Generalized Zero-shot Learning with Multi-source Semantic Embeddings for Scene Recognition

Xinhang Song[1,2], Haitao Zeng[1,3], Sixian Zhang[1,2], Luis Herranz [4], Shuqiang Jiang[1,2]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]China University of Mining and Technology, Beijing, 100083, China
[4]Computer Vision Center, Barcelona, Spain
xinhang.song,hangtao.zeng,sixian.zhang@vipl.ict.ac.cn;lherranz@cvc.uab.es;sqjiang@ict.ac.cn

ABSTRACT

Recognizing visual categories from semantic descriptions is a promising way to extend the capability of a visual classifier beyond the concepts represented in the training data (i.e. seen categories). This problem is addressed by (generalized) zero-shot learning methods (GZSL), which leverage semantic descriptions that connect them to seen categories (e.g. label embedding, attributes). Conventional GZSL are designed mostly for object recognition. In this paper we focus on zero-shot scene recognition, a more challenging setting with hundreds of categories where their differences can be subtle and often localized in certain objects or regions. Conventional GZSL representations are not rich enough to capture these local discriminative differences. Addressing these limitations, we propose a feature generation framework with two novel components: 1) multiple sources of semantic information (i.e. attributes, word embeddings and descriptions), 2) region descriptions that can enhance scene discrimination. To generate synthetic visual features we propose a two-step generative approach, where local descriptions are sampled and used as conditions to generate visual features. The generated features are then aggregated and used together with real features to train a joint classifier. In order to evaluate the proposed method, we introduce a new dataset for zero-shot scene recognition with multi-semantic annotations. Experimental results on the proposed dataset and SUN Attribute dataset illustrate the effectiveness of the proposed method.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**.

KEYWORDS

scene recognition, zero-shot learning, region description

ACM Reference Format: Xinhang Song, Haitao Zeng, Sixian Zhang, Luis Herranz, Shuqiang Jiang. 2020. Generalized Zero-shot Learning with Multi-source Semantic Embeddings for Scene Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20), October 12-16, 2020, Seattle, WA,USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413568

1 INTRODUCTION

Visual recognition has experienced a remarkable progress thanks to deep learning, achieving comparable or even surpassing humans in certain benchmarks, such as object classification [11] and scene classification [12, 44]. The major factor in this success is the availability of large scale annotated image datasets [33, 44]. However, these conventional image classfiers are also constrained by the training data, which limits the number of categories that can be recognized, and their effectivity also relies on having enough images per category. In contrast, the real world contains a much larger number of categories and its distribution has long tail where just few or no images are available for some of them. In addition collecting and annotating such amount of categories is also impractical.

Zero-shot learning (ZSL) methods can predict categories from semantic descriptions, without requiring explicit visual data. This enables visual recognition to be extended beyond the categories with training images (i.e. from *seen* to *unseen* categories). The semantic descriptions connects seen and unseen categories at a shared semantic level. Such descriptions can be represented in terms of attributes [26, 32], word embeddings [25, 35] or even collected from texts [6, 45]. The most common approach consists of aligning the visual representation of the image with the semantic representation of the category, typically through an intermediate mapping between the visual and semantic spaces [19, 42] or in a common intermediate space. The predicted class is then predicted as the closest neighbor in this shared space. Originally, ZSL focused on making predictions over unseen categories only. A more challenging setting is generalized zero-shot learning (GZSL) where the prediction is over all categories (seen and unseen). This more challenging setting requires addressing the inherent bias towards seen categories in the prediction due to the lack of visual data in unseen categories. A promising approach to address this problem is feature generation [13, 38, 45], where synthetic features of unseen categories are generated from the semantic descriptions and then combined with real data from seen categories to train a joint classifier on all categories.

Scene recognition is a fundamental problem in computer vision. Scenes are generally more complex and abstract than scenes, because they are composed of objects themselves as well as other *stuff* (e.g. sea, sky), not arranged in certain spatial layouts (but

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413568

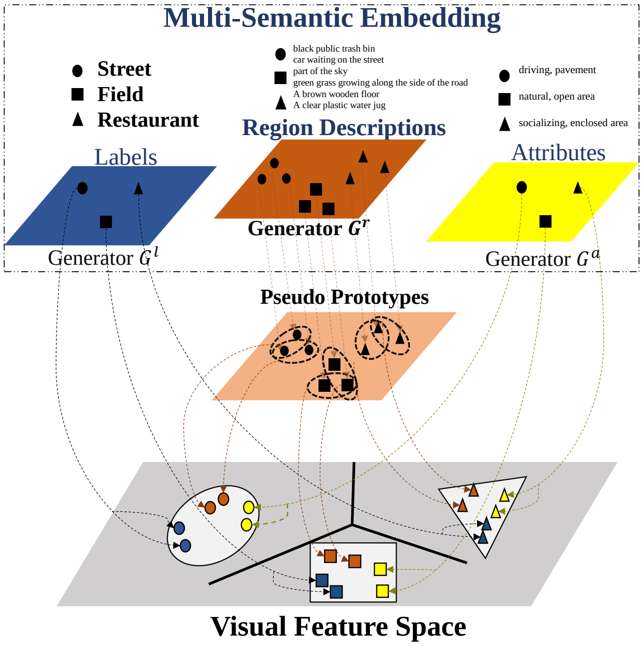

Figure 1: Overview of multi-source semantic embedding aggregation in visual space. 1) semantic representations such as scene attributes, labels and descriptions are first obtained to generated features; 2) different global features are aggregated as training data in visual feature space.

which can vary greatly within the same category). The particular characteristics of scenes have not been explicitly studied in the context of zero-shot learning. In particular, current scene recognition benchmarks contain hundreds of categories where the difference between many categories is subtle and often localized in some particular discriminative object (e.g. hotel room vs bedroom). Thus, attributes, captions and other global semantic descriptions fall short in describing the complexity of scenes and often do not capture the discriminative aspects that allow effective classification of very similar categories. In this paper we focus on this particular case, and propose a framework that captures richer and more discriminative semantic representations for zero-shot recognition by including two novel components: (1) multiple sources of semantic information about categories, and (2) region-level descriptions and a novel feature generation framework to generate synthetic visual features from them.

