PMB: Compositional Attribute-object Understanding with Pronouns

Runyi Yang[1,2] Zirui Wu[1,3] Yongliang Shi[1] Xin Wu[5] Guyue Zhou[1]
Yurong Chen[4] Hongbin Zha[5] Hao Zhao[1] *

Abstract

Deep neural networks, as highly non-linear end-to-end models, still struggle to recognize compositional attribute-object pairs in a zero-shot manner. State-of-the-art methods leverage pre-trained language models to generate regression targets so that the embeddings are better anchored in the feature space. However, we note that current text encoder outputs are not regularized and thus may lose the rich structure. To this end, we introduce pronouns so that regression targets are augmented from adjectives (e.g., running) to adjective-pronoun pairs (e.g., running something). Meanwhile, we design a first-in-first-out memory bank for every and each attribute/object, which intrinsically regularizes the regression target. We evaluate our framework on three large-scale datasets: MIT-States, UT-Zappos, and VAW-CZSL, demonstrating clear improvements. Codes, data, and models will be made publicly available.

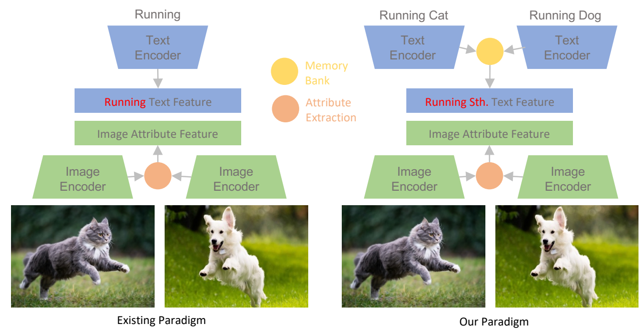


Figure 1. **Left:** Existing methods generate the text attribute, object and pair embeddings by utilizing multiple sub-networks (e.g. of `running` with a text attribute encoder). An attribute extraction module generates image attribute embedding of `running` from two images. **Right:** Our proposed method regularizes the output from the text embedding. We use a text encoder to generate the pair embedding like `running cat` and maintain a memory bank to output text attribute embedding `running something`.

1. Introduction

Closed-set visual recognition [38] has seen large progress since the advent of deep learning. However, deep neural networks, which are highly nonlinear due to a large stack of non-linearly activated modules, may learn spurious correlations that lead to confident wrong predictions on certain samples [19]. This limitation is better shown by its limited success in the zero-shot compositional attribute-object understanding setting. As shown in Fig. 1, if two images of `running cat` and `running dog` are presented to a deep neural network, it might learn the pattern of `four-leg animals in the air` for the concept `running`. If this spurious correlation is established through an uninterpretable nonlinear mapping, recognizing brand new composition (e.g. `running man`) in the zero-shot setting becomes challenging.

So a recent state-of-the-art method [39] proposes to leverage the rich semantic structure hidden in the large-scale pre-trained language models. As shown in the left panel of Fig. 1, the text encoder generates the regression target for `running`. An attribute extraction module would produce an image attribute embedding from those two images. Enforcing the text and image embeddings to be closer facilitates zero-shot recognition as one can input newly composed texts (e.g. `running man`) into the text encoder during test time.

However, we identify a limitation of this existing paradigm: because both text and image encoders are optimized during training, the rich structure of the pre-trained language model may be broken after convergence. In other words, the outputs of text encoders are not regularized. To this end, we propose the idea of **Pronoun Memory Bank** or **PMB**. Specifically, the text inputs are augmented from `running` to `running cat/dog`. We maintain a memory bank that corresponds to the text features of all seen pairs of `running something`. A temporally averaged version of this `running something` text feature bank functions as the regression target in our method. As in other memory-based methods like MoCo [7] or mean teachers [41], our PMB imposes regularization to the regression

*Corresponding Author
1 Institute for AI Industry Research, Tsinghua University
2 Imperial College London, yang.yang23@imperial.ac.uk
3 System Thrust, HKUST(GZ) 4 Intel Labs 5 Peking University

target so that better scalability can be achieved. Although not shown in Fig. 1, object regression targets (e.g. `cat`) are also generated by the average of a queue `someadj. cat` in the memory bank consisting of the text features of `running cat`, `sleeping cat` and others.

To summarize, we have the following contributions:

- We propose a new framework named pronoun memory bank or PMB for zero-shot compositional attribute-object understanding. First-in-first-out memory banks generate averaged regression targets for each attribute or object, as an effective regularization.
- We evaluate our PMB method on public benchmarks MIT-States, UT-Zappos, and VAW-CZSL and achieve state-of-the-art results with a great margin. Codes, data, and models will be released.

2. Related Work

Visual Attribute. Visual attributes have been widely used in understanding visual properties of objects. As a middle-level concepts, visual attributes is used to describe objects [39], human faces [22], scenes [30, 37], human activities [30], which benefit many downstream tasks of computer vision, such as recognition [5], image retrieval [47], semantic representation [48].

Attribute-augmented semantic hierarchy bridges gap between semantics and intention retrieval [47], therefore, visual attribute is regarded as cue to discover and model the intra-concept visual variance for learning extensive models within any concept [5]. Parikh et al. [29] firstly model relative attributes to learn a ranking function for each attribute that indicates the relative strength of the attribute presence in them. Following the formulation, a set of ranking functions are learned to facilitate the interactive image search [14]. In order to recognize unseen objects, Nagarajan et al. [27] model attributes as operators to learn a semantic embedding that explicitly factors out attributes from their accompanying objects.

