VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

Wenjia Xu[1,7,8] Yongqin Xian[2] Jiuniu Wang[5,7,8] Bernt Schiele[3] Zeynep Akata[3,4,6]

[1] Beijing University of Posts and Telecommunications [2] ETH Zurich
[3] Max Planck Institute for Informatics [4] University of Tübingen [5] City University of Hong Kong
[6] Max Planck Institute for Intelligent Systems
[7] University of Chinese Academy of Sciences [8] Aerospace Information Research Institute, CAS

Abstract

Human-annotated attributes serve as powerful semantic embeddings in zero-shot learning. However, their annotation process is labor-intensive and needs expert supervision. Current unsupervised semantic embeddings, i.e., word embeddings, enable knowledge transfer between classes. However, word embeddings do not always reflect visual similarities and result in inferior zero-shot performance. We propose to discover semantic embeddings containing discriminative visual properties for zero-shot learning, without requiring any human annotation. Our model visually divides a set of images from seen classes into clusters of local image regions according to their visual similarity, and further imposes their class discrimination and semantic relatedness. To associate these clusters with previously unseen classes, we use external knowledge, e.g., word embeddings and propose a novel class relation discovery module. Through quantitative and qualitative evaluation, we demonstrate that our model discovers semantic embeddings that model the visual properties of both seen and unseen classes. Furthermore, we demonstrate on three benchmarks that our visually-grounded semantic embeddings further improve performance over word embeddings across various ZSL models by a large margin. Code is available at https://github.com/wenjiaXu/VGSE

1. Introduction

Semantic embeddings aggregated for every class live in a vector space that associates different classes even when visual examples of these classes are not available. Therefore, they facilitate the knowledge transfer in zero-shot learning (ZSL) [1, 28, 42, 59] and are used as side-information in other computer vision tasks like fashion trend forecast [4, 23, 64], face recognition and manipulation [11, 27, 29], and domain adaptation [10, 24].

Human annotated attributes [19, 36, 55], characteristic properties of objects annotated by human experts, are widely used as semantic embeddings [61, 62]. However, obtaining attributes is often a labor-intensive two-step process. First, domain experts carefully design an attribute vocabulary, e.g., color, shape, etc., and then human annotators indicate the presence or absence of an attribute in an image or a class (as shown in Figure 1). The labeling effort devoted to human-annotated attributes hinders its applicability of performing zero-shot learning for more datasets in realistic settings [30].

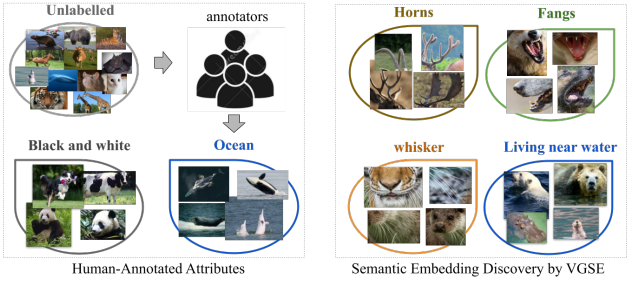


Figure 1. Human-annotated attributes (left) are labor-intensive to collect, and may neglect some local visual properties shared between classes. We propose to discover semantic embeddings via visually clustering image patches and predicting the class relations.

Previous works tackle this problem by using word embeddings for class names [31, 38], or semantic embeddings from online encyclopedia articles [3, 39, 67]. Though they model the semantic relation between classes without using human annotation, some of these relations may not be visually detectable by machines, resulting in a poor performance in zero-shot learning. Similarly, discriminative visual cues may not all be represented in those semantic embeddings.

To this end, we propose the Visually-Grounded Semantic Embedding (VGSE) Network to discover semantic embeddings with minimal human supervision (we only use category labels for seen class images). Our network explicitly explores visual clusters that relate image regions from different categories, which is useful for knowledge transfer between classes under zero-shot learning settings (see our

learnt clusters in Figure 1). To fully unearth the visual properties shared across different categories, our model discovers semantic embeddings by assigning image patches into various clusters according to their visual similarity. Besides, we further impose class discrimination and semantic relatedness of the semantic embeddings, to benefit their ability in transferring knowledge between classes in ZSL.

To sum up, our work makes the following contributions. (1) We propose a visually-grounded semantic embedding (VGSE) network that learns visual clusters from seen classes, and automatically predicts the semantic embeddings for each category by building the relationship between seen and unseen classes given unsupervised external knowledge sources. (2) On three zero-shot learning benchmarks (i.e. AWA2, CUB, and SUN), our learned VGSE semantic embeddings consistently improve the performance of word embeddings over five SOTA methods. (3) Through qualitative evaluation and user study, we demonstrate that our VGSE embeddings contain rich visual information like fine-grained attributes, and convey human-understandable semantics that facilitates knowledge transfer between classes.

2. Related Work

Zero-shot Learning aims to classify images from novel classes that do not appear during training. Existing ZSL methods usually assume that both the seen and unseen classes share a common semantic space, thus the key insight of performing ZSL is to transfer knowledge from seen classes to unseen classes. To assign the image to a semantic class embedding, many classical approaches learn a compatibility function to associate visual and semantic space [1, 20, 48, 58]. Recent works mainly focus on synthesizing image features or classifier weights with a generative model [43, 60, 61], or training enhanced image features extractors with visual attention [66, 68] or local prototypes [62].

Semantic embeddings are crucial in relating different categories with shared characteristics, i.e., the semantic space. Despite their importance, semantic embeddings are relatively under-explored in zero-shot learning. Human-annotated attributes [19, 36, 55, 59], i.e., the properties of objects such as color and shape, are the most commonly used semantic embeddings in zero-shot learning. Though the attributes can be discriminative for each class, their annotation process is labor-intensive and require expert knowledge [50, 55, 65]. We propose to discover visual properties through patch-level clustering over image datasets, and predict semantic embeddings automatically, where no additional human annotation is required except for the class labels of seen class images.

