composition Zero Shot Learning (CZSL) [16, 18, 20, 21, 24, 27, 32, 35]–is an active area of research. There has been significant progress on CZSL methods in recent years, however, as our experiments demonstrate, their performance degrades in natural cluttered scenes, as illustrated in Fig.1. The main reason in these cases is the interference from the other potential confusing elements. For example, in Fig. 1(B.1), the SOTA methods are not able to detect the object of interest given its size relative to image; and while the bird is the object of interest in Fig. 1(B.2), the surrounding context dominated by the green leaves results in an incorrect association of the color attribute to the object.

The poor performance of the SOTA methods can be attributed to the dominant confounding elements thereby impeding the right O-A composition prediction. This in turn is due to the bias towards seen O-A composition during training time. Generalization to more realistic cases as seen in Fig. 1(B) is crucial for the widespread use of CZSL.

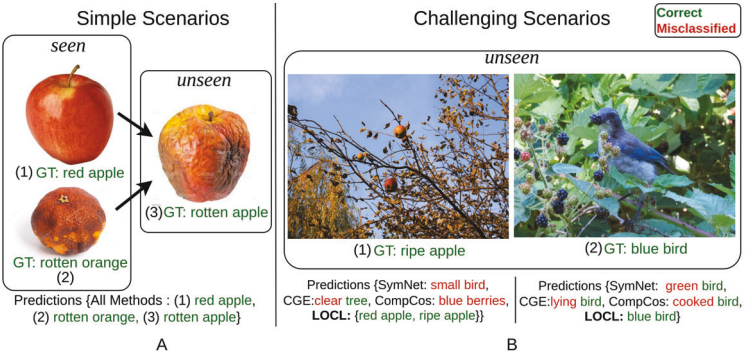

Figure 1: The object of interest shown in images A.1, A.2, and A.3 presents simple scenarios where all SOTA (SymNet [16], CGE [21], CompCos [17]) methods make correct O-A associations. However, for the same object (*apple*) in a more cluttered scene in image B.1, these methods fail. Even in cases where there is a dominant object of interest, such as a *bird* in (B.2), where there is significant background clutter, most of the SOTA methods have incorrect O-A associations.

Inspired by these limitations, we propose Learning Object-Attribute Composition using Localization (LOCL). Our model (LOCL) leverages spatially-localized learning, which is not present in the existing CZSL networks. It is reasonable to ask *Why not first localize the objects and then associate the attributes?* In principle, this can be done, however, the SOTA methods for object detection and localization use extensive datasets for their training. Hence it will not be possible to meaningfully test the CZSL with pre-trained detectors. The images shown in Fig. 1 are from existing datasets for CZSL methods [11, 21, 36]. *We note that all the experiments reported in this paper use the datasets that are created for evaluating CZSL approaches.*

Existing SOTA object-attribute detection approaches do not take into account the possibility of scene attributes confounding with correct O-A composition prediction [16, 18, 21]. These methods are designed to work with wholistic image features [20, 24, 27, 32, 35]. Some recent work address this issue by partitioning the image into regions [7, 34] or equal-size grid cells [8, 12, 37], but they are not very effective in capturing distinctive object features.

Our approach towards better generalization of CZSL to more challenging images (Fig. 1.B) with background clutter is to leverage localized feature extraction in O-A composition. Specifically, we adopt a two step approach. First, a Localized Feature Extractor (LFE) associates an object with its attribute by reducing the interference arising from additional attribute-related visual cues occurring elsewhere in the image. The CZSL benchmark datasets *do not* contain any localization information. As noted before, off-the-shelf object detectors can be inadvertently exposed [26] to unseen OA compositions. Therefore, we developed a weakly supervised method for localized feature extraction. Second, the composition classifier uses the localized distinctive visual features to predict an O-A pair.

The proposed LOCL outperforms competing CZSL methods on *all* existing datasets – including the more challenging CGQA – providing a strong evidence in favor of its applicability to more realistic scenes. Further, the performance of all existing methods improve when our localized feature extractor is included as a pre-processing module – although LOCL still outperforms these methods.

