Learning to Predict Visual Attributes in the Wild

Khoi Pham[1][*]   Kushal Kafle[2]   Zhe Lin[2]   Zhihong Ding[2]   Scott Cohen[2]

Quan Tran[2]   Abhinav Shrivastava[1]

[1]University of Maryland, College Park   [2]Adobe Research

[1]{khoi,abhinav}@cs.umd.edu   [2]{kkafle, zlin, zhding, scohen, qtran}@adobe.com

Abstract

Visual attributes constitute a large portion of information contained in a scene. Objects can be described using a wide variety of attributes which portray their visual appearance (color, texture), geometry (shape, size, posture), and other intrinsic properties (state, action). Existing work is mostly limited to study of attribute prediction in specific domains. In this paper, we introduce a large-scale in-the-wild visual attribute prediction dataset consisting of over 927K attribute annotations for over 260K object instances. Formally, object attribute prediction is a multi-label classification problem where all attributes that apply to an object must be predicted. Our dataset poses significant challenges to existing methods due to large number of attributes, label sparsity, data imbalance, and object occlusion. To this end, we propose several techniques that systematically tackle these challenges, including a base model that utilizes both low- and high-level CNN features with multi-hop attention, reweighting and resampling techniques, a novel negative label expansion scheme, and a novel supervised attribute-aware contrastive learning algorithm. Using these techniques, we achieve near 3.7 mAP and 5.7 overall F1 points improvement over the current state of the art. Further details about the VAW dataset can be found at https://vawdataset.com/

1. Introduction

Learning to predict visual attributes of objects is one of the most important problems in computer vision and image understanding. Grounding objects with their correct visual attribute plays a central role in a variety of computer vision tasks, such as image retrieval and search [32], tagging, referring expression [35], visual question answering (VQA) [4, 33], and image captioning [7].

While several existing works address attribute prediction, they are limited in many ways. Objects in a visual
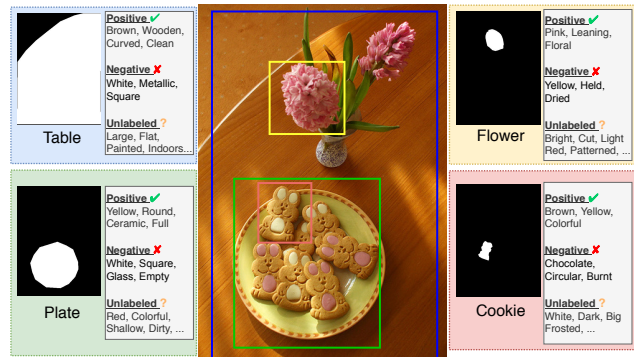

Figure 1. **Example annotations in our dataset.** Each possible attribute-object category pair is annotated with at least 50 examples consisting of explicit positive and negative labels.

scene can be described using a vast number of attributes, many of which can exist independently of each other. Due to the variety of possible object and attribute combinations, it is a daunting task to curate a large-scale visual attribute prediction dataset. Existing works have largely ignored large-scale visual attribute prediction in-the-wild and have instead focused only on domain-specific attributes [19, 43], datasets consisting of very small number of attribute-object pairs [51], or are rife with label noise, ambiguity and label sparsity [38]. Similarly, while attributes can form an important part of related tasks such as VQA, captioning, referring expression, these works do not address the unique challenges of attribute prediction. Existing work also fails to address the issue of partial labels, where only a small subset of all possible attributes are annotated. Partial labels and the lack of explicit negative labels make it challenging to train or evaluate models for large-scale attribute prediction. To address these problems, we propose a new large-scale visual attribute prediction dataset for images in the wild that includes both positive and negative annotations.

Our dataset, called visual attributes in the wild (VAW), consists of over 927K explicitly labeled positive and negative attribute annotations applied to over 260K object instances (with 620 unique attributes and 2,260 unique ob-

ject phrases). Due to the number of combinations possible, it is prohibitively expensive to collect exhaustive attribute annotations for each instance. However, we ensure that every attribute-object phrase pair in the dataset has a *minimum* of 50 positive and negative annotations. With density of 3.56 annotations per instance, our dataset is 4.9 times denser compared to Visual Genome while also providing negative labels. Additionally, annotations in VAW are visually-grounded with segmentation masks available for 92% of the instances. Formally, our VAW dataset proposes attribute prediction as a long-tailed, partially-labeled, multi-label classification problem.

We explore various state-of-the-art methods in attribute prediction and multi-label learning and show that the VAW dataset poses significant challenges to existing work. To this end, we first propose a strong baseline model that considers both low- and high-level features to address the heterogeneity in features required for different classes of attributes (*e.g.*, *color* vs. *action*), and is modeled with multi-hop attention and an ability to localize the region of the object of interest by using partially available segmentation masks.

We also propose a series of techniques that are uniquely suited for our problem. Firstly, we explore existing works that address label imbalance between positive and negative labels. Next, we describe a simple yet powerful scheme that exploits linguistic knowledge to expand the number of negative labels 15-fold. Finally, we propose a supervised contrastive learning approach that allows our model to learn more attribute discriminative features. Through extensive ablations, we show that most of our proposed techniques are model-agnostic, producing improvements not only on our baseline but also other methods. Our final model is called Supervised Contrastive learning with Negative-label Expansion (SCoNE), which surpasses state-of-the-art models by 3.5 mAP and 5.7 overall F1 points.