Semantic Embeddings with Minimal Supervision is drawing attention in image classification [6, 9, 26, 40, 45], transfer learning [10, 37, 54] and low-shot learning problems [3, 25, 33, 44, 50, 65]. Semantic embeddings collected from text corpora are alternatives to manual annotations, which include word embeddings learned from large corpora [31, 38, 49, 63], semantic relations such as knowledge graphs [9, 26, 56], and semantic similarities [12, 57], etc. More recently, [3, 39, 41, 67] collect attribute-class associations from online encyclopedia articles that describe each category. The semantic similarity can be encoded by a taxonomical hierarchy or by incorporating co-occurrence statistics of words within the document. However, this may not reflect visual similarity, e.g., *sheep* is semantically close to *dog* since they often co-occur in online articles, while visually *sheep* is closer to a *deer*. We focus on discovering visually-grounded semantic embeddings in the image space, and further incorporate the semantic relations between classes into our semantic embedding for better zero-shot knowledge transfer.

Learning Visual Properties from Image Patches. Previous attempts for discovering middle-level representations for classification include exploring image-level embeddings by learning binary codes or classeme representations [6, 40, 51], and further introducing humans in the loop to discover localized and nameable attributes [18, 35]. However, those methods discover properties depicted in the whole image, which might result in a combination of several semantics covering several objects (parts) that are hard to interpret [35]. Visual transformer [17] and BagNets [8] showed that image patches can work as powerful visual words conveying visual cues for class discrimination. Bag of visual words (BOVW) models [13, 47] propose to cluster image patches to learn a codebook and form image representations. However, BOVW extracts hand-crafted features followed by k-means clustering, while we learn clustering in an end-to-end manner via deep neural networks. Considering the above problem, we propose to learn visual properties by clustering image patches, and predict the semantic embeddings with the visual properties depicted by patch clusters.

More closely related to our work are the ones learning discriminative image regions that can represent each class through clustering of local patches [15, 16, 45, 46], e.g., finding representative elements to discriminate one class from others. Instead of picking up the most salient patches in each class, we aim to learn visual properties that are shared among different classes for most of the image patches appearing in the dataset. Besides, unlike some above methods that divide an image into a grid of square patches, we propose to use segmentation-based region proposals to obtain semantic image regions (e.g., the entire head could represent one semantic region).

3. Visually-Grounded Semantic Embedding

We are interested in the (generalized) zero-shot learning task where the training and test classes are disjoint sets. The training set $\{(x_n, y_n) | x_n \in X^s, y_n \in Y^s\}_{n=1}^{N_s}$ consists of images x_n and their labels y_n from the seen classes Y^s. In

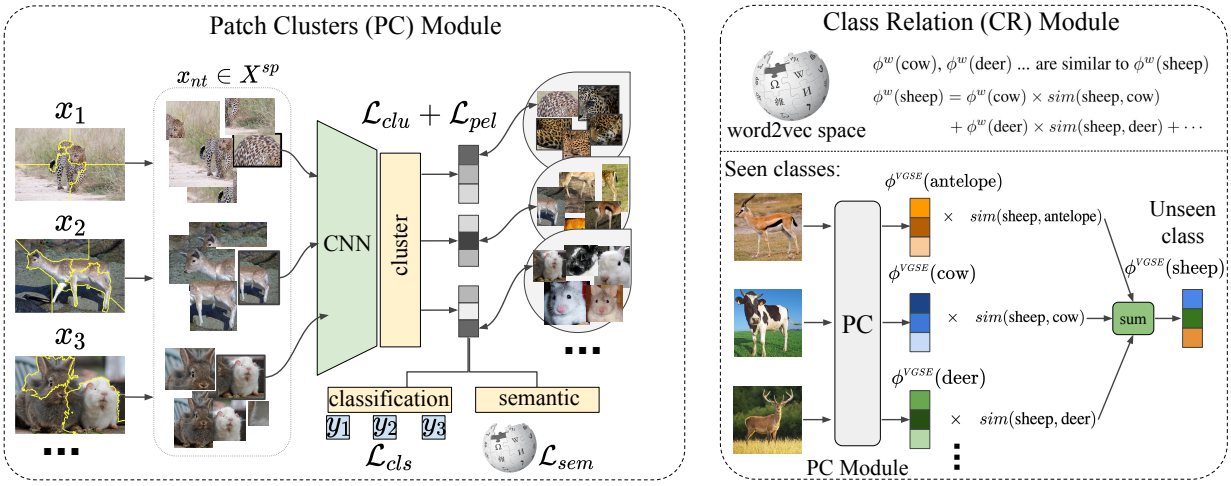


Figure 2. Our visually-grounded semantic embedding network consists of two modules. The Patch Clustering (PC) module learns clusters from patch images, and predicts semantic embeddings for seen classes with their images. The Class Relation (CR) module predicts the unseen class embeddings $\phi^{VGSE}(y_m)$ using unseen and seen class relations learned from external knowledge, e.g., word2vec. For instance, the embedding for unseen class *sheep* is predicted using the semantic embeddings of the seen classes, e.g., *antelope*, *cow*, *deer*, and so on.

the ZSL setting, test images are classified into unseen classes Y^u, and in the GZSL setting, into both Y^s and Y^u with the help of a semantic embedding space, e.g., human annotated attributes. Since human-annotated attributes are costly to obtain, while prior unsupervised semantic embeddings are incomplete to describe the rich visual world, we propose to automatically discover a set of D_v visual clusters as the semantic embedding, denoted by $\Phi^{VGSE} \in \mathbb{R}^{(|Y^u|+|Y^s|)\times D_v}$. The semantic embeddings for seen classes $\{\phi^{VGSE}(y)\,|y \in Y^s\}$, describing diverse visual properties of each category, are learned on seen classes images X^s. The semantic embeddings for unseen classes $\{\phi^{VGSE}(y)\,|y \in Y^u\}$ is predicted with the help of unsupervised word embeddings, e.g., w2v embeddings for class names $\Phi^w \in \mathbb{R}^{(|Y^u|+|Y^s|)\times D_w}$.