For attribute research, a variety of datasets are developed. Patterson et al. [32] discover and annotate visual attributes for the COCO dataset for deeper object understanding. Transient attribute database [15] is created for high-level understanding and editing. SUN Attribute Database [31] is the first large-scale scene attribute database. Pham et al. [34] introduce a in-the-wild visual attribute prediction dataset, and describe a multitude of attributes which portray their visual appearance, geometry, and other intrinsic properties.

Zero-shot Learning. Given high-level semantically meaningful attributes [3] and textual descriptions [17] of seen object classes, Zero-shot Learning (ZSL) aims to complete relevant downstream tasks including recognition and visual search, etc. With ideas from manifold learning, Changpinyo et al. [2] introduce a set of "phantom" object classes to align the semantic space to the model space that concerns itself with recognizing visual features. Natural language offers a general and flexible interface for describing objects in visual attribute space, so vision and language are combined in ZSL [28] to represent object and attribute as linguistic word embedding vectors to recognize unseen attribute-object pair. Besides, [9] compose sentences that describe novel objects and their interactions with other objects. To evaluate ZSL approaches, Chao et al. [3] develop a performance metric called the Area Under Seen-Unseen accuracy Curve. In light of this, Liu et al. [21] propose a Deep Calibration Network (DCN) to map visual features of images and semantic representations of class prototypes to a common embedding space.

Compositional Zero-shot Learning. Unlike ZSL, Compositional Zero-Shot Learning (CZSL) entails that the model learns to compose unseen concepts from primitive components that have already been learned [23]. Most approaches to CZSL learn the embedding of object-attribute pair in image feature space [25], and require hundreds of training examples, while Purushwalkam et al. [36] propose task-driven modular networks to learn the joint compatibility between the input image and the pair by learning a representation. To exploit rich dependency structure of different states, objects and their compositions, Naeem et al. [26] propose the Compositional Graph Embedding (CGE) that learns image features, compositional classifiers and latent representations of visual primitives in an end-to-end manner. Compositional Cosine Graph Embeddings (Co-CGE) [24] use the score of unseen composition as margins in a cosine similarity-based loss and as weights in the adjacency matrix of the graphs. OADis [39] utilizes auxiliary networks to explicitly focus on separating attributes and object features in the visual space, and achieved state-of-the-art performance.

3. Compositional Attribute-Object Understanding with Pronouns

Previous methods (e.g. OADis [39]) apply learnable sub-networks onto embeddings generated by language models and the outputs of these sub-networks serve as regression targets, as shown in Fig. 3 (a). As as those MLPs shown in Fig. 3 (a) are not fixed, the regression targets change from iteration to iteration. Thus, the networks from the image part are optimized toward inconsistent targets. This inconsistency severely challenges the scalability of the image encoding network. To this end, we propose to use memory banks that smooth the regression targets over time so they are more consistent in each iteration, as shown in Fig. 3 (b).

We illustrate our overall system in Sec. 3.2. Pronoun Memory Bank is introduced in Sec. 3.3. The basic visual components are introduced in Sec. 3.4. We quantitatively

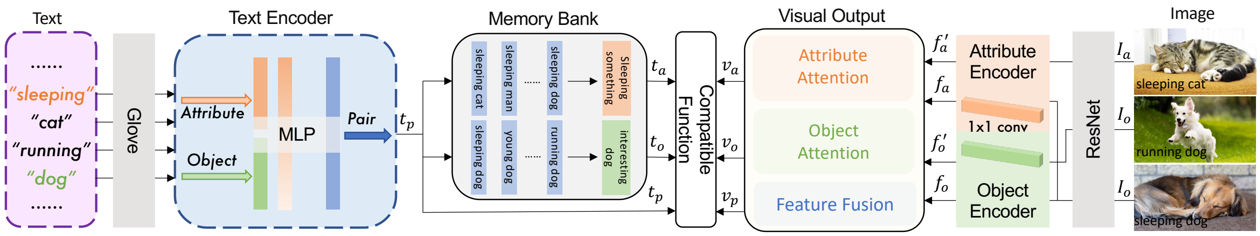
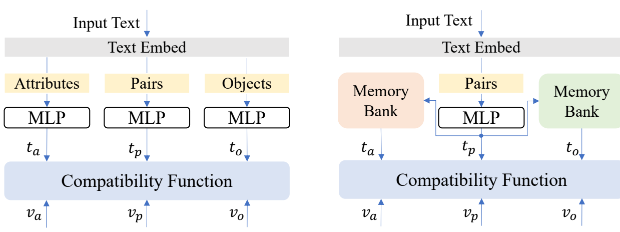


Figure 2. **System Overview:** Given three images, we use pretrained image encoding backbone to extract features and utilize attribute and object encoders to generate visual features $f_a, f_o, f_a^{'}$ and $f_o^{'}$. The final visual embeddings are computed by feature fusion model and attention models. After preprocessing text labels by pretrained word embedding, text features of attributes and objects are composed to text pair features by the Text Encoder. Pronoun Memory Bank is proposed to represent text features of attributes/objects with pronouns. Thus, vision embeddings v_p, v_a, v_o and text embeddings t_p, t_a, t_o are compatible.



(a) OADis (b) PMB (Ours)

Figure 3. Comparison of OADis and our proposed methods (PMB) from the text branch.

demonstrate that our Pronoun Memory Bank design can significantly improve the scalability of the image branch in Sec. 4, using a larger image encoding backbone leads to stable performance improvements on three datasets while it is not the case for OADis.