Our paper makes the following major contributions:

- We create a new large-scale dataset for visual attribute prediction in the wild (VAW) that addresses many shortcomings in existing literature and demonstrate that VAW poses considerable difficulty to existing algorithms.
- We design a strong baseline model for attribute prediction using existing visual attention technique. We further extend this baseline to our novel attribute learning paradigm called Supervised Contrastive learning with Negative-label Expansion (SCoNE) that considerably advances the state of the art.
- Through extensive experimentation, we show the efficacy of both our proposed model and our proposed techniques.

2. Related Work

Some of the earliest work related to attribute learning stem from a desire to learn to describe objects rather than predicting their identities [17, 16, 6, 18]. Since then, extensive work has sought to explore several aspects of object attributes, including attribute-based zero-shot object classification [41, 30, 2], relative attribute comparison [50, 58, 10], and image search [57, 39]. While research in compositional zero shot learning [66, 53, 49, 29] also tackle object attributes, they target transformation of 'states' of objects, treat each instance as having only *one* state, and focus on predicting unseen compositions rather than the prediction of complete set of attributes for each object instance. Several works have focused on attribute learning in specific domains such as animals, scenes, clothing, pedestrian, human facial and emotion attributes [72, 81, 19, 45, 43, 37]. In contrast, we seek to explore attribute prediction for *unconstrained* set of objects.

Only a limited number of work have sought to explore *general* attribute prediction. COCO Attributes [51] is an attempt to develop in-the-wild attribute prediction dataset, however, it is very limited in scope, covering only 29 object categories. Similarly, a portion of the the Visual Genome (VG) [38] dataset consists of attribute annotations. However, attributes in VG are not a central focus of the work and therefore they are very sparsely labeled, noisy, and lack negative labels, making it unsuitable to be used as a standalone attribute prediction benchmark. Despite this, attribute annotations from VG are used to train attribute-aware object detectors for downstream vision-language tasks [3, 31, 78]. By introducing the VAW dataset, the research community can use its dense attribute annotations in conjunction with VG and our attribute learning techniques to train better attribute prediction models. Several recent works have also sought to take advantage of massive amount of data in VG to curate datasets for specific challenges [27, 70]. In a similar vein, we also start by leveraging existing sources of clean annotations to develop our VAW dataset.

VAW can be cast as a multi-label classification problem which has been extensively studied in the research community [15, 13, 44, 54]. Multi-label learning involving missing labels poses a greater challenge, but is also extensively studied [28, 40, 69, 74]. In many cases, missing labels are assumed to be *negative* examples [61, 8, 60, 76] which is unsuitable for attribute prediction, since most of the attributes are not mutually exclusive. Some others attempt to predict missing labels by training expert models [56], which is also infeasible for a large-scale problem like ours.

Data imbalance naturally arises in datasets with large set of labels. As expected, label imbalance exists in our VAW dataset, therefore techniques designed to learn from imbalanced data are also related to our explorations. These works can be divided into two main approaches: cost-sensitive learning [26, 67, 14] and resampling [55, 22, 9, 20, 47]. We utilize both of these techniques in our final model.

Attention is a highly effective technique in image clas-

sification, captioning, VQA, and domain-specific attribute prediction [68, 73, 46, 28, 62, 79, 54]. In our VAW dataset, most of the objects are annotated with their segmentation mask, which allows us to guide the attention map to ignore irrelevant image regions. We also use additional attention maps to allow our model to properly explore the surrounding context of the object.

Contrastive learning has recently gained a lot of traction as an effective self-supervised learning technique [71, 5, 23, 11]. While originally intended to be used in self-supervised setting, recent works have expanded contrastive learning to be used in supervised setting [36]. Motivated by these works, we propose an extension of supervised contrastive loss to allow it to work in a multi-label setting required for VAW. To the best of our knowledge, ours is the first attempt to apply contrastive loss for multi-label learning.

3. Visual Attributes in the Wild (VAW) Dataset

In this section, we describe how we collect attribute annotations and present statistics of the final VAW dataset. In general, we aim to overcome the limitations of VG on the attribute prediction task which includes noisy labels, label sparsity, and lack of negative labels to create a dataset applicable for training and testing attribute classification models.

3.1. Data collection

VAW is created based on the VGPhraseCut and GQA datasets, both of which leverage and refine annotations from Visual Genome. VGPhraseCut is a referring expression dataset that provides high-quality attribute labels and per-instance segmentation mask, while GQA is a VQA dataset that presents cleaner scene graph annotations.

Step 1: Extraction from VGPhraseCut and GQA

Our goal is to build a dataset that allows us to predict the maximal number of attributes *commonly* used to describe objects in the wild. From VGPhraseCut, we select attributes that appear within more than 15 referring phrases. After manually cleaning ambiguous/hard to recognize attributes, we obtain a set of 620 unique attributes which are used throughout the rest of the process. Next, we extract more instances from GQA that are labeled with these attributes. We further take advantage of the referring expressions from VGPhraseCut to collect a reliable negative label set: given an image, for instances that are not selected by an attribute referring phrase, we assign that attribute as a negative label for the instance. This step allows us to collect 220,049 positive and 21,799 negative labels.

Step 2: Expand attribute-object coverage

In this step, we seek to collect additional annotations for every *feasible* attribute-object pair that may be lacking annotations. We define feasible pair as those with at least 1 positive example in our dataset. We ensure that every feasible pair has at least 50 (positive or negative) annotations. To keep the annotation cost in check, we do not annotate pairs that already have 50 or more annotations. This expansion enriches our dataset with more positives and negatives for every attribute across different objects, allowing for better training and evaluation of classification models. This step adds 156,690 positive and 455,151 negative annotations.