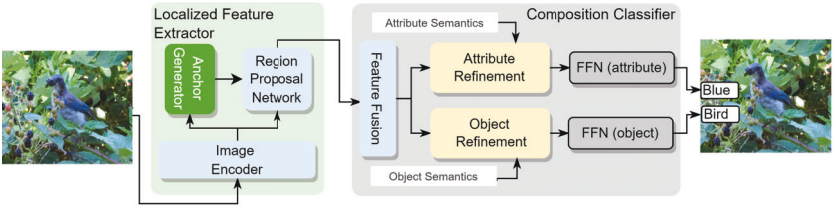

Figure 2: LOCL architecture. The Localized Feature Extractor (Section: 3.1) generates proposals that are likely to contain objects. These proposals are refined with the object and attribute semantics using Composition Classifier (Section: 3.2).

2 Related Work

Existing work in object attribute (O-A) CZSL task typically assume that the images of the object of interest present in uncluttered context. This assumption is also true for the initial CZSL datasets [11, 36]. As a result, most of the CZSL methods [1, 18, 20, 21, 22, 23, 24, 27, 32, 35] perform quite well on uncluttered scenes. As noted before, some of the methods reduce the interference from confounding elements by partitioning the image. [37] is designed for datasets with a dominant object with a clear background. [7] partitions the image to equal-sized grid cells and relies on aligning the attribute semantic and visual features. Further, the dataset used in [7, 37] consist of one object type,e.g., face or bird images.

The O-A problem is considered as a matching problem in a latent space [20, 23, 24, 32]. For this matching task, [24] proposes a modular network with a dynamic gating whereas [20] defines objects and attributes using a support vector machine (SVM). On the other hand, [16, 22] consider attributes as functional operation over objects. More recently [1, 18, 21, 27, 35] focus on the relationships among attributes and objects. [1] disentangle attributes and objects with metric based learning. [21] learns attribute-object dependence in a semi-supervised graph structure where unseen combinations are considered connected during training. This is extended in [18, 31, 35] where all possible combination of objects and attributes are considered during inference.

In general, the performance of current methods drop significantly on images with background clutter. Taking inspiration from the generic pipeline of weakly supervised object detection (WSOD) [2, 3, 4, 13, 14, 19, 29], LOCL consists of a region proposal network with pseudo-label generation module that leverages supervision from linguistic embeddings in a novel contrastive framework. This feature extraction can be utilized to improve the existing network's performance in images with background clutter as shown in Table 3.

3 Approach

The primary issue to be addressed is the scene complexity, that require that the methods are able to make the correct associations during the training phase and predict the unseen configuration during testing. The proposed method is intuitive and straightforward in creating a weakly supervised framework that is modular and generalizes well. The LOCL extracts localized features of object regions in the image, which allows it to learn useful OA relationships and also suppress spurious O-A relations from the background clutter.

First, we pre-train a Localized Feature Extractor $LFE(.)$ (Sec. 3.1) network to extract features from multiple regions of the image. Second, the pre-trained $LFE(.)$ along with a Composition Classifier $CC(.)$ (Sec. 3.2) network learns to detect the O-A composition. *The*

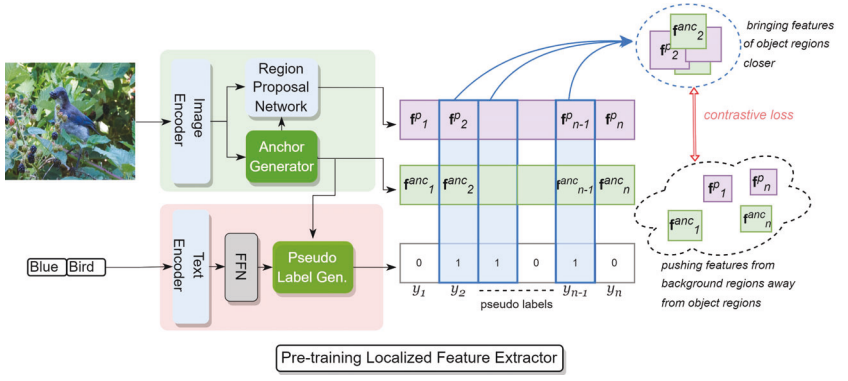

Figure 3: Summary of pre-training the localized feature extractor.The image encoder and region proposal are jointly trained to generate features of object of interest. During training time, we use text embeddings to generate pseudo labels to train the image encoder and region proposal using contrastive learning. At the test time, the learned image encoder and region proposal network are used to generate features from object regions.

key insight is to leverage the features from regions containing the object of interest to learn accurate O-A associations. Fig. 2 summarizes the overall LOCL architecture.