3.1. Task Formulation

Compositional zero-shot learning (CZSL) [25] aims to recognize the novel compositional labels that are not observed during training. This is particularly challenging because different attributes can drastically change the visual appearance of an object, making it difficult for classifiers to identify it accurately. In this task, all attribute labels $\mathcal{A}$ and object labels $\mathcal{O}$ compose a label space domain $\mathcal{T} = \{(a_i, o_j)|a_i \in \mathcal{A}, o_j \in \mathcal{O}\}$ which contains all pair labels. Each pair label $p_i \in \mathcal{P}$ is a composition of attribute $a_j \in \mathcal{A}$ and $o_k \in \mathcal{O}$. Since not all pair label makes sense, such as `flying cheese`, the pair label space is a subset of text label space $\mathcal{P} \subset \mathcal{T}$.

Given an image I_i corresponding to a pair label p_i, the train set is denoted by $\mathcal{S}_t = \{(I_i, p_i)|I_i \in \mathcal{I}_t, p_i \in \mathcal{P}_s\}$, where $\mathcal{I}_t$ contains all images for training, and seen pairs $\mathcal{P}_s$ is a subset of $\mathcal{P}$. The target of CZSL task is to train a model $\mathcal{M} : \mathcal{I} \rightarrow \mathcal{P}$, enabling to predict both seen pairs $\mathcal{P}_s$ and unseen pairs $\mathcal{P}_u$ i.e., $\mathcal{P}_s \cap \mathcal{P}_u = \emptyset$ and $\mathcal{P}_s \cup \mathcal{P}_u = \mathcal{P}$. Following previous works [36,44], we study this problem in the Generalized CZSL setting which has both seen $\mathcal{P}_s$ and unseen $\mathcal{P}_u$ pairs in the validation and test sets.

3.2. System Overview

Denote that the visual embeddings of attribute, object and pair are v_a, v_o and v_p, and text embeddings of those are t_a, t_o and t_p respectively.

The entire architecture is presented in Fig. 2 and is composed of two distinct parts separated by a compatible function. The left part corresponds to the text-based component, whereas the right part represents the visual component. In the text part, a single Multi Layer Perception (MLP) is employed to create pair embeddings t_p, and a Pronoun Memory Bank is utilized to produce the attribute embeddings t_a and object embeddings t_o. In the visual part, the visual component employs visual attribute and object encoders to generate attribute features and object features respectively. A feature fusion model is proposed to aggregate the attribute and object features into visual pair embeddings v_p. Moreover, to ensure alignment between the output of the Pronoun Memory Bank and the visual component, attention modules are employed to produce the visual attribute embeddings v_a and object embeddings v_o.

3.3. Pronoun Memory Bank (PMB)

In this section, we introduce the pronoun concept and the Pronoun Memory Bank to represent attribute and object embeddings.

Extending Pronoun to Adjectives. The usage of pronouns in natural language is a well-established linguistic phenomenon that allows speakers to refer to a previously mentioned nouns without emphasizing it [12, 40, 43]. For example, we would say `there's something running on the street` to emphasize the attribute `running` and pay little attention to what is running, and the `something` is the pronoun. In this task, the representation of attributes (adj.) is similar to the that of objects (noun). Thus, we propose extending this concept to include adjectives. We would say `there's an interesting dog` to emphasize the object `dog` and pay little attention to its attribute, and the `interesting` is the adjective version of pronoun. Note that we just use an example to show

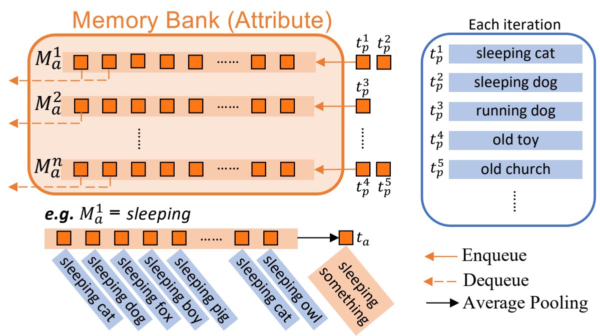


Figure 4. **Memory Bank Architecture.** We propose a FIFO memory bank to implicitly represent the concept of pronouns. All text embeddings are classified at each iteration according to attribute and object labels. Then they are enqueued in the memory bank, with the label of the attribute and object serving as the queue item, and dequeue the same number of the previous features.

the concept of pronoun in natural languages, so it doesn't mean that the embedding of the words `interesting` and `something` are used to represent pronoun.

Pronoun Memory Bank Architecture. We design a first-in-first-out Memory Bank to store the pair text embeddings. Then these embeddings could be used to represent the pronoun of attributes and objects as shown in Fig. 4. The memory bank in consideration has dimensions $n \times n_m \times d_f$, where n denotes the number of entities in the system, n_m represents the size of the memory queue, and d_f is the dimension of each feature vector stored in the queue. Here, n can either refer to the number of attributes ($n = n_{attr}$) or the number of objects ($n = n_{obj}$).