Our Visually-Grounded Semantic Embedding (VGSE) Network (see Figure 2) consists of two main modules. (1) The Patch Clustering (PC) module takes the training dataset as input, and clusters the image patches into D_v visual clusters. Given one input image x_n, PC can predict the cluster probability $a_n \in \mathbb{R}^{D_v}$ indicating how likely the image would contain the visual property appearing in each cluster. (2) Since unseen class images cannot be observed during training, we propose the Class Relation (CR) module to infer the semantic embeddings of unseen classes. Finally, the learned semantic embedding Φ^{VGSE} can be used to perform downstream tasks, e.g., Zero-Shot Learning.

3.1. Patch Clustering (PC) Module

Patch image generation. Patch-level embeddings allow us to explore the visual properties that appear in local image regions [17, 55], e.g., the shape and texture of animal body parts or the objects in scenes. To obtain image patches that cover the entire semantic image region (e.g. an animal head), we segment an image into regularly shaped regions via an unsupervised compact watershed segmentation algorithm [32]. As shown in Figure 2, for each image x_n, we find the smallest bounding box that fully covers each segment and crop x into N_t patches $\{x_{nt}\}_{t=1}^{N_t}$ that cover different parts of the image. The number of patches N_t is empirically set to be around 9, as we observed in initial experiments that larger patches may include too many attributes, while smaller patches will be too tiny to contain any visual attribute. In this way, we reconstruct our training set consisting of image patches $\{(x_{nt}, y_n)\,|x_{nt} \in X^{sp}, y_n \in Y^s\}_{n=1}^{N_s}$, here $|X^{sp}| = N_s N_t$, and N_s is the train set size.

Patch clustering. Our patch clustering module is a differentiable middle layer, that simultaneously learns image patch representations and clustering. As shown in Figure 2 (left), we start from a deep neural network that extracts patch feature $\theta(x_{nt}) \in \mathbb{R}^{D_f}$, where we use a ResNet [22] pretrained on ImageNet [14] as in other ZSL models [59, 61]. Afterwards, a clustering layer $H : \mathbb{R}^{D_f} \rightarrow \mathbb{R}^{D_v}$ converts the feature representation into cluster scores:

$$a_{nt} = H \circ \theta(x_{nt}) \,, \tag{1}$$

where a_{nt}^k (the k-th element of a_{nt}) indicates the probability of assigning image patch x_{nt} to cluster k, e.g., the patch clusters of spotty fur, fluffy head in Figure 2.

A pretext task can be adopted to obtain semantically meaningful representations [21, 34, 53] in an unsupervised manner. Our pretext task [53] enforces the image patch x_{nt} and its neighbors being predicted to the same clusters. We

retrieve nearest patch neighbors of x_{nt} as X^{sp}_{nb} by the $\mathcal{L}_2$ distance of patch features $\|\theta(x_{nt}) - \theta(x_i)\|_2$, where $x_i \in X^{sp}$ and $x_i \neq x_{nt}$. The clustering loss is defined as

$$\mathcal{L}_{clu} = -\sum_{x_{nt}\in X^{sp}} \sum_{x_i \in X^{sp}_{nb}} log(a^T_{nt} a_i)\,, \tag{2}$$

where $a_i = H \circ \theta(x_i)$. $\mathcal{L}_{clu}$ imposes consistent cluster assignment for x_{nt} and its neighbors. To avoid all images being assigned to the same cluster, we follow [53] to add an entropy penalty as follows:

$$\mathcal{L}_{pel} = \sum_{k=1}^{D_v} \bar{a}^k_{nt} \log \bar{a}^k_{nt}\,, \quad \bar{a}^k_{nt} = \frac{1}{N_s N_t} \sum_{x_{nt}\in X^{sp}} a^k_{nt}\,, \tag{3}$$

ensuring that images are spread uniformly over all clusters.

Class discrimination. To impose class discrimination information into the learnt clusters, we propose to apply an cluster-to-class layer $Q : \mathbb{R}^{D_v} \rightarrow \mathbb{R}^{|Y^s|}$ to map the cluster prediction of each image to the class probability, i.e., $p(y|x_{nt}) = softmax(Q \circ \theta(x_{nt}))$. We train this module with the following cross-entropy loss,

$$\mathcal{L}_{cls} = -\log \frac{\exp(p(y_n|x_{nt}))}{\sum_{\hat{y}\in Y^s} \exp(p(\hat{y}|x_{nt}))}\,. \tag{4}$$

Semantic relatedness. We further encourage the learned visual clusters to be transferable between classes, to benefit the downstream zero-shot learning tasks. We learn clusters shared between semantically related classes, e.g., *horse* share more semantic information with *deer* than with *dolphin*. We implement this by mapping the learned cluster probability to the semantic space constructed by w2v embeddings Φ^w. The cluster-to-semantic layer $S : \mathbb{R}^{D_v} \rightarrow \mathbb{R}^{D_w}$ is trained by regressing the w2v embedding for each class,

$$\mathcal{L}_{sem} = \|S \circ a_{nt} - \phi^w(y_n)\|_2\,, \tag{5}$$

where y_n denotes the ground truth class, and $\phi^w(y_n) \in \mathbb{R}^{D_w}$ represents the w2v embedding for the class y_n.

The overall objective for training the model is as follows:

$$\mathcal{L} = \mathcal{L}_{clu} + \lambda \mathcal{L}_{pel} + \beta \mathcal{L}_{cls} + \gamma \mathcal{L}_{sem}\,. \tag{6}$$

Predict seen semantic embeddings. After we learned the visual clusters, given one input image patch x_{nt}, the model extracts the feature $\theta(x_{nt})$ followed by predicting the cluster probability $a_{nt} = H \circ \theta(x_{nt}) \in \mathbb{R}^{D_v}$ where each dimension indicates the likelihood that the image patch x_{nt} being assigned to a certain cluster learned by this module.

The image embedding $a_n \in \mathbb{R}^{D_v}$ for x_n is calculated by averaging the patch embedding in that image:

$$a_n = \frac{1}{N_t} \sum_{t=1}^{N_t} a_{nt}\,. \tag{7}$$

Similarly, we calculate the semantic embedding for y_n by averaging the embeddings of all images belonging to y_n:

$$\phi^{VGSE}(y_n) = \frac{1}{|I_i|} \sum_{j\in I_i} a_j\,, \tag{8}$$

where I_i is the indexes of all images belonging to class y_n, and a_j denotes the image embedding of the j-th image.