Problem Setting: Let $\{I, T_o, T_a, (a,o)\}$ be the training dataset with N samples, where I is the input image. T_o, T_a are the list of all object and attribute labels, respectively, and (a, o) is the tuple of attribute-object pairs in the image. The O-A pairs labels used during training are categorised as seen pairs. The goal of CZSL trained model is to take in an input image I and predict $(\hat{a}, \hat{o})$. The O-A pairs labels used during inference are novel and unseen. H*ere the seen and unseen object-attribute pairs are mutually exclusive.*

Proposed Network: The proposed LOCL is as $(\hat{a}, \hat{o}) = CC\ (\ LFE\ (I),\ T_a,\ T_o)$, where $LFE(.)$ and $CC(.)$ are trainable networks. LOCL is trained in two stages. In the first stage, given an image I, pre-training of $LFE(.)$ is done to generate multiple localized features. The details of the $LFE(.)$ module are discussed in Sec. 3.1. The output of the trained network $LFE(.)$ is a list of n features of object regions identified in the image.

In the second stage, out of these n features, r features ($r < n$) are input to the composition classifier (Sec. 3.2) to make the final prediction of attribute and object present in the image.

3.1 Pre-training Localized Feature Extractor($LFE(.)$)

Our Localized Feature Extractor network $LFE(.)$ is a combination of an image encoder (ResNet-50 [5]), text encoder, Region Proposal Network (RPN), and a pseudo label generator, as shown in Fig. 3. The RPN is inspired from F-RCNN [26]. It is trained from scratch using a contrastive learning framework. It generates proposals features for regions in the image that has high likelihood of object presence. The pseudo label generator creates labels to supervise the visual space that have high semantic similarity with ground truth O-A pair.

Given the input image I, the image encoder generates feature map $\mathbf{F} \in \mathbb{R}^{H' \times W' \times C}$ where H', W' and C are the height, width and channel dimensions. Then n valid anchors (based on input image size, in our case for an image of 256×256, $n = 576$,) are generated on the input image [26]. Anchors are a set of rectangular boxes with different aspect ratio and scale generated at **each pixel of the input image** [26]. Corresponding to each anchor, a list of

features is pooled from $\mathbf{F}$. The pooled anchor features are $[\mathbf{f}_1^{anc}, \mathbf{f}_2^{anc}, ..., \mathbf{f}_n^{anc}] \in \mathbb{R}^C$ shown as output of "Anchor Generator" in Fig. 3

The text encoder generates semantic pair embedding from the input text label $(a : Blue, o : Bird)$ pair. With the help of these semantic pair embedding, we generate pseudo ground truth labels to train $LFE(.)$ network with weak supervision [28]. In the following, we refer to "pseudo ground truth labels" as "pseudo labels" for simplicity.

Pseudo Label Generator: The ground truth O-A semantic pair embeddings generated from text encoder are projected through fully connected layers (FFN) into a common subspace as visual embeddings. The output of FFN is denoted by $\mathbf{f}_{ao}^{text} \in \mathbb{R}^C$, where "$ao$" index is the ground truth (in our case one per image). Here the length C of semantic embedding equal to channel dimension C of visual feature vector $\mathbf{F}$. Now to generate pseudo labels, a cosine similarity score is computed between each $\mathbf{f}_k^{anc}$ and $\mathbf{f}_{ao}^{text}$.

$$\phi_k = \frac{\mathbf{f}_{ao}^{text} \cdot \mathbf{f}_k^{anc}}{||\mathbf{f}_{ao}^{text}|| \, ||\mathbf{f}_k^{anc}||} \;\forall\; \mathbf{f}_k^{anc}, \; where \;\; (\phi = [\phi_1, \phi_2, .., \phi_k, .., \phi_n]) \tag{1}$$

$$\mathbf{y} = \begin{cases} 1 & argsort(\phi)[0:l] \\ 0 & for\ all\ other\ indexes \end{cases} \tag{2}$$

where $\mathbf{y} = [y_1, y_2, .., y_k, .., y_n]$, l top anchors are selected out of n based on cosine similarity score ϕ. They are assigned with label 1 in $\mathbf{y}$ and rest are assigned 0 as shown above with Eq. 2. Here each y_k represents the presence/absence of object of interest regions in the image. Intuition here is that $\mathbf{f}_k^{anc}$'s which contains the object will lie closer to $\mathbf{f}_{ao}^{text}$ in feature space.