Step 3: Expand long-tailed attribute set

In this step, we aim to collect additional annotations for the long-tailed attributes. Long-tailed attributes are associated with very few object categories, which is either due to the attribute not being frequently used by humans or the attribute is only applied to a small set of objects. Hence, given a long-tailed attribute and a known object that it applies to, we first expand its set of *possibly applied objects* by using the WordNet ontology. For example, while *playing* may only be applied to *child* in the training set, it *could* also be applicable to other closely related object categories like *man, woman, person*. After we find candidate object categories for a given long-tail attribute, we ask humans to annotate randomly sampled images from these candidates with either positive or negative label for the given attribute. This step adds additionally 16,239 positive and 57,751 negative annotations pertaining to all long-tailed attributes.

3.2. Statistics

Our final dataset consists of 620 attributes describing 260,895 instances from 72,274 images. Our attribute set is diverse across different categories, including *color, material, shape, size, texture, action*. On the annotated instances, our dataset consists of 392,978 positive and 534,701 negative attribute labels. The instances from VGPhraseCut (occupy 92% in the dataset) are provided with segmentation masks which can be useful in attribute prediction. We split the dataset into 216,790 instances (58,565 images) for training, 12,286 instances (3,317 images) for validation, and 31,819 instances (10,392 images) for testing. We split the dataset such that the test set has higher annotation density per object, which allows for more thorough testing. In particular, our test set has an average of 7.03 annotations per instance compared to 3.02 in the training set.

In Table 1, we compare the statistics of the VAW dataset with other in-the-wild and domain-specific visual attribute datasets. Compared to existing work, VAW fills an important gap in the literature by providing a domain-agnostic, in-the-wild visual attribute prediction dataset with denser annotations, explicit negative labels, segmentation masks, and large number of attribute and object categories.

4. Methodology

In this section, we will describe components of our strong baseline model along with the Supervised Contrastive learning with Negative-label Expansion (SCoNE)

Dataset	VAW	Visual Genome [38]	COCO Attributes [51]	EMOTIC [37]	WIDER [43]	iMaterialist [19]
# attributes	620	68,111	196	26	14	228
# instances	260,895	3,843,636	180,000	23,788	57,524	1,012,947
# object categories	2,260	33,877	29	1 (person)*	1 (person*)	1 (clothes)*
# attribute anno. per instance	3.56	0.73	≥ 20	26	14	16.17
Negative Labels	Yes	No	Yes	Yes	Yes	Yes
Segmentation masks	Yes	No	No	No	No	Yes
Domain	In-the-wild	In-the-wild	In-the-wild	Emotions	Pedestrian	Fashion

Table 1. **Statistics of VAW with other attribute in-the-wild and domain-specific datasets.** *person (resp. *clothes) category may represent multiple categories including {boy, girl, man, woman, *etc.* } (resp. {shirt, pants, top, *etc.* }). While Visual Genome is the largest among these in terms of number of attribute annotations, it is sparsely labeled. Other datasets are either fully annotated for domain-specific attributes or more densely labeled but covering few object categories.

algorithm that helps our model to learn more attribute discriminative features. A depiction of our strong baseline model is shown in Figure 2.

4.1. Preliminaries

Problem formulation: Let $\mathcal{D} = \{I_i, g_i, o_i; Y_i\}_{i=1}^N$ be a dataset of N training samples, where I_i is an object instance image (cropped using its bounding box), g_i is its segmentation mask, o_i is the category phrase of the object for which we want to predict attributes, and $Y_i = [y_{i,1}, ..., y_{i,C}]$ is its C-class label vector with $y_c \in \{1, 0, -1\}$ denoting whether attribute c is positive, negative, or missing respectively. Our goal is to train a multi-label classifier that, given an input image and the object name, can output the confidence score for all C attribute labels.

Image feature representation: Given an image I of an object o, let $f_{img}(I) \in \mathbb{R}^{H \times W \times D}$ be the D-dimensional image feature map with spatial size $H \times W$ extracted using any CNN backbone architecture. In our model, we use the output of the penultimate layer of ResNet-50 [24].

4.2. Image-object feature composition

Prior models for attribute prediction mostly tackle domain-specific settings or a limited number of object categories [51, 54, 56]. Hence, these works are able to employ object-agnostic attribute classification. However, because our VAW dataset contains attribute annotations across a diverse set of object categories, incorporating object embedding as input can help the model learn to avoid infeasible attribute-object combinations (*e.g. parked dog*).

There are multiple ways to compose the image feature map with the object embedding [52, 65, 3]. Here, we opt for a simple object-conditioned gating mechanism, which we find to be consistently better than concatenation used in [3, 31]. Let $\phi_o \in \mathbb{R}^d$ be the object embedding vector, $f_{comp}(f_{img}(I), \phi_o) \in \mathbb{R}^D$ be the composition module that takes in the image feature map and object embedding. We implement f_{comp} with a gating mechanism as follows:

$$f_{comp}(f_{img}(I), \phi_o) = f_{img}(I) \odot f_{gate}(\phi_o), \quad (1)$$

$$f_{gate}(\phi_o) = \sigma(W_{g2} \cdot \text{ReLU}(W_{g1}\phi_o + b_{g1}) + b_{g2}), \quad (2)$$

$\odot$ is the channel-wise product, $\sigma(\cdot)$ is the sigmoid function, $f_{gate}(\phi_o) \in \mathbb{R}^D$ is broadcasted to match the feature map spatial dimension and is a 2-layer MLP. Intuitively, f_{gate} acts as a filter that only selects attribute features relevant to the object of interest and suppresses incompatible attribute-object pairs.