3.2. Class Relation (CR) Module

While seen semantic embeddings can be estimated from training images using Eq. 8, how to compute the unseen semantic embeddings is not straightforward since their training images are not available. As semantically related categories share common properties, e.g., *sheep* and *cow* both live on grasslands, we propose to learn a Class Relation Module to formulate the similarity between seen classes Y^s and unseen classes Y^u. In general, any external knowledge, e.g., word2vec [31, 38] or human-annotated attributes, can be utilized to formulate the relationship between two classes. Here we use word2vec learned from a large online corpus to minimize the human annotation effort. Below, we present two solutions to learn the class relations: (1) directly averaging the semantic embeddings from the neighbor seen classes in the word2vec spaces, (2) optimizing a similarity matrix between unseen and seen classes.

Weighted Average (WAvg). For unseen class y_m, we first retrieve several nearest class neighbours in seen classes by the similarity measured with $\mathcal{L}_2$ distance over w2v embedding space, and we denote the neighbor classes set as Y^s_{nb}. The semantic embedding vector for y_m is calculated as the weighted combination [5] of seen semantic embeddings:

$$\phi^{VGSE}(y_m) = \frac{1}{|Y^s_{nb}|} \sum_{\tilde{y}\in Y^s_{nb}} sim(y_m, \tilde{y}) \cdot \phi^{VGSE}(\tilde{y})\,, \tag{9}$$

$$sim(y_m, \tilde{y}) = \exp(-\eta \|\phi^w(y_m) - \phi^w(\tilde{y})\|_2)\,, \tag{10}$$

where exp stands for the exponential function and η is a hyperparameter to adjust the similarity weight. We denote our semantic embeddings learned with weighted average strategy as `VGSE-WAvg`.

Similarity Matrix Optimization (SMO). Given the w2v embeddings $\phi^w(Y^s) \in \mathbb{R}^{|Y^s|\times D_w}$ of seen classes and embedding $\phi^w(y_m)$ for unseen class y_m, we learn a similarity mapping $r \in \mathbb{R}^{|Y^s|}$, where r_i denotes the similarity between the unseen class y_m and the i-th seen class. The similarity mapping is learned via the following optimization problem:

$$\begin{aligned} &\min_r \left\|\phi^w(y_m) - r^T \phi^w(Y^s)\right\|_2 \\ &\text{s.t.} \quad \alpha < r < 1 \quad and \quad \sum_{i=1}^{|Y^s|} r_i = 1. \end{aligned} \tag{11}$$

Here α is the lower bound which can be either 0 or -1, indicating whether we only learn positive class relations or we learn negative relations as well. We base this mapping on the assumption that semantic embeddings follow linear analogy, e.g., $\phi^w(\text{king}) - \phi^w(\text{man}) + \phi^w(\text{woman}) \approx \phi^w(\text{queen})$, which holds for w2v embeddings and our semantic embeddings ϕ^{VGSE}. After the mapping is learned, we can predict the semantic embeddings for the unseen class y_m as:

$$\phi^{VGSE}(y_m) = r^T \phi^{VGSE}(Y_s)\,, \tag{12}$$

where the value of each discovered semantic embedding for unseen class y_m is the weighted sum of all seen class semantic embeddings. We denote our semantic embeddings learned with similarity matrix optimization (SMO) as VGSE-SMO.

4. Experiments

After introducing the datasets and experimental settings, we demonstrate that our VGSE outperforms unsupervised word embeddings over three benchmark datasets and this phenomenon generalizes to five SOTA ZSL models (§4.1). With extensive ablation studies, we showcase clustering with images patches is effective for learning the semantic embeddings, and demonstrate the effectiveness of the PC module and CR module (§4.2). In the end, we present visual clusters as qualitative results (§4.3, §4.4).

Dataset. We validate our model on three ZSL benchmark datasets. AWA2 [59] is a coarse-grained dataset for animal categorization, containing $30,475$ images from 50 classes, where 40 classes are seen and 10 are unseen classes. CUB [55] is a fine-grained dataset for bird classification, containing $11,788$ images and 200 classes, where 150 classes are seen and 50 are unseen classes. SUN [36] is also a fine-grained dataset for scene classification, with $14,340$ images coming from 717 scene classes, where 645 classes are seen and 72 are unseen classes.

Implementation details. Specifically, in the patch clustering (PC) module we learn seen-semantic embeddings with train set (seen classes) proposed by [59], the unseen-class embeddings are predicted in the class relation (CR) module without seeing unseen images. We adopt ResNet50 [22] pretrained on ImageNet1K [14] as the backbone. The cluster number D_v is set as 150 for three datasets. For the Weighted Average module in Eq. 9, we set η as 5 for all datasets, and use 5 neighbors for all datasets. For the similarity matrix optimization in Eq. 11, we set α as -1 for AWA2 and CUB, and as 0 for SUN. More details are in the supplementary.

Semantic embeddings for ZSL. To be fair, we compare our VGSE semantic embeddings with other alternatives using the same image features and ZSL models. All the image features are extracted from ResNet101 [22] pretrained on ImageNet [14]. We follow the data split provided by [59]. The semantic embeddings are L2 normalized following [59]. All ablation studies use the SJE [2, 62] as the ZSL model as it is simple to train. Besides, we verify the generalization ability of our semantic embeddings over five state-of-the-art ZSL models with their official code. The non-generative models include SJE [2], APN [62], GEM-ZSL [28], learning a compatibility function between image and semantic embeddings. The generative approaches consist of CADA-VAE [43] and f-VAEGAN-D2 [61], learning a generative model that synthesizes image features of unseen classes from their semantic embeddings. Note that for all ZSL models, we use the same hyperparameters as proposed in their original papers for all semantic embeddings with no hyperparameter tuning.

4.1. Comparing with the State-of-the-Art

We first compare our semantic embeddings VGSE-SMO with the unsupervised word embeddings w2v [31] on three benchmark datasets and five ZSL models. We further compare ours with other state-of-the-art methods that learn semantic embeddings with less human annotation.