Region proposal Network (RPN): The RPN branch shown in Fig. 3 is inspired from FasterRCNN [26]. The RPN generated proposals are used to pool a list of features from the feature map $\mathbf{F}$. The pooled features are $[\mathbf{f}_1^p, \mathbf{f}_2^p, ..., \mathbf{f}_n^p] \in \mathbb{R}^C$. Now the anchor features $[\mathbf{f}_1^{anc}, \mathbf{f}_2^{anc}, ..., \mathbf{f}_n^{anc}]$, proposal features $[\mathbf{f}_1^p, \mathbf{f}_2^p, ..., \mathbf{f}_n^p]$ and pseudo label $y = [y_1, y_2, .., y_n]$ are used to train the function $LFE(.)$ using contrastive learning as explained in next section.

Contrastive Pre-training: Recall that the current benchmark CZSL datasets [11, 17, 36] do not have ground truth bounding boxes for objects of interest. For this reason, we use both the anchor features and proposal features to localize the objects of interest. This is different from regular object detection networks [26]. The pseudo label $\mathbf{y}$ informs the network which anchor features are likely to represent object(s) of interest. Using contrastive learning, we train the regional proposal network branch to localize the object with weak supervision. The goal of contrastive learning is to maximize the similarity between similar feature vectors and minimize the similarity between the dissimilar feature vectors. Here, the objective is to maximize the cosine similarity between $< \mathbf{f}_k^{anc}, \mathbf{f}_k^p >$ features where there is a possibility of object being present and minimize in all other cases. The contrastive objective function is:

$$\mathcal{L}_{CON} = \sum_{k=1}^{n} (1 - y_k) * d_k^2 + y_k * \max(0, 1 - d_k^2), \tag{3}$$

where $*$ is element wise multiplication, y_k tells us which features have the possibility of having an object (Eq. 2), d_k is the cosine distance between $< \mathbf{f}_k^{anc}, \mathbf{f}_k^p >$.

$$d_k = \frac{\mathbf{f}_k^p \cdot \mathbf{f}_k^{anc}}{||\mathbf{f}_k^p|| \;\; ||\mathbf{f}_k^{anc}||} \; \forall \, k = [1, 2, ..n] \tag{4}$$

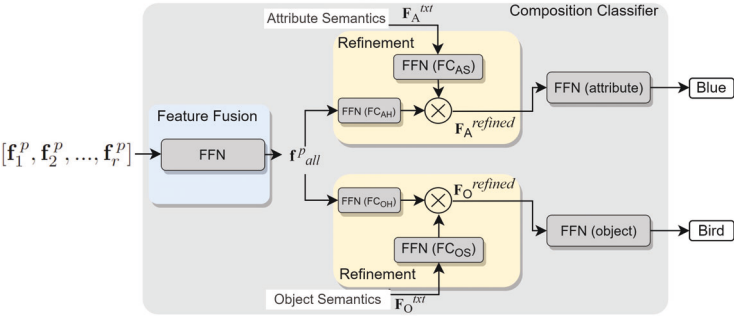

Figure 4: Composition Classifier $CC(.)$ architecture. The proposal features $[\hat{\mathbf{f}}_1^p, \hat{\mathbf{f}}_2^p, ..., \hat{\mathbf{f}}_r^p]$ are the outputs from $LFE(.)$ which are combined into a single representation $\mathbf{f}_{all}^p$. The attribute and object semantics are the semantic encoding of all attributes and objects under consideration. Two branches predict attribute and object from semantically refined $\mathbf{f}_{all}^p$.

Along with contrastive loss, we optimize binary cross entropy over the objectness score predicted by region proposal network. The overall loss function is:

$$\mathcal{L}_{total} = \alpha * \mathcal{L}_{CON} + \beta * \mathcal{L}_{BCE}(o, \phi), \tag{5}$$

where, α and β are empirically-determined scaling parameters, o is the objectness score from RPN and ϕ is the cosine distance from Eq. 1. Once the $LFE(.)$ network is trained, the output of trained model are the proposal feature vectors $[\hat{\mathbf{f}}_1^p, \hat{\mathbf{f}}_2^p, ..., \hat{\mathbf{f}}_n^p]$ along with objectness score $\hat{o} = [\hat{o_1}, \hat{o_2}, .., \hat{o_n}]$. This learnt parameter objectness score ensures selection of features with object information, thereby minimizing the interference from potential confusing elements as shown in Fig. 1

3.2 Composition Classifier $CC(.)$

The ability to learn individual representation of O-A in visual domain is crucial for transferring knowledge from seen to unseen O-A associations. Existing SOTA works [17, 18, 21, 27, 33] use homogeneous features from whole image as without localizing the object, they ignore the discriminative visual features of object and its attributes. Our Composition Classifier network $CC(.)$ leverages the distinctive features extracted by $LFE(.)$ to predict the object and its corresponding attribute as shown in Fig. 4. It is challenging to associate right attribute with the object by using homogeneous features, as there can be interference from prominent confounding elements like examples shown in Fig1B. Similarly in Fig. 5 (row-2, column-1 image), the attribute "*green color*" is so prominent, and using homogeneous features can lead to wrong O-A prediction.