4.3. Object localization and multi-attention module

Relevant object localizer. An object bounding box can contain both the relevant object and other objects or background. Hence, it is desirable to learn a smarter feature aggregation that can suppress all irrelevant image regions. We propose to leverage the availability of the object segmentation mask in the VAW dataset to achieve this.

Let $X \in \mathbb{R}^{H \times W \times D}$ be the image-object composed feature map, the relevant object region G is localized using a 2-stacked convolutional layers f_{rel} with kernel size 1, followed by spatial softmax:

$$g = f_{\text{rel}}(X), \ \ g \in \mathbb{R}^{H \times W}, \quad (3)$$

$$G_{h,w} = \frac{\exp(g_{h,w})}{\sum_{h,w} \exp(g_{h,w})}, \ \ G \in \mathbb{R}^{H \times W}. \quad (4)$$

We can then pool the image feature vector as

$$Z_{\text{rel}} = \sum_{h,w} G_{h,w} X_{h,w}. \quad (5)$$

G is learned with direct supervision from the object mask whenever available with the following loss:

$$\mathcal{L}_{\text{rel}} = \sum_{h,w}(G_{h,w} \times (1 - M_{h,w})) - \lambda_{\text{rel}}(G_{h,w} \times M_{h,w}), \quad (6)$$

where M is the ground truth object binary mask. Rather than requiring G to exactly match the object mask, we find it is better to penalize the network whenever its prediction falls outside of the mask. This frees the network to learn heterogeneous attention within the object region if necessary (*e.g.*, *black mirror* refers to its frame being *black* rather
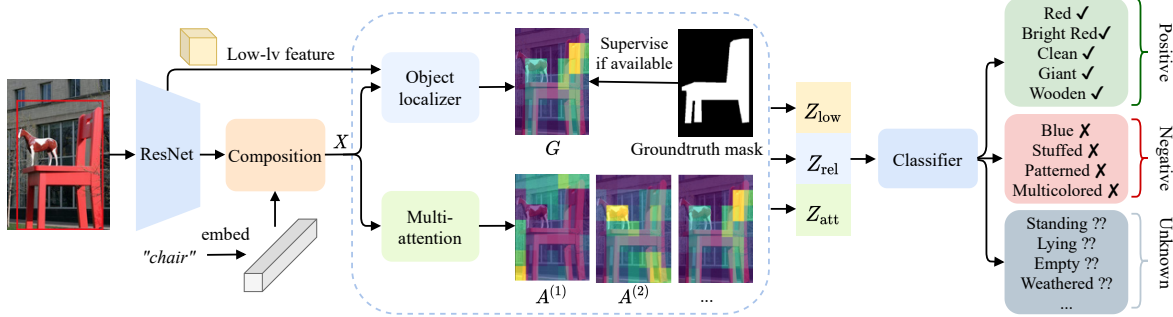
Figure 2. **Strong baseline attribute prediction model.** Feature map extracted from the input image is modulated with the object embedding which allows the model to learn useful attribute-object relationships (*e.g. ball* is *round*) and also to suppress infeasible attribute-object pairs (*e.g. talking table*). The image-object combined feature map X is used to infer the object region G and multiple attention maps $\{A^{(m)}\}$ which are subsequently used to aggregate features for classification. Here, Z_{low} and Z_{rel} respectively denotes low-level and image-object features aggregated inside the estimated object region, Z_{att} corresponds to image-object features pooled from the multiple attention maps. The classifier is trained with BCE loss on the explicit positive and negative labels. For the missing (unknown) labels, we find treating them as "soft negatives" and assigning them with very small weights in the BCE loss also helps improve results.

than its interior) instead of distributing its attention uniformly over the object. Hence, by setting λ_{rel} to a small positive constant less than 1, we prioritize the need for G to *not* attend to non-object pixels over the need to uniformly attend to object pixels.

Multi-attention module. Object localization is beneficial for recognizing several attributes such as *color, material, texture,* and *shape*, but might be too restrictive for attributes that require attention to different object parts or the background. For example: *bald-headed* or *bare-footed* requires looking at a person's head or foot; distinguishing different activities (e.g., *jumping* vs *crouching*) might require information from surrounding context. Therefore, we utilize a free-form multi-attention mechanism to allow our algorithm to attend to features at different spatial locations.

There are two extreme cases to apply spatial attention [28]: (1) one attention map for all attributes and (2) one attention map per attribute [54]. The first approach is similar to using the object foreground which is unlike what we are aiming for. The latter allows for more control but does not scale well with large number of attributes. Hence, we opt for a hybrid multi-attention idea as in [28].