VGSE surpasses w2v by a large margin. The results shown in Table 1 demonstrate that our VGSE-SMO semantic embeddings significantly outperform word embedding w2v on all datasets and all ZSL models. Considering the non-generative ZSL models, VGSE-SMO outperform w2v on all three datasets by a large margin. In particular, on AWA2 dataset, when coupled with GEM-ZSL, our VGSE-SMO boosts the ZSL performance of w2v from 50.2% to 58.0%. On the fine-grained datasets CUB and SUN, VGSE-SMO achieves even higher accuracy boosts. For example, when coupled with the APN model, VGSE-SMO increases the ZSL accuracy of CUB from 22.7% to 28.9%, and the accuracy of SUN from 23.6% to 38.1%. These results demonstrate that our approach not only works well on generic object categories, but also has great potential to benefit the challenging fine-grained classification task. VGSE improves the GZSL performance of both seen and unseen classes, yielding a much better harmonic mean (e.g., when trained with SJE, VGSE-SMO improves over the harmonic mean of w2v by 8.0% on AWA2, 10.3% on CUB, and 7.6% on SUN). These results indicate that our VGSE facilitates the model to learn a better compatibility function between image and semantic embeddings, for both seen and unseen classes.

Our VGSE semantic embeddings show great potential on generative models as well. In particular, VGSE coupled with f-VAEGAN-D2 surpasses all other methods by a wide margin on SUN and CUB datasets, i.e., we obtain 35.0% vs 32.7% (w2v) on CUB, and 41.1% vs 39.6% (w2v) on SUN. As our embeddings are more machine detectable than w2v, introducing visual properties to the conditional GAN will allow them to generate more discriminative image features.

VGSE outperforms SOTA weakly supervised ZSL semantic embeddings. We compare VGSE with other works that learn ZSL semantic embeddings with less human annotation.

	ZSL Model	Semantic Embeddings	Zero-Shot Learning AWA2 T1	Zero-Shot Learning CUB T1	Zero-Shot Learning SUN T1	Generalized Zero-Shot Learning AWA2 u	AWA2 s	AWA2 H	CUB u	CUB s	CUB H	SUN u	SUN s	SUN H
Generative	CADA-VAE [43]	w2v [31]	49.0	22.5	37.8	38.6	60.1	47.0	16.3	39.7	23.1	26.0	28.2	27.0
		VGSE-SMO (Ours)	52.7	24.8	40.3	46.9	61.6	53.9	18.3	44.5	25.9	**29.4**	29.6	29.5
	f-VAEGAN-D2 [61]	w2v [31]	58.4	32.7	39.6	46.7	59.0	52.2	23.0	44.5	30.3	25.9	33.3	29.1
		VGSE-SMO (Ours)	61.3	**35.0**	**41.1**	45.7	66.7	54.2	**24.1**	45.7	**31.5**	25.5	**35.7**	**29.8**
Non-Generative	SJE [2]	w2v [31]	53.7	14.4	26.3	39.7	65.3	48.8	13.2	28.6	18.0	19.8	18.6	19.2
		VGSE-SMO (Ours)	62.4	26.1	35.8	46.8	72.3	56.8	16.4	44.7	28.3	28.7	25.2	26.8
	GEM-ZSL [28]	w2v [31]	50.2	25.7	-	40.1	80.0	53.4	11.2	**48.8**	18.2	-	-	-
		VGSE-SMO (Ours)	58.0	29.1	-	49.1	78.2	60.3	13.1	43.0	20.0	-	-	-
	APN [62]	w2v [31]	59.6	22.7	23.6	41.8	75.0	53.7	17.6	29.4	22.1	16.3	15.3	15.8
		VGSE-SMO (Ours)	**64.0**	28.9	38.1	**51.2**	**81.8**	**63.0**	21.9	45.5	29.5	24.1	31.8	27.4

Table 1. Comparing our VGSE-SMO, with w2v semantic embedding over state-of-the-art ZSL models. In ZSL, we measure Top-1 accuracy (**T1**) on unseen classes, in GZSL on seen/unseen (**s/u**) classes and their harmonic mean (**H**). Feature Generating Methods, i.e., f-VAEGAN-D2, and CADA-VAE generating synthetic training samples, and SJE, APN, GEM-ZSL using only real image features.

Semantic Embeddings	External knowledge	Zero-shot learning AWA2	CUB	SUN
w2v [31]	w2v	58.4	32.7	39.6
ZSLNS [39]	T	57.4	27.8	-
GAZSL [67]	T	-	34.4	-
Auto-dis [3]	T	52.0	-	-
CAAP [5]	T and H	55.3	31.9	35.5
VGSE-SMO (Ours)	w2v	**61.3** $\pm$ 0.3	**35.0** $\pm$ 0.2	**41.1** $\pm$ 0.3

Table 2. Comparing with state-of-the-art methods for learning semantic embeddings with less human annotation (T: online textual articles, H: human annotation) using same image features and ZSL model (f-VAEGAN-d2 [61]).

Semantic Embeddings	Zero-shot learning AWA2	CUB	SUN
k-means-SMO	54.5 $\pm$ 0.4	15.0 $\pm$ 0.5	25.2 $\pm$ 0.4
ResNet-SMO	55.3 $\pm$ 0.2	15.4 $\pm$ 0.1	25.1 $\pm$ 0.1
$\mathcal{L}_{clu} + \mathcal{L}_{pel}$ (baseline + SMO)	56.6 $\pm$ 0.2	16.7 $\pm$ 0.2	26.3 $\pm$ 0.3
$+\mathcal{L}_{cls}$	61.2 $\pm$ 0.1	23.7 $\pm$ 0.2	30.5 $\pm$ 0.2
$+\mathcal{L}_{sem}$ (VGSE-SMO)	**62.4** $\pm$ 0.3	**26.1** $\pm$ 0.3	**35.8** $\pm$ 0.2
VGSE-WAvg	57.7 $\pm$ 0.2	25.8 $\pm$ 0.3	35.3 $\pm$ 0.2

Table 3. Ablation study over the PC module reporting ZSL T1 on AWA2, CUB, and SUN (mean accuracy and std over 5 runs). The baseline is the PC module with the cluster loss $\mathcal{L}_{clu}$ and $\mathcal{L}_{pel}$. Our full model VGSE-SMO is trained with two additional losses $\mathcal{L}_{cls}$, $\mathcal{L}_{sem}$. Two kinds of semantic embeddings learned from k-means clustering and pretrained ResNet are listed below for comparison.