During training, for every composed pair of text embeddings t_p, which includes information about the j^{th} attribute and a random object, we update the j^{th} queue, denoted by $\mathbf{M}_a^j = [t_p^1, t_p^2, ..., t_p^{n_m}]$, to indicate the j^{th} attribute. We utilize a moving average of $\mathbf{M}_a^j$ to represent all available objects as a pronoun. Similar to the attribute memory bank, the object memory bank denoted by $\mathbf{M}_o$, is updated in the same way. The Pronoun Memory Bank is obtained by combining $\mathbf{M}_a$ and $\mathbf{M}_o$. The mechanism is shown in Figure 4. Thus the j^{th} attribute feature and the k^{th} object feature could be represented as:

$$t_a^j = \frac{1}{n_m}\sum_{i=1}^{n_m} \mathbf{M}_a^{j,i} \quad \text{and} \quad t_o^k = \frac{1}{n_m}\sum_{i=1}^{n_m} \mathbf{M}_o^{k,i} \qquad (1)$$

Attr-Obj Pronoun Representation. As shown in Fig. 2, from the text part, we utilize a pretrained text embed to extract the word embedding. Then an MLP is used to compose the attribute and the object embedding, and output the pair embedding t_p. We utilize Pronoun Memory Bank to store all pair embeddings and adopt the moving average method to output the final embeddings of attribute-pronoun t_a and pronoun-object t_o. Taking the attribute-pronoun as an example, obtaining a series of pair embeddings of `running dog`, `running cat`, `running human`, ..., we take the average of these embeddings to represent the attribute-pronoun `running something`. In this way, t_p is composed of an MLP, and t_a, t_o are obtained from the t_p through Pronoun Memory Bank. Thus the regression targets are regularized, leading to improved scalability and superior performance on the larger image encoding backbone.

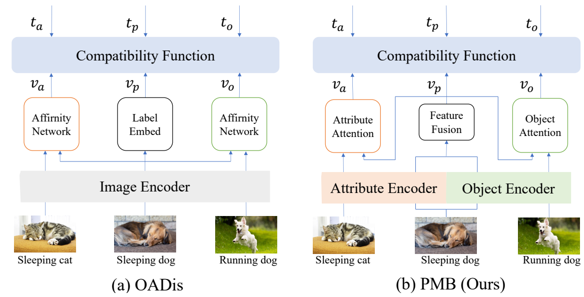


Figure 5. Comparison of OADis and our proposed method (PMB) from the image branch.

3.4. Visual Embedding Network Architecture

Integrating a Pronoun Memory Bank design into the OADis [39] framework is not feasible. Applying Pronoun Memory Bank leads to regression target embeddings corresponding to adjective-(pro)noun pairs. For example, Averaging the language embeddings like `running dog`, `running cat` and `running man` leads to a regression target corresponding to `running something`. This is quite different from OADis's attribute target, which corresponds to `running` (i.e. t_a in Fig. 3 (a)). This calls for a network architecture change to the image branch. Using OADis's design to extract attribute-only image feature (i.e. v_a in Fig. 5 (a)) is no longer a choice naturally compatible with the regression target `running something`. So, we modify to use two separate image encoders for attribute and object instead of the only image encoder of OADis. This modification can be used to conveniently generate image features corresponding to adjective-noun pairs.

Attribute and Object Encoder. We first use the second last layer before Pooling of a pretrained ResNet [8]. Attribute Encoder and Object Encoder share the same structure which is a 1×1 convolutional layer. Input three images I_p, I_a, and I_o that corresponding to the pair label (e.g. `sleeping dog`), attribute label (e.g. `sleeping cat`) and object label (e.g. `sleeping dog`). The attribute encoder generates f_a, $f_a^{'}$ from I_p and I_a and the object encoder generates f_o, $f_o^{'}$ from I_p and I_o respectively.

Feature Fusion Model. Bilinear models were first introduced by [42] to separate style and content, and [20] used Bilinear Pooling for image captioning. Inspired by [6, 35, 42], we propose to use Feature Fusion Model based

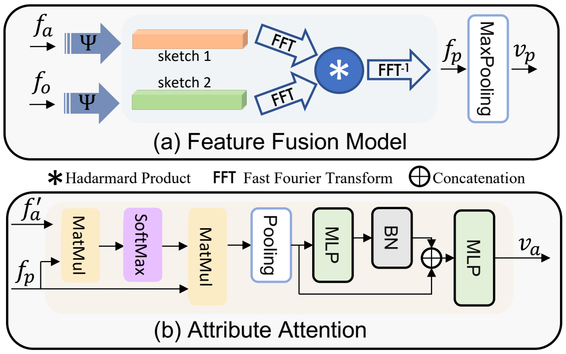


Figure 6. **Visual Output.** We adopt compact bilinear pooling method in the Feature Fusion Model. Ψ is the Count Sketch [4] function and FFT is fast Fourier transformation. In the image attention model, we only demonstrate the attribute attention structure because the object branch is symmetric to attribute branch and object attention has the same structure.

on Bilinear models to integrate attributes and objects of an image, as seen in Fig. 6 (a).

Feature Fusion Model first projects the attribute features f_a and the object features f_o to a lower dimensional space (using Count Sketch [4]) and then convolving both vectors by using element-wise product in Fast Fourier Transform (FFT) space. The Inverse Fast Fourier Transform (FFT^{-1}) outputs a composed feature map f_p.

$$f_p = \Phi(f_a, f_o) \tag{2}$$

where f_p, f_a and f_o are in the same shape of $n \times 7 \times 7$, n is the dimension of feature vector space. Φ is the function to compute the fusion results of two matrices.

$$\begin{aligned} \Phi =& \text{FFT}^{-1}(\text{FFT}(\phi_a) \otimes \text{FFT}(\phi_o)) \\ &\Psi(f, h, s) : f \rightarrow \phi \end{aligned} \tag{3}$$

where $\otimes$ is the hadarmard product [10]. Denoted that Ψ is the transform function to project f^a, f^o into a lower dimensional space by using Count Sketch [4] and thus we get ϕ^a and ϕ^o in the shape of n. Techonically, we initialize two vectors $h \in \{1, ..., n\}^n$ and $s \in \{-1, 1\}^n$, where h maps each index i in the input f to an index j in the output ϕ and s contains either -1 or 1 for each index.