We extract M attention maps $\{A^{(m)}\}_{m=1}^{M}$ from X using $f_{\text{att}}^{(m)}$ which has the same architecture as f_{rel}:

$$E^{(m)} = f_{\text{att}}^{(m)}(X), \quad E^{(m)} \in \mathbb{R}^{H \times W}, m = 1, ..., M \quad (7)$$

$$A_{h,w}^{(m)} = \frac{\exp(E_{h,w}^{(m)})}{\sum_{h,w} \exp(E_{h,w}^{(m)})}, \quad A_{h,w}^{(m)} \in \mathbb{R}^{H \times W}. \quad (8)$$

This is partly similar to [80] where object parts are localized using learned embeddings of these parts. Because the VAW dataset does not have part annotations for every attribute, this approach is not usable in our case. Similar to [28], we employ the following divergence loss to encourage

these attention maps to focus on different regions:

$$\mathcal{L}_{\text{div}} = \sum_{m \neq n} \frac{\langle E^{(m)}, E^{(n)} \rangle}{\|E^{(m)}\|_2 \|E^{(n)}\|_2}. \quad (9)$$

Using the computed M attention maps, we aggregate M feature vectors $\{r^{(m)}\}_{m=1}^{M}$ from X and pass them through a projection layer to obtain their final representations:

$$r^{(m)} = \sum_{h,w} A_{h,w}^{(m)} X_{h,w}, \quad r^{(m)} \in \mathbb{R}^{D}, \quad (10)$$

$$z_{\text{att}}^{(m)} = f_{\text{proj}}^{(m)}(r^{(m)}), \quad z_{\text{att}}^{(m)} \in \mathbb{R}^{D_{\text{proj}}}. \quad (11)$$

Our final multi-attention feature is the concatenation of all individual attention features:

$$Z_{\text{att}} = \text{concat}([z_{\text{att}}^{(1)}, ..., z_{\text{att}}^{(M)}]). \quad (12)$$

4.4. Loss function and training paradigm

Our final feature vector is the concatenation of the localized object and the multi-attention feature. In addition, we also find using low-level feature from early blocks improves accuracy for low-level attributes (*color, material*). Therefore, we also pool low-level features from the estimated object region G to construct Z_{low}. The input to the classification layer is $[Z_{\text{low}}, Z_{\text{rel}}, Z_{\text{att}}]$, and we use a linear classifier with C output logit values followed by sigmoid.

Let $\hat{Y} = [\hat{y}_1, ..., \hat{y}_C]$ be the output of the classification layer. We apply the following reweighted binary cross-entropy loss that takes data imbalance into account:

$$\mathcal{L}_{\text{bce}}(Y, \hat{Y}) = -\sum_{c=1}^{C} w_c \big(\mathbb{1}_{[y_c=1]} p_c \log(\hat{y}_c) \quad (13)$$
$$+ \mathbb{1}_{[y_c=0]} n_c \log(1 - \hat{y}_c) \big),$$

where w_c, p_c, n_c are respectively the reweighting factors for attribute c, its positive, and its negative example. Let n_c^{pos} and n_c^{neg} be the number of positives and negatives of attribute c. First, we want w_c to reflect the importance of the rare attributes by setting w_c inversely proportional to its number of positive examples: $w_c \propto 1/(n_c^{pos})^\alpha$ and normalize so that $\sum_c w_c = C$ [14] (α is a smoothing factor). Second, we want to balance between the effect of positive and negative examples. We apply the same idea by setting $p_c \propto 1/(n_c^{pos})^\alpha$, $n_c \propto 1/(n_c^{neg})^\alpha$ and normalize so that $p_c + n_c = 2$. As a result, the ratio between the positive and negative becomes $p_c/n_c = (n_c^{neg}/n_c^{pos})^\alpha$, which helps balance out their effect based on their frequency.

Our re-weighted BCE (termed RW-BCE) is different from [15], where the authors propose to reweigh each sample based on its proportion of available labels (*i.e.*, an instance with less number of available labels is assigned a larger weight). We posit this is not ideal because the number of labels for an instance should not affect loss computation (*e.g.*, loss for *red* should be the same between a *red car* instance and a *large shiny red car* instance). Our final loss is a combination of all loss functions presented above:

$$\mathcal{L} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{rel}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}. \tag{14}$$

Empirically, we find applying repeat factor sampling (RFS) [20, 47] with RW-BCE works well. RFS is a method that defines a repeat factor for every image based on the rarity of the labels it contains. Therefore, we employ both RW-BCE and RFS (referred as RR) in training our model.

4.5. Negative label expansion

While our dataset provides unprecedented amount of explicitly labeled negative annotations, the amount of possible negatives still far outnumbers the number of possible positive attributes. Because many attributes are mutually exclusive (*i.e.*, presence of attribute *clean* implies absence of attribute *dirty*), we seek to use existing linguistic and external knowledge tools to expand the set of negative annotations.

Consider attribute type A (*e.g.*, *material*), the following observations can be made about its attributes: (1) there exists *overlapping* relation between some attributes due to their visual similarity or them being hierarchically related (*e.g.*, *wooden* overlaps with *wicker*); (2) there exists *exclusive* relation where two attributes cannot appear on the same object (*e.g.*, *wet* vs. *dry*). Therefore, for an object labeled with attribute $a \in A$, we can generate negative labels for it from the set $\{a' \in A \mid \neg overlap(a,a') \lor exclusive(a,a')\}$.

We classify the attributes into types and construct their overlapping and exclusive relations using existing ontology from a related work [21], WordNet [48], and the relation edges from ConceptNetAPI [59]. We further expand the overlapping relations based on the co-occurrence (by using conditional probability) of the attribute pairs (*e.g.*, *white* and

beige are similar and often mistaken by human annotators). More details about how these relations are constructed are presented in the supplementary material. Our negative label expansion scheme allows to add 5.9M negative annotations to our training set. Aside from the extra negatives, one strong point of this approach is that when we want to label a novel attribute class, we can use the same approach to discover its relationship with existing attributes in the dataset and attain free negatives for the new class.