CAAP [5] learns the unseen semantic embeddings with the help of w2v and the human annotated attributes for seen classes. Auto-Dis [3] collects attributes from online encyclopedia articles that describe each category, and learn attribute-class association with the supervision of visual data and category label. GAZSL [67] and ZSLNS [39] learn semantic embeddings from wikipedia articles.

The results shown in Table 2 demonstrate that our VGSE embedding, using only w2v as external knowledge, surpasses all other method that uses textual articles on three datasets. In particular, our VGSE-SMO achieves an accuracy of 61.3% on AWA2, improving the closest semantic embedding w2v by 2.9%. On SUN, we also outperform the closest semantic embedding w2v by 1.5%.

4.2. Ablation study

We provide ablation studies for our PC and CR modules.

Is PC module effective? We first ask if learning semantic embeddings through clustering is effective in terms of ZSL accuracy, when compared to other alternatives. We compare our semantic embeddings against the following baselines:

ResNet features are extracted by feeding image patch x_{nt} to a pretrained ResNet50. We follow Eq. 7 and Eq. 8 to predict semantic embeddings for seen classes. *K-means clustering* is an alternative for our clustering model. We cluster the patch images features $\theta(x_{nt})$ learned from our PC module into D_v visual clusters. The patch embedding a_{nt}^k is defined as the cosine similarity between the patch feature $\theta(x_{nt})$ and the cluster center. In both cases the unseen semantic embeddings are predicted with our SMO module.

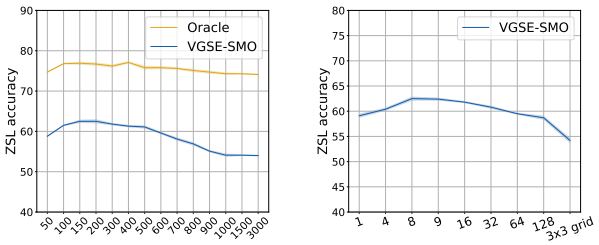
We ablate our losses and compare our VGSE-SMO with the two alternatives, then report ZSL results on three benchmark datasets in Table 3. First, the k-means-SMO achieves on par results with our baseline model trained with only the cluster losses $\mathcal{L}_{clu}$ and $\mathcal{L}_{pel}$ [53], the reason we adopt [53] instead of k-means is that we can easily train the network with our proposed losses in an end-to-end manner. Second, the addition of the classification loss $\mathcal{L}_{cls}$ leads to notable improvement over the baseline model trained with $\mathcal{L}_{clu}$ and $\mathcal{L}_{pel}$, and the semantic relatedness loss $\mathcal{L}_{sem}$ further improve the performance of our semantic embeddings, e.g., in total, we gain

5.8%, 9.4% and 9.5% improvement on AWA2, CUB, and SUN, respectively. The result demonstrates that imposing class discrimination and semantic relatedness leads to better performance in the ZSL setting. Third, our VGSE−SMO embeddings improve over the ResNet−SMO embeddings by 7.1%, 10.7% and 10.7% on AWA2, CUB, and SUN, respectively. We conjuncture that the visual clusters learned in our model is shared among different classes and lead to better generalization ability when the training and testing sets are disjoint (see qualitative results in Figure 1 and Section 4.3).

How many clusters are needed? To measure the influence of the cluster number D_v on our semantic embeddings, we train the PC module with various D_v (results shown in Figure 3a). When the unseen semantic embeddings are predicted under an oracle setting (predicted from the unseen class images), various dimension D_v does not influence the classification accuracy on unseen classes (the orange curve). While under the ZSL setting where unseen semantic embeddings are predicted from class relations (VGSE−SMO), the cluster numbers influence the ZSL performance. Before the cluster number increases up to a breaking point ($D_v = 200$), the ability of the semantic embeddings is also improved (from 58.4% to 62.5%), since the learned clusters contain visually similar patches from different classes, which can model the visual relation between classes. However, increasing the number of clusters leads to small pure clusters (patches coming from one single category), resulting in poor generalization between seen and unseen classes.

SMO vs WAvg. We compare our two class relation functions VGSE−WAvg and VGSE−SMO in Table 3 (Row 7 and 6). The results demonstrate that VGSE−WAvg works on par with VGSE−SMO on SUN and CUB datasets, with $< 0.5\%$ performance gap. While on AWA2 dataset, VGSE−SMO yields better ZSL performance (with 62.4%) than VGSE−WAvg (with 57.7%). The results indicate that predicting the unseen semantic embeddings with the weighted average of a few seen classes semantic embeddings (VGSE−WAvg) is working well for fine-grained datasets since the visual discrepancy between classes is small. However, for coarse-grained dataset AWA2, the class relation function considering all the seen classes embeddings (VGSE−SMO) works better.

Ablation over patches. We further study if using patches for clustering is better than using the whole image, and how many patches do we need from one image. The experiment results in Figure 3b demonstrate that with the patch number increase from 1 (single image clustering) to 9, the ZSL performance increases as well, since the image patches used for semantic embedding learning contain semantic object parts and thus result in better knowledge transfer between seen and unseen classes. However, for a large N_t, the patches might be too tiny to contain consistent semantic, thus resulting in performance dropping, e.g., the ZSL accuracy on AWA2 drops from 62.4% ($N_t = 9$) to 58.7% ($N_t = 128$). We also



(a) cluster number D_v (b) patch number N_t

Figure 3. (a) Influence of the cluster number $D_v = 50, \ldots, 3000$. In the oracle setting, we feed unseen classes images to the PC module to predict unseen semantic embeddings. (b) Influence of the patch number N_t we used per image with the watershed segmentation for obtaining our VGSE−SMO class embeddings. $N_t = 1$ uses the whole image (no patches). "3×3 grid" crops the image into 9 square patches. Both plots report ZSL accuracy with SJE model trained on AWA2 dataset (mean and std over 5 runs).