$$\begin{aligned} \phi &= \{\phi(1), \phi(2), ...\phi(n)\} \\ \phi(i) &= \sum_{j}^{h(j)=i} s(j) f(j) \end{aligned} \tag{4}$$

Attribute and Object Attention. In computer vision, visual attention aims to focus on specific images or subregions [1, 16, 45]. And in compositional zero-shot learning tasks, image attributes and objects tend to attract different attention. For example, to distinguish `sleeping dog` and `running dog`, visual attention prefers to focus on the different motion states than only on the object feature. The attention module is used to extract similar features between two images, the attribute attention model structure is shown in Fig. 6 (b). As the architecture for visual attribute and object embedding output is symmetry, we could get the v_o by substituting f_a, f_a' for f_o, f_o'.

First, input two image features $f_a, f_a' \in \mathbb{R}^{n \times 49}$ (reshape $n \times 7 \times 7$ to $n \times 49$), and compute the feature relevance with a cosine distance.

$$R = \frac{f_1^T f_2}{\|f_1\| \|f_2\|} \tag{5}$$

Based on the relevance matrix $R \in \mathbb{R}^{49 \times 49}$, we apply softmax to normalize the feature map to the attention score. The similarity between two images could be represented as:

$$s(\lambda, R) = \sum_{i=1}^{d} \frac{e^{\lambda r_{ij}}}{\sum_{j=1}^{d} e^{\lambda r_{ij}}} \tag{6}$$

where $d = 49$ is the dimension of the space of Relevance matrix, r_{ij} is a element in R with location of the i^{th} row and the j^{th} column, λ is the inverse temperature parameter of the softmax function.

The output similarity contains rich covariance information. Our target of this module is to output a vector containing information on the similarity. For example, assume the input f_a, f_a' is `running cat` and `running dog`, the output should be a vector with the same dimension as the inputs, and that captures the semantic meaning of `running something`. Taking the attribute attention as an example, the attention on output would be

$$\text{Attn}_a = f_a \cdot s(\lambda, R(f_a, f_a')) \tag{7}$$

where $(\cdot)$ is the matrix product, and we use MaxPooling and MLPs to final output the visual attribute embedding v_a.

3.5. Compatibility Function

Compatibility Score. Following previous work [30,34,39], we use cosine similarity to measure the final prediction for each pair, and use cross entropy to calculate the final compatibility score. For visual embeddings like $v \in \mathbb{R}^n$ and text embeddings like $T \in \mathbb{R}^{m \times n}$, where m is the total number of text embeddings, we calculate the cosine similarity of the prediction embeddings and all text pair embeddings.

$$C(v, T) = cos(v, T) = \frac{vT^T}{\|v\| \|T\|} \tag{8}$$

Getting the final prediction matrix, the training process is supervised by a cross-entropy loss.

$$\mathcal{L}(v, T) = \frac{e^{C(v,T)}}{\sum_{p \in \mathcal{T}} e^{C(v,p)}} \tag{9}$$

Training Loss. All text embeddings $t_p \in \mathbb{R}^n$ compose of the total text embeddings $T \in \mathbb{R}^{m \times n}$. The main loss $\mathcal{L}_p$ is computed by T and visual embeddings v_p through the compatibility function.

$$\mathcal{L}_p = \mathcal{L}(v_p, T) \tag{10}$$

Embeddings with the same attribute or the same object are regularized and optimized by visual attribute-pronoun features v_a and attribute memory bank $\mathbf{M}_a$, visual pronoun-object features v_o and object memory bank $\mathbf{M}_o$. The loss functions are represented as:

$$\begin{aligned} \mathcal{L}_a &= \mathcal{L}(v_a, \mathbf{M}_a) \\ \mathcal{L}_o &= \mathcal{L}(v_o, \mathbf{M}_o) \end{aligned} \tag{11}$$

The total loss is the weighted sum of the above, where λ_1, λ_2 are hyperparameters. As the architecture for attribute and object embedding output is symmetry, we set $\lambda_1 = \lambda_2 = \lambda$, and λ is 0.25 in this paper. Additionally, we run ablations on λ in the supplementary.

$$\mathcal{L}_{total} = \mathcal{L}_p + \lambda_1 \mathcal{L}_a + \lambda_2 \mathcal{L}_o \tag{12}$$

Inference. In the validation or test process, we derive a prediction by searching the pair label that yields the highest cosine similarity, Given an image, using and attribute encoder and object encoder to generate f_a and f_o, we could get the visual pair embedding v_p through the feature fusion model, the result is shown by $\text{Pred}(v_p)$.

$$\text{Pred}(v_p) = \arg\max_{p \in \mathcal{P}} C(v_p, T_+) \tag{13}$$

$T_+ \in \mathbb{R}^{m_+ \times n}$ contains all seen and unseen text pair labels, where $m_+ > m$. It is worth mentioning that our model works in the generalized Compositional Zero-Shot Learning setting, all reachable classes of seen and unseen are predicted.

4. Experiment

4.1. Datasets and Metrics

Datasets. Our experiments are conducted on three datasets: MIT-states [11], UT-Zappos [46] and VAW-CZSL [39]. MIT-states [11] contains 63440 images covering 115 attributes and 245 objects. Each image is attached to an attribute-object pair label and there are 1262 classes of pairs in total. We use 1262 pairs/30338 images for training and 800 pairs/12995 images for testing. UT-Zappos [46] contains 12 types of attributes and 16 types of objects. We use 83 pairs/22998 images as the train set and 36 pairs/2914 images as the test set. VAW-CZSL [39] is a dataset with a much larger output space of 440 attributes and 541 objects. We use 11175 pairs/72203 images for training and 4019 pairs/10856 images for testing.