4.6. Supervised contrastive learning

[75] shows with success that imbalanced learning could benefit from self-supervised pretraining on both labeled and unlabeled data, where a network can be better initialized by originally avoiding strong label bias due to data imbalance. Also motivated by [36], we propose to use supervised contrastive (SupCon) pretraining for our attribute learning with partial labels problem, where we extend the SupCon loss from a single-label to a multi-label setting.

We perform mean-pooling inside the feature map X to obtain $x \in \mathbb{R}^D$. We follow the design of SimCLR [11] and add a projection layer to map $z = Proj(x) \in \mathbb{R}^{128}$. The projection layer is an MLP with hidden size 2048 and is only used during pretraining. In a multi-label setting, it is not trivial how to pull two samples together since they can share some labels but different in terms of other labels. Motivated by [49, 25], we propose to represent each attribute c as a matrix $A_c \in \mathbb{R}^{128 \times 128}$ that linearly projects z into an attribute-aware embedding space $z_c = A_c z$, which is then ℓ_2-normalized onto the unit hypersphere. With this, samples that share the same attribute can have their respective attribute-aware embeddings pulled together.

In the pretraining stage, we construct a batch of $2N$ sample-label vector pairs $\{I_i, Y_i\}_{i=1}^{2N}$, where I_{2k} and I_{2k-1} ($k = 1...N$) are two views (from random augmentation) of the same object image and $Y_{2k} = Y_{2k-1}$. Let $z_{i,c}$ be the c-attribute-aware embedding of I_i, and $B(i) = \{c \in C : Y_{i,c} = 1\}$ is the set of positive attributes of I_i. We reuse notations from [36]: $K \equiv \{1...2N\}$, $A(i) \equiv K \setminus \{i\}$, $P(i,c) \equiv \{p \in A(i) : Y_{p,c} = Y_{i,c}\}$ and use the following SupCon loss

$$\mathcal{L}_{\text{sup}} = \sum_{i=1}^{2N} \sum_{c \in B(i)} \frac{-1}{|P(i,c)|} \sum_{p \in P(i,c)} \log \frac{\exp(z_{i,c} \cdot z_{p,c}/\tau)}{\sum_{j \in A(i)} \exp(z_{i,c} \cdot z_{j,c}/\tau)}. \tag{15}$$

Linear transformation using A_c, followed by dot product in the SupCon loss, implements an inner product in the embedding space of z, which can be interpreted as finding part of z that encodes the attribute c [25]. Therefore, our approach fits nicely into the multi-label setting where an image embedding vector z can simultaneously encode multiple attribute labels that can be probed by these linear transformations for contrasting in the SupCon loss.

After the pretraining stage, we keep the backbone encoder and the image-object composition module and fine-tune them along with the classification layer.

While SupCon is designed to be used for pretraining, empirically, we find it hampers the multi-attention module ability to focus on specific regions. To reconcile this difference, we find it is empirically better to minimize $\mathcal{L}_{\text{sup}}$ jointly with the other loss. For other models that do not use attention (vanilla ResNet), we find SupCon pretraining still effective.

5. Experiments

In this section, we discuss the evaluation metrics and report the results of our attribute prediction framework and other related baselines on the VAW dataset. The implementation details are presented in the supplementary material.

5.1. Algorithms for VAW Dataset

We consider the following baselines and state-of-the-art multi-label learning approaches and compare them to our SCoNE algorithm. We made our best attempt to modify the authors' implementation (if available) to include the image-object composition module in section 4.2. All models use ResNet-50 as their backbone and use BCE loss (except LSEP and ResNet-Baseline-CE) for training. Empirically, we find treating missing labels as negatives and assigning them with very small weights in the BCE loss also improves results. Hence, we apply this for all methods.

- **ResNet-Baseline:** ResNet-50 followed by image-object composition and classification layer.
- **ResNet-Baseline-CE:** Similar as above, but uses softmax cross entropy loss. This is used by [3, 31] to train attribute prediction head for object detectors on Visual Genome.
- **Strong Baseline (SB):** The combination of our image-object composition, multi-attention, and object localizer.
- **LSEP** [43] **:** Uses ranking loss and label threshold estimation method to predict which attributes to output.
- **ML-GCN** [13] **:** Uses graph convolution network to predict classifier weights based on the 100-d GloVe embeddings of the attribute names. Label correlation graph is constructed following the authors' implementation.
- **Durand *et al*. (Partial BCE + GNN)** [15] **:** BCE loss reweighted by the authors' reweighting scheme. Graph neural network is applied on the output logits.
- **Sarafianos *et al*.** [54] **:** one of the SOTAs in pedestrian attribute prediction that also uses multi-attention.

5.2. Evaluation metrics

We employ the following metrics to measure attribute prediction from different perspectives. Detailed discussion of these metrics can be found in the supplementary.
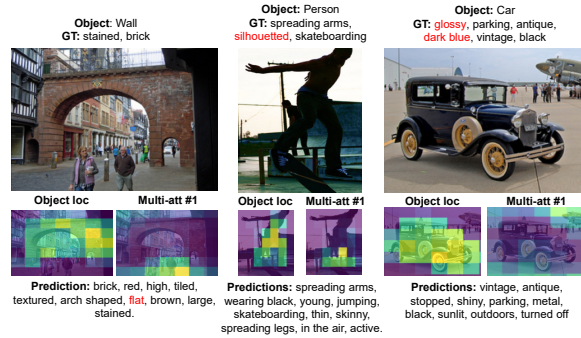

Figure 3. **Qualitative results.** Examples of predictions from SB+SCoNE. We show the object name and its ground truth positive attribute labels above the image. The object localized region, attention map #1, and model top-10 predictions are shown below. Red text represents missed or incorrect predictions.

mAP: mean average precision over all classes. mAP is a popular metric for attribute prediction and multi-label learning [51, 13, 64]. Because VAW is partially labeled, we only evaluate mAP on the annotated data as in [64].

mR@15: mean recall over all classes at top 15 predictions in each image. Recall@K is often used in datasets that are not exhaustively labeled such as scene graph generation [77, 12]. This is also used in multi-label learning [13, 44, 28] under the name 'per-class recall'.