The input to $CC(.)$ is a set of top r ($r < n$) proposal feature vectors from pre-trained $LFE(.)$ $[\mathbf{f}_1^p, \mathbf{f}_2^p, ..., \mathbf{f}_r^p] \in \mathbb{R}^{r \times C}$ sorted in descending order based on objectness score $o = [o_1, o_2, .., o_r]$. The proposal feature vectors are fused into a single visual feature $\mathbf{f}_{all}^p \in \mathbb{R}^{1 \times C}$ using weighted average with learnable parameters [15] as shown in Fig. 4. Input to the learnable parameters has $r \times C$ dimension. r is the number of proposals and C is the number of channels. We swap the dimensions at the input to fuse different proposals together. The output of it is a single feature vector of length C. The fusion operation is a learnable weighted average operation, that learns to create joint representation from features of object regions. Next, $\mathbf{f}_{all}^p$ is projected to two different representation via two feed forwards networks FC_{AH} and FC_{OH} layers as shown in Fig. 4. Similarly, the semantic embeddings of all attributes

	# Images			# Objects	#Attributes	# OA Pairs
	Train	Val	Test			
MIT-States [20]	30k	10k	13k	245	115	1962
UT-Zappos [36]	23k	3k	3k	12	16	116
CGQA [21]	26k	7k	5k	870	453	9378

Table 1: Comparison of different CZSL datasets [20, 21, 36]

$\mathbf{F}_A^{text} = [\mathbf{f}_{a1}^{text}, \mathbf{f}_{a2}^{text}.., \mathbf{f}_{ai}^{text}]$ and all objects $\mathbf{F}_O^{text} = [\mathbf{f}_{o1}^{text}, \mathbf{f}_{o2}^{text}.., \mathbf{f}_{oj}^{text}]$ (generated from T_a and T_o) are projected via feed forward network FC_{AS} and FC_{OS}. Here T_a and T_o are the list of all attributes and objects in the dataset respectively, and $i = len(T_a)$, $j = len(T_o)$. $len(.)$ is length. Inspired by [9, 10, 16, 30, 31], the visual feature projections are refined as shown below. The particular choice of refinement by one-to-one multiplication is an empirical choice. We explore different refinement techniques in Table 3 of supplementary materials. This all is done by *Refinement* block as shown in Fig. 4.

$$\mathbf{F}_O^{refined} = FC_{OH}(\mathbf{f}_p^{all}) \circ FC_{OS}(\mathbf{F}_O^{txt}) \;\; ; \;\; \mathbf{F}_A^{refined} = FC_{AH}(\mathbf{f}_p^{all}) \circ FC_{AS}(\mathbf{F}_A^{txt}) \tag{6}$$

where, "$\circ$" represents element-wise multiplication and the feed forward network are two fully connected layers with ReLu activation. This refinement aggregates semantic information with the visual features. These refined features are passed through another feed forward network (Fig. 4) and softmax layers to make the final decision on the object $\hat{o}$ and attribute $\hat{a}$ present in the image I. To train the composition classifier function $CC(.)$, we optimize binary cross entropy function over the O-A prediction.

4 Experiments

Table 1 summarizes the datasets used. In the MIT-states dataset the images are of natural objects collected using an older search engine with limited human annotation causing a significant label noise [1]. For UT-Zappos [36] the simplicity of the images (one object with white background) makes it unsuitable to work in natural surroundings. We performed experiments on very recently released Compositional GQA (CGQA) dataset [6, 17]. This dataset is proposed to evaluate existing CZSL models in a more realistic challenging scenarios with background clutter. More details about the datasets can be found in supplementary materials section 4.