Metrics. We adopt the evaluation protocol [36] and report the Area Under the Curve (AUC) (in %) between the accuracy on seen and unseen compositions with different bias terms, which are positively relevant to the performance of unseen pairs and negative relevant to that of seen pairs. In addition, the best harmonic mean is reported when the bias is balanced. Furthermore, we also present the accuracy of attributes and objects to show the improvement through the regularization of attribute regression and object regression.

Training Details. Our image features are extracted from the ResNet101 [8] pre-trained on ImageNet [38]. We use pretrained text embedding GloVe [33] to process the words to vectors. The embed dimension for MIT-States and VAW-CZSL datasets is 300, and for UTZappos dataset is 100. The text encoder is an MLP of one hidden layer and the feature shapes of the input are 600 and that of the output is the embed dimension. We use Adam Optimizer [13] with an initial learning rate of $3e^{-4}$ and the decay factor of 0.1. We train our model on NVIDIA 3090 GPUs.

4.2. Quantitative Results

We evaluate our PMB method on public benchmarks MIT-states [11], UT-Zappos [46] and VAW-CZSL [39]. Due to the utilization of Pronoun Memory Bank, the image encodings of out model networks are optimized towards more consistent targets. As evidenced by our experiments, switching the image backbone from ResNet18 to ResNet101 hardly improves or hurts the performance of compositional understanding under the OADis [39] framework. Our model has better consistent performance and achieves state-of-the-art result on all datasets.

MIT-States. Our PMB method shows its robustness against considerable noise in the MIT-states dataset. It achieves a test AUC of 7.3% and a validation AUC of 8.8%, which is a significant improvement from the previous state-of-the-art OADis [39] of 5.9% and 7.6% AUC on test and validation set respectively as seen in Table 1. It is worth mentioning that we have better scalability by using backbone ResNet101, but if we replace the backbone of OADis, the evaluation metrics will barely improve. Overall, our model outperforms other models on all metrics. Besides, our model could see a slight improvement using ResNet18.

UT-Zappos. We show our results of AUC of 31.7% on the test set and 40.7% on the validation set, which overtakes all other models on all metrics. Besides, due to better consistancy of the regression targets, we could also get better performance using ResNet18. However, it is hard to make much improvement since not all labels (7/36 attribute labels) have appeared in the train set, so training, validation, and test are not always highly relevant. We need to pay attention to the over-fitting problem and balance the performance on the validation set and test set.

Table 1. We show results on MIT-states [11] and UT-Zappos [45]. Following [36], we use AUC in % between seen and unseen compositions with different bias terms, along with Val, Test, attribute and object accuracy. HM is Harmonic Mean.

	MIT-States							UT-Zappos						
Model	Val@1	Test@1	HM	Seen	Unseen	Attribute	Object	Val@1	Test@1	HM	Seen	Unseen	Attribute	Object
AttrOpr [27]	2.5	2.0	10.7	16.6	18.4	22.9	24.7	29.9	22.8	38.1	55.5	54.4	38.6	70.0
LabelEmbed+ [27]	3.5	2.3	11.5	16.2	21.2	25.6	27.5	35.5	22.6	37.7	53.3	58.6	40.9	69.1
TMN [36]	3.3	2.6	11.8	22.7	17.1	21.3	24.2	35.9	28.4	44.0	58.2	58.0	40.8	68.4
CompCos [24]	6.9	4.8	16.9	26.9	24.5	28.3	31.9	40.8	26.9	41.1	57.7	62.8	43.3	73.0
Symnet [18]	4.5	3.4	13.8	24.8	20.0	26.1	25.7	27.4	27.7	42.5	56.7	61.6	44.0	70.6
GraphEmb [26]	7.2	5.3	18.1	28.9	25.0	27.2	32.5	33.9	24.7	38.9	58.8	61.0	44.0	72.6
OADis [39]+ResNet18	7.6	5.9	18.9	31.1	25.6	28.4	33.2	40.8	30.0	44.4	59.5	65.5	46.5	75.5
OADis [39]+ResNet101	7.4	5.6	18.4	29.8	27.5	30.8	35.4	40.0	30.1	45.3	59.3	64.6	46.6	75.3
PMB+ResNet18	7.5	5.9	19.4	31.6	25.2	28.0	33.3	39.7	31.0	45.8	60.4	65.4	46.7	75.0
PMB+ResNet101	**8.8**	**7.3**	**20.9**	**35.0**	**28.8**	**31.3**	**37.2**	**40.7**	**31.7**	**45.9**	**60.8**	**65.0**	**46.3**	**73.7**