F1@15: as the above metric may be biased towards infrequent classes, we also report overall F1 at top 15 predictions in each image. Because VAW is partially labeled, we only evaluate the prediction of label that has been annotated.

mA: mean balanced accuracy over all classes using threshold at 0.5 to separate between positive and negative prediction. This metric is used in pedestrian and human facial attribute works [54, 42].

5.3. Results

Overall results are shown in Table 2, where SB and SB+SCoNE are compared with other baselines and state-of-the-art algorithms. In overall, SB is better than other baselines in almost all metrics except for mR@15 where it is lower than ResNet-Baseline-CE. This shows that the object localizer and multi-attention are effective in attribute prediction. ResNet-Baseline-CE, which is adopted by [3, 31], has good recall but very low precision (mAP and F1). This is in contrast to ResNet-Baseline which is trained with BCE.

SB+SCoNE substantially improves over SB in all metrics and clearly surpasses available algorithms by a large margin. It is particularly effective in long-tail attributes where it outperforms its closest competitor (other than SB) by 5 mAP points, and is also highly effective in detecting color and material attributes where it is nearly 7-8 mAP points higher than the next-best method. This shows that our attribute learning paradigm, including the negative la-

Methods	Overall				Class imbalance (mAP)			Attribute types (mAP)						
	mAP	mR@15	mA	F1@15	Head	Medium	Tail	Color	Material	Shape	Size	Texture	Action	Others
LSEP [44]	61.0	50.7	67.1	62.3	69.1	57.3	40.9	56.1	67.1	63.1	61.4	58.7	50.7	64.9
ML-GCN [13]	63.0	52.8	69.5	64.1	70.8	59.8	42.7	59.1	64.7	65.2	64.2	62.8	54.7	66.5
Partial-BCE + GNN [15]	62.3	52.3	68.9	63.9	70.1	58.7	40.1	57.7	66.5	64.1	65.1	59.3	54.4	65.9
ResNet-Baseline [51]	63.0	52.1	68.6	63.9	71.1	59.4	43.0	58.5	66.3	65.0	64.5	63.1	53.1	66.7
ResNet-Baseline-CE [3, 31]	56.4	55.8	50.3	61.5	64.6	52.7	35.9	54.0	64.6	55.9	56.9	54.6	47.5	59.2
Sarafianos *et al.* [54]	64.6	51.1	68.3	64.6	72.5	61.5	42.9	62.9	68.8	64.9	65.7	62.3	56.6	67.4
Strong Baseline (SB)	65.9	52.9	69.5	65.3	73.6	62.5	46.0	64.5	68.9	67.1	65.7	66.1	57.2	68.7
SB + SCoNE (Ours)	**68.3**	**58.3**	**71.5**	**70.3**	**76.5**	**64.8**	**48.0**	**70.4**	**75.6**	**68.3**	**69.4**	**68.4**	**60.7**	**69.5**

Table 2. **Experimental results compared with baselines and SOTA multi-label learning methods.** The top box displays results of multi-label learning methods; the middle box shows results of models from attribute prediction works and our strong baseline; the last row shows performance of our SCoNE algorithm applied onto the strong baseline.

Methods	mAP	mR@15	mA	F1@15
Strong Baseline (SB)	65.9	52.9	69.5	65.3
+ Negative	67.7	54.3	70.0	69.6
+ Neg + SupCon	68.2	55.2	70.3	70.0
+ Neg + SupCon + RR (SCoNE)	**68.3**	**58.3**	**71.5**	**70.3**
ResNet-Baseline	63.0	52.1	68.6	63.9
+ SCoNE	66.4	56.8	70.7	68.8

Table 3. **Ablation study.** We show how each of our proposed techniques help improve overall performance.

bel expansion, supervised contrastive loss, reweighted and resampling scheme is clearly effective in attribute learning.

5.4. Ablation studies

Table 3 shows effect of different components of SCoNE. Starting from our SB, we can see that each of our model choices substantially improves its performance, with the biggest mAP improvements provided by our negative label expansion scheme. Each of the components of SCoNE also stacks additively, with our final model performing 2.4 mAP, 5.4 mR@15, and 5 F1@15 points over SB. Moreover, the components of SCoNE are *model agnostic*. We verify that by enhancing our ResNet-Baseline with SCoNE, which also improves its mAP and mR@15 by 3.4 and 4.7 points.

Due to lack of space, several additional ablation experiments are included in the supplementary material.

5.5. Qualitative examples

Figure 3 shows qualitative results of our SB+SCoNE model, which clearly showcases its various strengths. Firstly, we clearly show a robust ability of the model to predict a variety of attribute types of different objects with good accuracy. Next, our object localizer shows remarkable ability to find the correct object of interest and ignore background and other distracting objects (in 3a, the ground is not attended by the object localizer). Next, our multi-attention module often works to complement our object localizer by attending to relevant image regions that may be outside of the object region. For example, in figure 3b, the activity

skateboarding is easier to predict if our model can look at the skateboard, but it is outside the *person* region. Here, our multi-attention correctly learns to look at appropriate image regions that can help our model determine the *person* has an attribute *skateboarding*. More results can be found in the supplementary material.