Both the localized feature extractor $LFE(.)$ and Composition Classifier $CC(.)$ are trained on all the datasets. To pre-train $LFE(.)$, an efficient contrastive pre-training framework is used with a margin distance of 1. Then this pre-trained network is used with $CC(.)$ to do an end-to-end training of LOCL. We have provided the complete implementation details in the supplementary materials. However, during inference time, $LFE(.)$ do not have any text embedding branch, hence it generates proposals for every potential object as shown in Fig. 6.

Following current methods [16, 21, 24], we evaluate our network's performance in Generalized Compositional Zero Shot Learning Protocol (GCZSL). Under this protocol, we draw seen class accuracy vs unseen class accuracy curve at different operating points of the network. These operating points are selected from a single calibration scaler that is added to our network's predictions of the unseen classes [16, 21, 24]. We report area under "seen class accuracy vs unseen class accuracy curve" (AUC) as our performance metrics. Additionally, we report our network's performance on Top-1 accuracy in seen and unseen classes.

Methods	MIT-States [11]			UT-Zappos [36]			CGQA [21]		
	Seen	Unseen	AUC	Seen	Unseen	AUC	Seen	Unseen	AUC
Attop [22]	14.3	17.4	1.6	59.8	54.2	25.9	11.8	3.9	0.3
LabelEmbed [20]	15	20.1	2.0	53.0	61.9	25.7	16.1	5	0.6
TMN [24]	20.2	20.1	2.9	58.7	60.0	29.3	21.6	6.3	1.1
SymNet [16]	24.2	25.2	3.0	49.8	57.4	23.4	25.2	9.2	1.8
CompCos [18]	25.3	24.6	4.5	59.8	62.5	28.1	28.1	11.2	2.6
ProtoProp [27]	-	-	-	62.1	65.5	34.7	26.4	18.1	3.7
BMP-Net [35]	**38.6**	21.7	6.0	**87.3**	64.5	**49.7**	-	-	-
CGE [21]	32.8	28.0	6.5	64.5	71.5	33.5	**31.4**	14	3.6
LOCL (Ours)	35.3	**36.0**	**7.7**	68.0	**76.7**	37.9	29.6	**26.4**	**4.2**

Table 2: Performance comparisons on MIT-States [11], UT-Zappos [36], CGQA [21] Datasets. '-' means unreported performance in a particular category. In all three datasets, LOCL significantly outperform current methods. Specially, for the more challenging (significant background clutter) CGQA dataset, the effectiveness of LFE is clearly demonstrated by its performance on the unseen O-A associations.

4.1 Results

LOCL outperforms current methods in the test set of all benchmark datasets in almost all categories as shown in Table 2. We evaluate LOCL's performance in terms of AUC under Generalized Compositional Zero Shot Learning (GCZSL) protocol. We also report Top-1 accuracy in seen and unseen classes and accuracy in detecting objects and attributes. In MIT-States [11] LOCL outperforms SOTA method by 8% on unseen class accuracy and 1.7% AUC. In UT-Zappos [36] LOCL's unseen class accuracy is 5.2% better than the SOTA method. Moreover, it almost *doubles* the unseen class accuracy while achieving 1.1% improvement in terms of AUC for the more challenging CGQA [21] dataset.

Current CZSL methods use homogeneous features from the backbone instead of using distinctive visual features of objects and attributes. While such techniques may work on simpler datasets like UT-Zappos [36], as evidenced by the high performance, the more realistic datasets such as CGQA pose challenges. Table 2 shows, LOCL achieves the best results on the challenging CGQA dataset. However, bias towards seen O-A compositions is a common issue [24] in current CZSL methods. Recent approaches [21, 35] have utilized a graph structure with message passing/blocking [35] or prior possible O-A knowledge [21] to reduce this bias. However they tend to be biased towards seen O-A pairs at inference as pointed out by the authors of [35]. In contrast, LOCL learns distinct object and attribute representations in the two separate branches of the $CC(.)$ and achieves high unseen class accuracy and AUC. In UT-Zappos, high AUC of [35] stems from high seen class accuracy with inferior unseen class performance. [35] do not evaluate their model CGQA dataset [21].