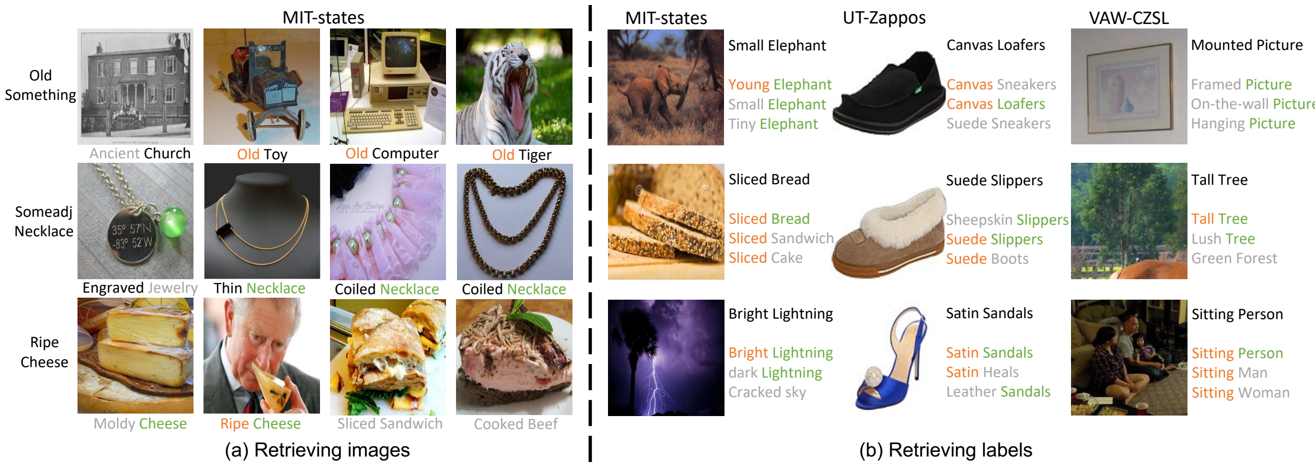


Figure 7. **Qualitative Results:** Left(a): We show good predictions for retrieving images from given labels. The first row focuses on the attribute `old`, the second row focuses on the object `necklace` and the third row is `ripe cheese`. Right(b): The top-3 predictions of our model for some examples from three datasets are shown and most of the predictions make sense. The words in black are ground truth, colored ones are good predictions and the grey ones are wrong to ground truth.

Table 2. We show results on VAW-CZSL [39]

	VAW-CZSL						
Model	Val@3	Test@3	HM	Seen	Unseen	Attribute	Object
AttrOpr [27]	1.4	1.4	9.1	16.4	11.7	13.7	34.9
LabelEmbed+ [27]	1.5	1.6	9.8	16.2	13.2	13.4	35.1
TMN [36]	2.2	2.3	11.9	19.9	15.4	15.9	38.3
Symnet [18]	2.3	2.3	12.2	19.1	15.8	18.6	40.9
CompCos [24]	3.1	3.2	14.2	23.9	18.0	16.9	41.9
GraphEmb [26]	2.7	2.9	13.0	23.4	16.8	16.9	40.8
OADis [39]	3.5	3.6	15.2	24.9	18.7	17.5	43.3
OADis+ResNet101 [39]	3.4	3.7	15.3	25.0	19.8	18.2	44.5
PMB+ResNet18	3.6	3.6	15.3	25.8	18.6	18.4	43.7
PMB+ResNet101	**4.4**	**4.5**	**17.2**	**26.0**	**22.3**	**20.4**	**47.1**

Figure 8. Prediction on hard-to-identify items: Labels on top is the ground truth, the yellow ones are the prediction of OADis and the blue ones are the prediction of our model, bold font style indicates the correct prediction.

VAW-CZSL. VAW-CZSL dataset poses a significant challenge due to its large number of samples and labels. As a result, the use of top-1 AUC as the sole evaluation metric for this task may be too strict. In this task, we use top-3 AUC to evaluate the VAW-CZSL dataset and demonstrate a significant improvement over previous state-of-the-art results. Specifically, our approach achieves an AUC of 4.4% and 4.5% for the validation and test sets, respectively, compared to the previous best results of 3.4% and 3.7%. Additionally, we observe improvements in attribute and object accuracy, with increases from 17.5% to 20.4% for attributes and from 43.3% to 47.1% for objects.

4.3. Qualitative Results

To perform a qualitative evaluation of our proposed PMB model, we present image retrieval results for a particular attribute and object, as depicted in Fig. 7 (a). Additionally, we showcase label retrieval results in Fig. 7 (b). Our findings in Fig. 8 demonstrate that our model exhibits enhanced capabilities for identifying challenging items.

Image retrieval. The examples are predicted to the labels shown below each image as shown in Fig. 7 (a), while the ground truth is marked on the left. The label `old something` is assigned to different objects with the attribute `old`, indicating that the model captures subtle variations in the meaning of attributes. The correct answers demonstrate that despite differences in appearance, shape, and color, the model correctly predicted the object as a `necklace`. However, the incorrect answer `jewelry` suggests that semantic similarity between object labels can confuse the model. In the third row, although only one image was predicted correctly, the other predicted labels were still semantically meaningful. These examples could illustrate that our PMB model works efficiently and robustly.

Label retrieval. In Fig. 7 (b), we present the results of retrieving the top-3 predicted labels corresponding to a set of given images. The labels are grouped into different categories based on their attributes such as color, shape, size, illuminance, and objects such as scenes, people, and animals. The predicted labels are reasonable and make sense for most of the images. However, one failed example is the `mounted picture` from VAW-CZSL, where the predicted labels are `framed picture`, which share a similar meaning with the ground truth. Additionally, the predicted labels `on-the-wall picture` and `hanging picture` are also semantically related to the input image. Thus, while there was an error in predicting the exact label, the predicted labels are still meaningful.

Prediction on challenging items. In this section, we present a set of images that pose a significant challenge for human observers to identify accurately, as illustrated in Fig. 8. However, our model was able to accurately predict the correct labels for these images, whereas the OADis [39] model struggled to perform well. This finding highlights our model's superior capability and robustness when dealing with difficult-to-identify items.

