In Proceedings of the European Conference on Computer Vision (ECCV), 2018

Attributes as Operators (Supplementary Material)

This document consists of supplementary material to support the main paper text. The contents include:

- Further analysis of the problems with the closed world setting (from Section 4.2 in the main paper) in the context of the MIT-States dataset.
- Architecture details of our LABELEMBED+ baseline model proposed in Section. 4.1 (Baselines) of the main paper.
- Variants of baseline models that add our proposed auxiliary regularizer (from Section 3.3).
- TSNE visualization of the joint semantic space described in Section 3.1 learned by our method.
- Our procedure to obtain the subset of UT-Zappos50K described in Section 4.1 (Datasets) of the main paper.
- Additional training details including hyper-parameter selection for all experiments in Section 4.2 of the main paper.
- Additional qualitative examples for retrieval on ImageNet from Section 4.3 of the main paper.

Attribute Affordances: Open vs. Closed world

As discussed in Section 4.2 of the main paper, recognition in the closed world setting is considerably easier than in the open world setting due to the reduced search space for attribute-object pairs.

Figure 1 highlights this difference for the MIT-States dataset. In the open world setting, each object would have many potential candidates for compositions (red region), but in the closed world case (blue region), this shrinks to a fraction of compositions. Overall this translates to a 2.8$\times$ higher chance of randomly picking the correct composition in the closed world. To make things worse, about 14% of the objects occur in the test set compositions that are dominated by a single attribute. For example, "tiger" affords 2 attributes, but one of those occurs in a single image, leaving *old tiger* as the only relevant composition. A model with poor attribute recognition abilities can still get away with a high accuracy in terms of the "closed world" setting as a result, giving a false sense of performance.

Previous work has focused only on the closed world setting, and as a result, compromised on the ability to perform well in the open world (Table 1 in the main text). Our model performs similarly well in both settings, indicating that it does not become dependent on the (artificially) simpler setting of the closed world.

LABELEMBED+ Details

In Section 4.1 (Baselines) of the main paper, we propose the LABELEMBED+ baseline as an improved baseline model, improving the LABELEMBED baseline presented in

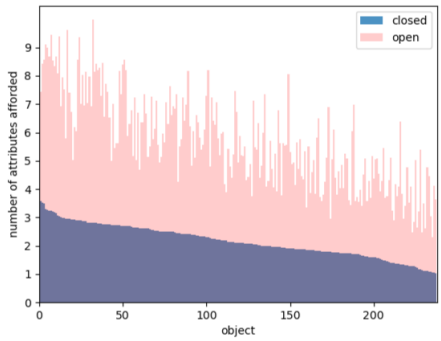

Fig. 1: Attribute affordances for objects. The closed world setting is easier overall due to the reduced number of attribute choices per object. In addition, about 14% of the objects are dominated by a single attribute affordance.

	MIT-States			UT-Zappos		
	closed	open	h-mean	closed	open	h-mean
LabelEmbed+ (1)	**14.9**	5.8	8.3	36.1	5.3	9.2
LabelEmbed+ (2)	**14.9**	5.7	8.2	37.4	9.4	15.0
LabelEmbed+ (3)	14.3	5.3	7.7	37.6	7.7	12.8
Ours	12.0	**11.4**	**11.7**	**38.1**	**29.7**	**33.4**

Table 1: Model capacity of baseline methods. The LABELEMBED+ baseline model with increasing model capacity (number of layers shown in brackets). Our model outperforms this baseline regardless of how many layers are involved, suggesting that model capacity is not the limiting factor.

the REDWINE paper by Misra et al. We present the details of the architecture of this baseline here. We use a two layer feed-forward network. Specifically, we concatenate the two primitive input representations of dimension D each, and pass it through a feed-forward network with the configuration (linear-relu-linear) and output dimensions $2D$ and D. Unlike REDWINE, we transform the image representation using a single linear layer, followed by a ReLU non-linearity. We do this to allow some flexibility on the side of the image representation (since we are not finetuning the network responsible for generating the image features themselves).

Here we report additional experiments where we vary the number of layers for this baseline (Table 1). We see that our model outperforms LABELEMBED+ regardless of how many layers are used. This suggests that our improvements are a result of learning a better composition model, and is not related to the network capacity of the baseline model.

	Original			Augmented Baselines		
	closed	open	h-mean	closed	open	h-mean
REDWINE	12.5	3.1	5.0	**12.7**	3.2	5.1
LABELEMBED	13.4	3.3	5.3	12.2	3.1	4.9
LABELEMBED+	**14.9**	5.7	8.2	8.1	7.4	7.7
OURS	12.0	**11.4**	**11.7**	12.0	**11.4**	**11.7**

Table 2: Baseline variants on MIT-States. The proposed auxiliary loss term, together with allowing attribute and object representations to be optimized during training, can also help the baselines learn a better composition model. Our complete model outperforms these baseline variants as well.

Baselines Variants

Next we include modifications to the baselines presented in Section 4.1 of the main paper that are inspired by components of our own model. Specifically, we allow trainable inputs and include the proposed auxiliary regularizer from Section 3.3.

Table 2 shows the results. The first column denotes the models as reported in the main paper, while the second column shows the models with extra components from our own model. Note that our model already includes these modifications, and we simply repeat its results on the "augmented" side for clarity. REDWINE and LABELEMBED are not severely affected because of the way the composition is interpreted—as a set of classifier weights. Extracting the attribute and object identity from *classifier weights* is less meaningful compared to extracting them from a general composition embedding. The auxiliary loss does however improve the embedding learning models. Our model outperforms all augmented variants of the baselines as well.