4.2 Ablation Study

Backbone and Localized Feature Extractor: Our Localized Feature Extractor $LFE(.)$ is modular and can easily be adapted to other methods. In Table 3, we show different SOTA methods' improved performances with our feature extraction. For the sake of fair comparison with SOTA methods: SymNet [16], ComCos [17] and CGE [21], we replace their backbones with our ResNet-50 pre-trained on a larger dataset [25]. As expected, both our backbone and $LFE(.)$ improve the existing networks' performances. This shows that the performance im-

Methods	Our BB	LFE	CGQA [21]			MIT-States [11]		
			Seen	Unseen	AUC	Seen	Unseen	AUC
SymNet [16]	✗	✗	25.2	9.2	1.8	24.2	25.2	3.0
	✓	✗	25.3	9.3	1.8	26.6	26.1	3.5
	✓	✓	27.7	13.5	2.0	28.7	27.7	3.8
CompCos [17]	✗	✗	28.1	11.2	2.6	25.3	24.6	4.5
	✓	✗	28.4	13.5	2.8	25.6	24.8	4.5
	✓	✓	28.9	16.7	2.9	27.9	26.7	5.1
CGE [21]	✗	✗	31.4	14.0	3.6	32.8	28	6.5
	✓	✗	31.4	19.3	3.8	33.3	28	6.5
	✓	✓	**31.9**	26.1	4.1	**36.3**	29.8	6.6
LOCL	✓	✓	29.6	**26.4**	**4.2**	35.3	**36.0**	**7.7**

Table 3: Performance of SOTA methods with our backbone (BB) and LFE. LFE significantly boosts all SOTA network performances specially in the CGQA dataset. Row 2 if each methods shows performance of all SOTA models with a common and better backbone.

provement is because of localized features generated from $LFE(.)$. All existing methods use ResNet-18 as the backbone following the seminal work of [20]. We recommend that CZSL networks should utilize stronger backbones for challenging datasets like CGQA [21]. However, our improved performance is not just coming from a stronger backbone. With $LFE(.)$, the performance boost is more significant (specially in terms of unseen class accuracy) than the performance boost with our backbone for the CGQA dataset. In particular, $LFE(.)$ increases the unseen class accuracy of CGE [21] in CGQA by *86%*, and other methods also get great improvement with $LFE(.)$. The performance improvement in the MIT-States dataset is less due to the noisy annotations [1] of this dataset. In summary, $LFE(.)$ improves three different architectures thus proving the effectiveness of localized feature extraction.

We additionally ablate LOCL's performance for different *number of proposals*, *number of pseudo labels*, *refinement techniques*, *margin distance for contrastive loss*, and *scaling parameters* α *and* β. All these experiments along with LOCL's performance on detecting individual objects and attributes are reported in the supplementary material.

Qualitative Results: We show results for unseen novel compositions with top-1 prediction in Fig. 5. They represent the scenes with clutter or confounding elements. For example, column-1 shows where the confounding element *color* attribute $< green >$ causes wrong O-A association for most of the SOTA methods while LOCL makes the right association. Similar in the first row. The last column shows where our network predictions do not match with the ground truth labels. In the top image our network focuses more on prominent object *clock*, while the image is labeled for $< red, bus >$. For row-2 image, it contains attribute categories like *size*, *color*, *texture* but the label only has attribute *size*. We should, however, note that the ground truth labels in these datasets contain only one O-A pair. This puts an artificial limitation on the evaluation metric even when the predictions are *perceivably correct* but does not match labels. We also show $LFE(.)$ proposals quality in eliminating background in Fig. 6

Multi O-A prediction: We extended LOCL to unconstrained setting than existing evaluation methods allows, i.e. detecting multiple O-A pairs. LOCL has flexibility of establishing the right O-A association for multiple objects in the scene. We are showing the top-3 O-A pairs prediction in Fig. 5 last row. As can been seen the prediction are perceivably correct but unavailability of ground-truth annotation in existing datasets puts a limit on quantitative

Figure 5: Qualitative results of LOCL. Row-1 & 2 (col-1,2,3)show correct predictions. Col-4 shows missed predictions, ground truth labels are marked with green box and our predictions in red box. The datasets has one O-A pair per image. Though our predictions are visually correct, do not match the ground-truth. This puts an artificial limit on the evaluation metric. Row-3 shows multiple O-A detections

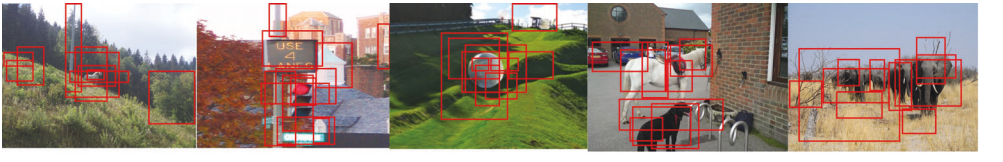

Figure 6: Proposals selected based on objectness score. We can see that the proposals are generated on the object of interest. Though LOCL is not designed for multi O-A, but in case of multiple objects, the proposals are distributed over multiple objects.

evaluation. Also, there can be multiple correct attribute related to one object as shown in the two central images in Fig. 5 last row. $< White >, < Standing >$ both are correct attribute for object $< Horse >$, similarly $< Dark >, < Large >$ are correct attributes for object $< Elephant >$. Multiple OA pairs detection is an interesting direction for future research.