Discussion. Thanks to the PMB design, our regression target is averaged over many different `old` objects during training and thus can well grasp the concept of `old`. During test time, `old` objects, buildings and animals are all recalled. As seen in figure 7, given an example of attribute `old`, it is obvious that an `old building` is featured of its poor color and uneven edges, and an `old computer` is yellow rather than discoloration. These two features share the same attribute `old` but would show a different appearance. Different from items, finding age information on animals e.g. `old tiger` is far more complicated than we could barely distinguish by our eyes. Learning with these averaged objects makes it possible to gain more semantic information of the attribute combined with different objects. Objects recognition is improved in the same way.

Table 3. Results with different feature fusion methods.

Fusion Methods	MIT-States		UT-Zappos		VAW-CZSL	
	Test@1	HM	Test@1	HM	Test@3	HM
Element-wise Sum+FC	7.1	20.1	30.9	44.8	4.0	16.7
Element-wise Product+FC	7.0	20.0	30.4	45.0	4.0	17.2
Concatenation+FC	7.0	20.4	30.1	43.2	4.1	16.5
Bilinear Pooling	**7.3**	**20.9**	**31.7**	**45.9**	**4.5**	**17.2**

Table 4. Results on different sizes of Memory Bank

Size	MIT-States		UT-Zappos		VAW-CZSL	
	Test@1	HM	Test@1	HM	Test@3	HM
10	7.0	21.0	29.0	44.0	4.4	16.8
1024	**7.3**	**20.9**	**31.7**	**45.9**	**4.5**	**17.2**
2048	6.9	20.8	30.2	43.7	4.3	13.7
Momentum [7]	7.1	21.2	30.9	44.5	4.4	16.5

4.4. Ablations

In this section, we ablate our PMB model with respect to different feature fusion methods and the size of the Pronoun Memory Bank.

Feature Fusion Methods. We compare the performance of non-bilinear and bilinear pooling methods in Tab. 3. For the main feature fusion model after Attribute and Object Encoders, we compare our Feature Fusion Model (bilinear pooling) with element-wise sum, element-wise product and concatenation with a fully connected layer.

Memory Bank Size. Tab. 4 presents the results of varying the size of the pronoun memory bank. Using an excessively large memory bank can lead to a decline in performance due to the inclusion of outdated information and increased storage requirements. Conversely, a memory bank that is too small is insufficient to represent the concept of pronouns adequately. Therefore, an appropriate range of memory bank sizes falls between 10 and 1024. Besides, we conducted experiments based on MoCo [7]. However, the results indicate that MoCo did not lead to any improvement in this task.

5. Conclusion

In this work, we propose a new framework for compositional zero-shot learning. We regularize the output of the text encoder as attribute-object pair embeddings, and use Pronoun Memory Bank to generate attribute and object embeddings by introducing pronoun concepts. The Pronoun Memory Bank makes the image encoders learn more consistent regression targets. Thus, our proposed method has good performance and better scalability on the larger image encoding backbone. Our experimental results demonstrate that the PMB framework achieves state-of-the-art performance on all three datasets.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 5

[2] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5327–5336, 2016. 2

[3] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European conference on computer vision*, pages 52–68. Springer, 2016. 2

[4] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002. 5

[5] Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277, 2014. 2

[6] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016. 4

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 8

[8] 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. 4, 6

[9] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2016. 2

[10] Roger A Horn. The hadamard product. In *Proc. Symp. Appl. Math*, volume 40, pages 87–169, 1990. 5

[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. 6, 7

[12] Hossein Karimi and Fernanda Ferreira. Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly journal of experimental psychology*, 69(5):1013–1040, 2016. 3

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, 115(2):185–210, 2015. 2

[15] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on graphics (TOG)*, 33(4):1–11, 2014. 2

[16] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 5

[17] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE international conference on computer vision*, pages 4247–4255, 2015. 2

[18] 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. 7

[19] Zhizhong Li and Derek Hoiem. Improving confidence estimates for unfamiliar examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2686–2695, 2020. 1

[20] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015. 4

[21] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. *Advances in neural information processing systems*, 31, 2018. 2

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2

[23] 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. 2

[24] 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*, 2022. 2, 7

[25] 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. 2, 3

[26] 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. 2, 7

[27] 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. 2, 7

[28] 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*, pages 8811–8818, 2019. 2

[29] Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, 2011. 2

[30] Seyoung Park, Bruce Xiaohan Nie, and Song-Chun Zhu. Attribute and-or grammar for joint parsing of human pose, parts and attributes. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1555–1569, 2017. 2, 5

[31] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012. 2

[32] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision*, pages 85–100. Springer, 2016. 2

[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 6

[34] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028, 2021. 2, 5

[35] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247, 2013. 4

[36] 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. 2, 3, 6, 7

[37] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5899–5908, 2018. 2

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1, 6

[39] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667, 2022. 1, 2, 4, 5, 6, 7, 8

[40] Robert Stalnaker. On the representation of context. In *Semantics and Linguistic Theory*, volume 6, pages 279–294, 1996. 3

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1

[42] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000. 4

[43] Heike Wiese and Horst J Simon. Grammatical properties of pronouns and their representation. *Pronouns–grammar and representation*, pages 1–21, 2002. 3

[44] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 3

[45] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 5, 7

[46] 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. 6

[47] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 33–42, 2013. 2

[48] Shuai Zheng, Ming-Ming Cheng, Jonathan Warrell, Paul Sturgess, Vibhav Vineet, Carsten Rother, and Philip HS Torr. Dense semantic image segmentation with objects and attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3214–3221, 2014. 2