6. Discussion, Future Works and Conclusion

VAW is a first-of-its kind large-scale object attribute prediction dataset in the wild. We explored various challenges posed by the VAW dataset and also discussed efficacy of current models towards this task. Our SCoNE algorithm proposed several novel algorithmic improvements that have helped us improve performance in the VAW dataset compared to our strong baseline by 2.4 mAP and and 5.4 mR@15 points. Despite our results, there are several outstanding challenges remaining to be solved in VAW.

Data Imbalance: Reweighting and resampling techniques have helped considerably improve the performance of tail categories in VAW dataset. However, even for our best model, mAP for tail categories still lags more than 25 points behind our head category. Similar to many vision and language problems [34], this is one considerable challenge for future works in this space.

Object-bias effect: Using object label as input is crucial to obtain good results in VAW, but it may also introduce object-bias in predictions. Ideally, an algorithm should be able to make robust predictions for compositionally novel instance. While not in scope of current paper, this can be explored in detail by redistributing train-test split in compositionally novel patterns [1, 53, 63].

In conclusion, we believe that VAW can serve as an important benchmark not only for attribute prediction in the wild, but also as a generic test for long-tailed multi-label prediction task with limited labels, data imbalance, out-of-distribution testing and bias-related issues.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018.

[2] Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5975–5984, 2016.

[3] 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.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[5] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.

[6] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010.

[7] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016.

[8] Serhat Selcuk Bucak, Rong Jin, and Anil K Jain. Multi-label learning with incomplete class assignments. In *CVPR 2011*, pages 2801–2808. IEEE, 2011.

[9] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[10] S. Chen and K. Grauman. Compare and contrast: Learning prominent visual differences. In *CVPR*, June 2018.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[12] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019.

[13] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019.

[14] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[15] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 647–657, 2019.

[16] Ali Farhadi, Ian Endres, and Derek Hoiem. Attribute-centric recognition for cross-category generalization. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2352–2359. IEEE, 2010.

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

[18] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *Advances in neural information processing systems*, pages 433–440, 2008.

[19] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Hartwig Adam, Matthew R Scott, and Serge Belongie. The imaterialist fashion attribute dataset. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[20] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019.

[21] Chi Han, Jiayuan Mao, Chuang Gan, Joshua B. Tenenbaum, and Jiajun Wu. Visual Concept Metaconcept Learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[22] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[23] 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.

[24] 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.

[25] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

[26] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[27] 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 Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

[28] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8776–8786, 2020.

[29] 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.

[30] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *Advances in neural information processing systems*, pages 3464–3472, 2014.

[31] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020.

[32] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[33] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 2017.

[34] Kushal Kafle, Robik Shrestha, and Christopher Kanan. Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*, 2:28, 2019.

[35] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[36] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[37] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1667–1675, 2017.

[38] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[39] Neeraj Kumar, Peter Belhumeur, and Shree Nayar. Facetracer: A search engine for large collections of images with faces. In *European conference on computer vision*, pages 340–353. Springer, 2008.

[40] Kaustav Kundu and Joseph Tighe. Exploiting weakly supervised visual patterns to learn from partial annotations. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[42] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016.

[43] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *European Conference on Computer Vision*, pages 684–700. Springer, 2016.

[44] Yuncheng Li, Yale Song, and Jiebo Luo. Improving pairwise ranking for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3617–3625, 2017.

[45] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[46] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016.

[47] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.

[48] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[49] 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.

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

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

[52] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

[54] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 680–697, 2018.

[55] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neu-

ral networks. In *European conference on computer vision*, pages 467–482. Springer, 2016.

[56] Minchul Shin. Semi-supervised learning with a teacher-student network for generalized attribute prediction. *arXiv preprint arXiv:2007.06769*, 2020.

[57] Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR 2011*, pages 801–808. IEEE, 2011.

[58] Yaser Souri, Erfan Noury, and Ehsan Adeli. Deep relative attributes. In *Asian conference on computer vision*, pages 118–133. Springer, 2016.

[59] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975*, 2016.

[60] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[61] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. In *Proceedings of the twenty-fourth AAAI conference on artificial intelligence*, pages 593–598, 2010.

[62] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7134–7143, 2019.

[63] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of good-hart's law. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[64] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847, 2017.

[65] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2019.

[66] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2019.

[67] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017.

[68] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE international conference on computer vision*, pages 464–472, 2017.

[69] Baoyuan Wu, Fan Jia, Wei Liu, Bernard Ghanem, and Siwei Lyu. Multi-label learning with missing labels using mixed dependency graphs. *International Journal of Computer Vision*, 126(8):875–896, 2018.

[70] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020.

[71] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[72] 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.

[73] 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, 2015.

[74] Hao Yang, Joey Tianyi Zhou, and Jianfei Cai. Improving multi-label learning with missing labels by structured semantic correlations. In *European conference on computer vision*, pages 835–851. Springer, 2016.

[75] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[76] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *International conference on machine learning*, pages 593–601, 2014.

[77] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *European Conference on Computer Vision*, pages 606–623. Springer, 2020.

[78] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021.

[79] Sanyi Zhang, Zhanjie Song, Xiaochun Cao, Hua Zhang, and Jie Zhou. Task-aware attention model for clothing attribute prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1051–1064, 2019.

[80] 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.

[81] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.