Semantic Embeddings	AWA2		CUB	
	T1	H	T1	H
w2v [31]	53.7 ± 0.2	48.8 ± 0.1	14.4 ± 0.3	18.0 ± 0.2
VGSE−SMO (w2v)	62.4 ± 0.1	56.8 ± 0.1	26.1 ± 0.2	28.3 ± 0.1
glove [38]	38.8 ± 0.2	38.7 ± 0.3	19.3 ± 0.2	13.4 ± 0.1
VGSE−SMO (glove)	46.5 ± 0.1	46.0 ± 0.1	25.2 ± 0.3	27.1 ± 0.2
fasttext [7]	47.7 ± 0.1	44.6 ± 0.3	-	-
VGSE−SMO (fasttext)	51.9 ± 0.2	53.2 ± 0.1	-	-
Attribute	62.8 ± 0.1	62.6 ± 0.3	56.4 ± 0.2	49.4 ± 0.1
VGSE−SMO (Attribute)	66.7 ± 0.1	64.9 ± 0.1	56.8 ± 0.1	50.9 ± 0.2

Table 4. Evaluating the external knowledge, i.e., word embeddings w2v [31], glove [38], fasttext [7], and the human annotated attributes, for our VGSE−SMO embeddings, e.g., VGSE−SMO (glove) indicates that CR module is trained with glove embedding. **T1**: top-1 accuracy in ZSL, **H**: harmonic mean in GZSL trained with SJE [2] on AWA2, and CUB (std over 5 runs).

compare the patches generated by watershed segmentation proposal with using 3×3 grid patches ($N_t = 9$), and we found that using watershed as the region proposal results in accuracy boost (8.2% on AWA2) compared to the regular grid patch, since the former patches tend to cover more complete object parts rather than random cropped regions.

Can we do better with human annotated attributes? Table 4 shows the performance of our model when different external knowledge is used to perdict the unseen class embeddings in the CR module. Nearly all of our conclusions from former section carry over, e.g., VGSE−SMO class embeddings outperform the other class embeddings by a large margin. For instance, we improve the ZSL accuracy over glove by 7.7% (AWA2) and 5.9% (CUB). Furthermore, VGSE−SMO (Attribute) also outperform Attribute on both AWA2 and CUB dataset, i.e., we achieve 66.7% (ZSL) on AWA2, compared to human attributes with 62.8%. The results demonstrate that our

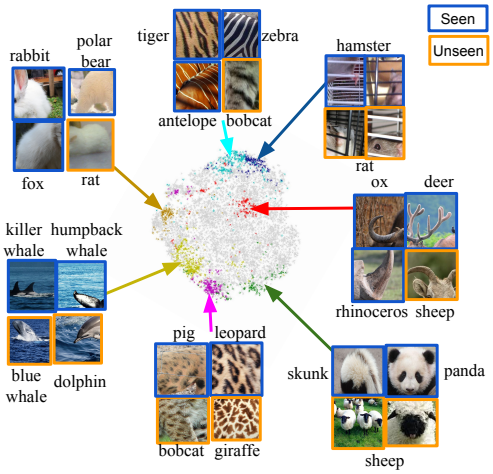


Figure 4. T-SNE embeddings of image patches from AWA2. Each colored dot region represents one visual cluster learnt by our VGSE model. We sample the seen (in blue) and unseen images (in orange) from the cluster center with their class names shown nearby.

VGSE-SMO embeddings coupled with visually-grounded information can not only outperform the unsupervised word embeddings, but also improve over human attributes in transferring knowledge under the zero-shot setting.

4.3. Qualitative Results

In Figure 4, we show the 2D visualization of image patches in the AWA2, where $10,000$ image patches are presented by projecting their embeddings a_{nt} onto two dimensions with t-SNE [52]. To picture their distribution on the embedding space, we sample several visual clusters (dots marked in the same color) and the image patches from the cluster center of both seen and unseen categories. Note that the unseen patches are not used to predict the unseen semantic embeddings, but only used for visualization here.

We observe that samples in the same cluster tend to gather together, indicating that the embeddings provide discriminative information. Besides, images patches in one cluster do convey consistent visual properties, though coming from disjoint categories. For instance, the *white fur* appears on *rabbit*, *polar bear*, and *fox* are clustered into one group, and the *striped fur* from *tiger*, *zebra*, and *bobcat* gather together because of their similar texture. We further observe that nearly all clusters consist images from more than one categories. For instance, the *horns* from seen classes *ox*, *deer*, *rhinoceros*, and unseen class *sheep*, that with slightly different shape but same semantic, are clustered together. Similar phenomenon can be observed on the *spotted fur* and *animals in ocean* clusters. It indicates that the clusters we learned contain semantic properties shared across seen classes, and can be transferred to unseen classes. Another interesting observation is that our VGSE clusters discover visual properties that my be neglected by human-annotated attributes, e.g., the *cage* appear for *hamsters* and *rat*, and the *black and white fur* not only appear on *gaint panda* but also on *sheeps*.

4.4. Human Evaluation

To evaluate if our VGSE conveys consistent visual and semantic properties, we randomly pick 50 clusters, each equipped with 30 images from the cluster center, and ask 5 postgraduate students without prior knowledge of ZSL to examine the clusters and answer the following three questions. Q1: Do images in this cluster contain consistent visual property? Q2: Do images in this cluster convey consistent semantic information? Q3: Please name the semantics you observed from the clusters, if your answer to Q2 is true. We do the same user study to 50 randomly picked clusters from the k-means clustering model. The results reveal that in 88.5% and 87.0% cases, users think our clusters convey consistent visual and semantic information. While for k-means clusters, the results are 71.5% and 71.0%, respectively. The user evaluation results agree with the quantitative results in Table 3, which demonstrates that the class embeddings containing consistent visual and semantic information can significantly benefit the ZSL performance. Interestingly, by viewing VGSE clusters, users can easily discover semantics and even fine-grained attributes not depicted by human-annotated attributes, i.e., the *fangs* and *horns* in figure 1. Note that the whole process, i.e., naming 50 attributes for 40 classes, took less than 1 hour for each user.