5 Conclusion

We present a novel two-step approach, LOCL, for recognizing O-A pairs. Our approach includes a robust local feature extractor followed by a composition classifier. LOCL is evaluated on benchmark datasets. Additionally, our experiments show that the local feature extractor improves the performance of current SOTA CZSL models by a significant margin of 12%. The code and trained model will be made available on Github at the time of publication.

6 Acknowledgement

The authors would like to thank Dr. Suya You, Army Research Laboratory (ARL), for the many discussions that contributed to the methods development. We also thank Mr. Pushkar

Shukla from Toyota Technological Institute, Chicago and Mr. Rahul Vishwakarma from Hitachi Labs America for their suggestions and critical reviews of the paper. This research was in part supported by NSF SI2-SSI award #1664172 and by US Army Research Laboratory (ARL) under agreement # W911NF2020157. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of US Army Research Laboratory (ARL) or the U.S. Government

References

[1] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1462–1473. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1010cedf85f6a7e24b087e63235dc12e-Paper.pdf.

[2] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 914–922, 2017.

[3] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2): 31–71, 1997.

[4] Nicolas Gonthier, Saïd Ladjal, and Yann Gousseau. Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts. *arXiv preprint arXiv:2008.01178*, 2020.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[7] Dat Huynh and Ehsan Elhamifar. Compositional zero-shot learning via fine-grained dense feature composition. *Advances in Neural Information Processing Systems*, 33: 19849–19860, 2020.

[8] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4483–4493, 2020.

[9] ASM Iftekhar, Satish Kumar, R Austin McEver, Suya You, and BS Manjunath. Gtnet: Guided transformer network for detecting human-object interactions. *arXiv preprint arXiv:2108.00596*, 2021.

[10] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5353–5363, 2022.

[11] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015.

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[13] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *European Conference on Computer Vision*, pages 350–365. Springer, 2016.

[14] Satish Kumar, Carlos Torres, Oytun Ulutan, Alana Ayasse, Dar Roberts, and BS Manjunath. Deep remote sensing methods for methane detection in overhead hyperspectral imagery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1776–1785, 2020.

[15] Satish Kumar, ASM Iftekhar, Michael Goebel, Tom Bullock, Mary H MacLean, Michael B Miller, Tyler Santander, Barry Giesbrecht, Scott T Grafton, and BS Manjunath. Stressnet: detecting stress in thermal videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 999–1009, 2021.

[16] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11316–11325, 2020.

[17] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. *arXiv preprint arXiv:2105.01017*, 2021.

[18] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5222–5230, 2021.

[19] R Austin McEver, Bowen Zhang, Connor Levenson, ASM Iftekhar, and BS Manjunath. Context-driven detection of invertebrate species in deep-sea video. *arXiv preprint arXiv:2206.00718*, 2022.

[20] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017.

[21] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021.

[22] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018.

[23] Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. Recognizing unseen attribute-object pair with generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8811–8818, 2019.

[24] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3593–3602, 2019.

[25] 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. *arXiv preprint arXiv:2103.00020*, 2021.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[27] Frank Ruis, Gertjan Burghours, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. *arXiv preprint arXiv:2106.00305*, 2021.

[28] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[29] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104 (2):154–171, 2013.

[30] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13617–13626, 2020.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[32] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[33] Guangyue Xu, Parisa Kordjamshidi, and Joyce Y Chai. Zero-shot compositional concept learning. *arXiv preprint arXiv:2107.05176*, 2021.

[34] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[35] Ziwei Xu, Guangzhi Wang, Yongkang Wong, and Mohan S Kankanhalli. Relation-aware compositional zero-shot learning for attribute-object pair recognition. *IEEE Transactions on Multimedia*, 2021.

[36] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5570–5579, 2017.

[37] Xiangyun Zhao, Yi Yang, Feng Zhou, Xiao Tan, Yuchen Yuan, Yingze Bao, and Ying Wu. Recognizing part attributes with insufficient data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 350–360, 2019.