Visualization of Composition Space

We visualize the common embedding space described in Section 3.1. Figure 2 contains the 2D TSNE projection of the 300D space generated by our method. The black points represent the embeddings of all unseen compositions in MIT-States. Each cluster (squared) represents the *span* of a single attribute operator—*i.e.*, the points in the vector-space that it can reach by transforming object vectors. Our composition model maintains a clear separation between several attribute-object compositions, despite many sharing the same object or attribute.

We also highlight three object superclasses, "fruit", "scenes" and "clothing", and plot all the compositions they are involved in to show which parts of this subspace are shared among different object classes. We see that common attributes like *old* and *new* are shared by many objects of each superclass, while more specialized attributes like *caramelized* for "fruit" are separated in this space.

UT-Zappos Subset Selection

As discussed in Section 4.1 (Datasets) of the main paper, we use a subset of the publicly available UT-Zappos50K dataset in our experiments. The attributes and annotations typically used in this dataset are *relative attributes*, which are not relevant for our

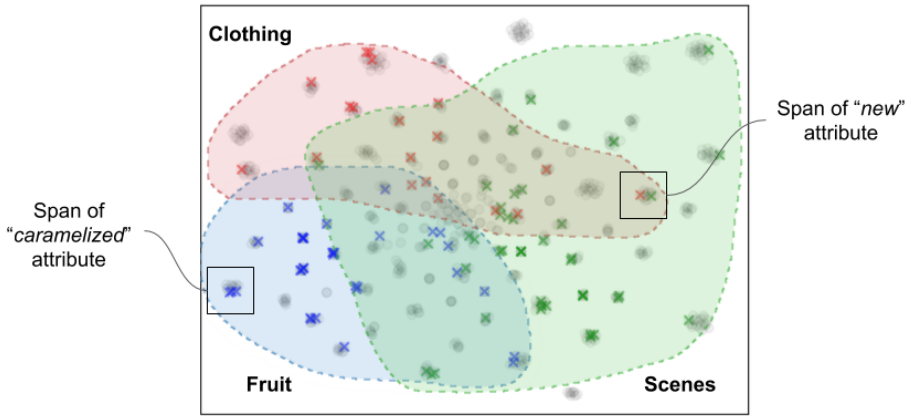

Fig. 2: TSNE visualization of our common embedding space. The span of an attribute operator represents the extent of its application to all objects that afford the attribute. Here each cluster of black points represents a single attribute's span. We see a strong separation between compositions in this semantic space, despite these compositions sharing attributes or objects. We highlight composition clusters for three common object superclasses, "fruit", "scenes" and "clothing". Each colored 'x' corresponds to a single composition involving an object from the respective superclass. For example, a blue 'x' is a composition of the form (attr, fruit) where *attr* is one of (peeled, diced, etc.) and *fruit* is one of (apple, banana, etc.).

experiments and are not applicable for comparisons to existing work. However, it also contains labels for *binary material attributes* that are relevant for our experiments.

Here we describe the process for generating the subset of UT-Zappos50K that we use in our experiments. These images have top level object categories of shoe type (*e.g.*, *high heel*, *sandal*, *sneaker*) as well as finer-grained shoe-type labels (*e.g.*, *ankle boots*, *knee-high boots* for the top-level *boots* category). We merge object categories that have fewer than 200 images per category into a single class (*e.g.*, all slippers are considered as one class), and discard the sub-classes that do not meet this threshold amount. We then discard all the images that do not have annotations for material attributes of shoes (*e.g.*, *leather*, *sheepskin*, *rubber*), which leaves us with $\sim$33K images. We randomly split this set of images into training and testing sets based on their attribute-object compositions. Our subset contains 116 compositions, over 16 attribute classes and 12 object classes.

Additional Training Details

We provide additional details to accompany Section 4.1 (Implementation Details) in the main paper. For our combined loss function, we take a weighted sum of all the losses, and select the weights using a validation set. We create this set for both our datasets by holding out a disjoint subset of 20% of the training pairs.

– For MIT-States, we train all models for 800 epochs. We set the weight of the auxiliary loss L_{aux} to 1000 for our model.

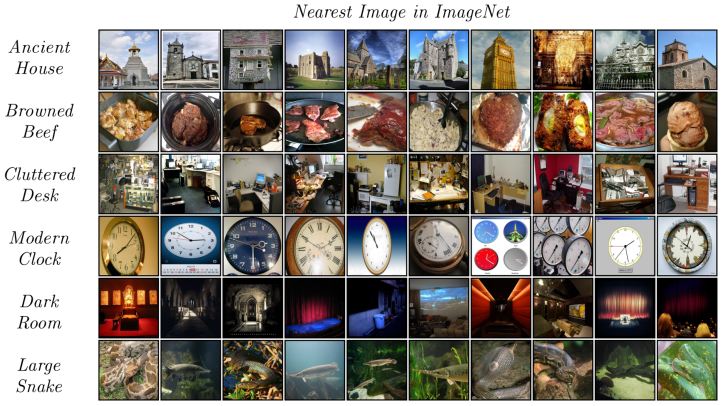

Fig. 3: Retrieval results on ImageNet images. Text queries of unseen compositions with top-10 image retrievals shown alongside. Note that the compositions are learned from a disjoint set of compositions on a disjoint dataset (MIT-States), then used to issue queries for images in ImageNet.

– For UT-Zappos, we train all models for 1000 epochs. We weight all regularizers equally.

The weight for L_{aux} is substantially higher for MIT-States, which may be necessitated by the low volume of training data per composition. On both datasets, we train our models with a learning rate of $1e-4$ and a batch size of 512.

Additional Qualitative Examples

Next we show additional qualitative examples from Section 4.3 for the unseen compositions. Figure 3 shows retrieval results on a diverse set of images from ImageNet, where object and attribute categories do not directly align with MIT-States. These examples are computed and displayed in the same manner as Figure 4 in the main paper.