5. Conclusion

We develop a Visually-Grounded Semantic Embedding Network (VGSE) to learn distinguishing semantic embeddings for zero-shot learning with minimal human supervision. By clustering image patches with respect to their visual similarity, our network explores various semantic clusters shared between classes. Experiments on three benchmark datasets demonstrate that our semantic embeddings predicted from the class-relation module are generalizable to unseen classes, i.e., achieving significant improvement compared with word embeddings when trained with five models in both ZSL and GZSL settings. We further show that the visually augmented semantic embedding outperforms other semantic embeddings learned with minimal human supervision. The qualitative results verify that we discover visually consistent clusters that generalize from seen to unseen classes and can unearth the fine-grained properties not depicted by humans.

Acknowledgements

This work has been partially funded by the ERC 853489 - DEXIM and by the DFG – EXC number 2064/1 – Project number 390727645.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *T-PAMI*, 2015. 1, 2

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 5, 6, 7

[3] Ziad Al-Halah and Rainer Stiefelhagen. Automatic discovery, association estimation and learning of semantic attributes for a thousand categories. In *CVPR*, 2017. 1, 2, 6

[4] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. Fashion forward: Forecasting visual style in fashion. In *ICCV*, 2017. 1

[5] Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *CVPR*, 2016. 4, 6

[6] Alessandro Bergamo, Lorenzo Torresani, and Andrew W Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In *NIPS*. Citeseer, 2011. 2

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 7

[8] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *ICLR*, 2019. 2

[9] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Generating visual representations for zero-shot classification. In *ICCV Workshops*, 2017. 2

[10] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *CVPR*, 2015. 1, 2

[11] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, 2018. 1

[12] Rudi L Cilibrasi and Paul MB Vitanyi. The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 2007. 2

[13] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3, 5

[15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013. 2

[16] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 2012. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 3

[18] Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*. IEEE, 2012. 2

[19] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*. IEEE, 2009. 1, 2

[20] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NeurIPS*, 2013. 2

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[23] Wei-Lin Hsiao and Kristen Grauman. Learning the latent" look": Unsupervised discovery of a style-coherent embedding from fashion images. In *ICCV*, 2017. 1

[24] Masato Ishii, Takashi Takenouchi, and Masashi Sugiyama. Zero-shot domain adaptation based on attribute information. In *Asian Conference on Machine Learning*. PMLR, 2019. 1

[25] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, and Xilin Chen. Learning discriminative latent attributes for zero-shot classification. In *ICCV*, 2017. 2

[26] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, 2019. 2

[27] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 1

[28] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. Goal-oriented gaze estimation for zero-shot learning. In *CVPR*, 2021. 1, 5, 6

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1

[30] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Field-guide-inspired zero-shot learning. In *CVPR*, 2021. 1

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *NeurIPS*, 2013. 1, 2, 4, 5, 6, 7

[32] Peer Neubert and Peter Protzel. Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In *ICPR*. IEEE, 2014. 3

[33] Ishan Nigam, Pavel Tokmakov, and Deva Ramanan. Towards latent attribute discovery from triplet similarities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

[34] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 3

[35] Devi Parikh and Kristen Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*. IEEE, 2011. 2

[36] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *IJCV*, 2014. 1, 2, 5

[37] Peixi Peng, Yonghong Tian, Tao Xiang, Yaowei Wang, Massimiliano Pontil, and Tiejun Huang. Joint semantic and latent attribute modelling for cross-class transfer learning. *T-PAMI*, 2017. 2

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 1, 2, 4, 7

[39] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton Van Den Hengel. Less is more: zero-shot learning from online textual documents with noise suppression. In *CVPR*, 2016. 1, 2, 6

[40] Mohammad Rastegari, Ali Farhadi, and David Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*. Springer, 2012. 2

[41] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010. 2

[42] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, 2019. 1

[43] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, 2019. 2, 5, 6

[44] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Augmented attribute representations. In *ECCV*. Springer, 2012. 2

[45] Ronan Sicre, Yannis Avrithis, Ewa Kijak, and Frédéric Jurie. Unsupervised part learning for visual recognition. In *CVPR*, 2017. 2

[46] Saurabh Singh, Abhinav Gupta, and Alexei A Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 2

[47] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003. 2

[48] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D Manning, and Andrew Y Ng. Zero-shot learning through cross-modal transfer. *NeurIPS*, 2013. 2

[49] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D Manning, and Andrew Y Ng. Zero-shot learning through cross-modal transfer. *NeurIPS*, 2013. 2

[50] Jie Song, Chengchao Shen, Jie Lei, An-Xiang Zeng, Kairi Ou, Dacheng Tao, and Mingli Song. Selective zero-shot classification with augmented attributes. In *ECCV*, 2018. 2

[51] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*. Springer, 2010. 2

[52] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008. 8

[53] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, 2020. 3, 4, 6

[54] Sirion Vittayakorn, Takayuki Umeda, Kazuhiko Murasaki, Kyoko Sudo, Takayuki Okatani, and Kota Yamaguchi. Automatic attribute discovery with neural activations. In *ECCV*, 2016. 2

[55] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1, 2, 3, 5

[56] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, 2018. 2

[57] Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. Flickr distance: a relationship measure for visual concepts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. 2

[58] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016. 2

[59] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *T-PAMI*, 2019. 1, 2, 3, 5

[60] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 2

[61] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, 2019. 1, 2, 3, 5, 6

[62] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *NeurIPS*, 2020. 1, 2, 5, 6

[63] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. *ACL*, 2020. 2

[64] Xun Yang, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua. Interpretable fashion matching with rich attributes. In *ACM SIGIR*, 2019. 1

[65] Felix X Yu, Liangliang Cao, Rogerio S Feris, John R Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013. 2

[66] Yunlong Yu, Zhong Ji, Yanwei Fu, Jichang Guo, Yanwei Pang, Zhongfei Mark Zhang, et al. Stacked semantics-guided attention model for fine-grained zero-shot learning. In *NeurIPS*, 2018. 2

[67] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018. 1, 2, 6

[68] Yizhe Zhu, Jianwen Xie, Zhiqiang Tang, Xi Peng, and Ahmed Elgammal. Semantic-guided multi-attention localization for zero-shot learning. In *NeurIPS*, 2019. 2