Our framework is based on feature generation via generative adversarial networks (GANs) [9, 30, 38], where the semantic description is an input condition (together with random noise) to the generator, whose output is a synthetic visual feature. In GANs, the generator improves over time by competing with another network (the discriminator). We consider three types of semantic descriptions: attributes, word embeddings and (category\region) textual descriptions. Depending on the source, the category is embedded in the attribute space, word vector space or in a language space, respectively (we refer to the embeddings as attribute embedding, label embedding and description embedding, respectively). By integrating them we can achieve a richer representation, since different sources may capture only part of the potentially discriminative information. We train specific generators (and discriminators) for each of this semantic sources and then merge them as training data for the joint classifier (see Fig. 1 about the overview of feature merging).

We also leverage local information in the form of region descriptions of the objects in the scene (as in dense captioning). In order to generate global synthetic visual features (also regarded as scene prototypes) from region descriptions, we propose a novel approach consisting of two generation steps. In the first step a number of region descriptions are sampled given the unseen category. Each of them is fed to the generator which outputs a visual feature for that particular. Finally the local visual features are pooled into a global visual feature which is used to train the classifier. While the generator of the second step is implemented as a neural network (i.e. the generator of a GAN), for the first step we implement it as a memory where we save embeddings extracted from images from a set of categories (not from unseen categories), and then generate embeddings by simply randomly selecting them (given the category).

In order to evaluate the proposed approach, we introduce a new benchmark for zero-shot scene recognition (ZSSR) with multiple types of semantic embeddings, such as scene attributes, scene label embedding, scene (category) descriptions and region descriptions. The proposed ZSSR dataset contains 100 seen categories and 100 unseen categories. Experimental results on ZSSR illustrate the effectiveness of the proposed methods. In addition, we also adapt the propose model to the SUN Attribute dataset [29], showing its effectiveness.

2 RELATED WORKS

Scene recognition has achieved impressive performance with the combination of deep networks and scene-centric large datasets (Places [44] in particular). While SUN attributes [29] is a commonly used dataset to evaluate ZSL, there has been little effort in studying the interplay between scene recognition and ZSL, and there are no approaches specific to zero-shot scene recognition.

Early approaches to ZSL model the relations between seen and unseen categories via attributes [1, 19] or word embeddings [7, 28, 35]. Text collected from websites such as Wikipedia have also been used [6, 45]. Lampert et al. [20] address attribute-based ZSL by using an intermediate layer of attribute classifiers from which the category is then inferred. Since the attributes are shared between seen and unseen categories, it is possible to infer unseen classes from the estimated attributes. However, the most common approach to ZSL is visual-semantic alignment. Images are embedded in a visual space (with a CNN) while categories are embedded in a continuous vector space, using word embeddings[7] or as attribute vectors [1]. These two different spaces are aligned using a mapping, typically a linear mapping, but could also be a non-linear one [35]. which are typically optimized using ranking [1, 7], Euclidean [35] or cross-entropy [22] losses. A key choice is the space where the alignment and subsequent nearest neighbor search takes place. Early methods use the lower dimensional semantic space, but it was argued that it can lead to the so-called hubness problem [4, 31]. Motivated by this problem, Zhang *et al.* [42] propose instead mapping semantic embeddings to the visual space, where the hubness problem is less apparent. Instead of explicitly alignment visual and semantic representations, ConSE [28] uses the semantic embedding of seen categories, weighted by the probability of the visual classifier to estimate the semantic embedding of an image of an unseen category. Changpingyo *et al.* [2] proposes synthesizing classifiers of unseen categories as a combination of phantom classifiers learned from seen data. Observing that many attributes are localized in certain regions of the image, some recent approaches explore local representations by using part detectors [6] or integrating attention mechanisms [15, 40, 47].

The more challenging GZSL setting adds an additional problem: combining predictions over seen and unseen classes in an unbiased way. Thus, calibrating seen and unseen scores is critical to achieve good GZSL performance [3]. However this requires either observing unseen images or a validation set with disjoint classes. Cosine similarity has been shown to achieve less biased similarities, improving performance in GZSL [22]. However, most state-of-the-art GZSL methods are based on the feature generation framework [37–39], which addresses the lack of visual data from unseen categories by generating synthetic data, used together with the available real data to directly train a classifier for all seen and unseen categories in a balanced way. The feature generators are typically variational autoencoders (VAEs) [37], GANs [38] or a combination of both [39].

Our approach also builds upon the GAN feature generation framework, but focuses both in the specific challenges of zero-shot scene recognition and proposes a framework to generate synthetic scene features in a bottom-up fashion starting from local semantic embeddings.

3 GZSL WITH MULTI-SEMANTIC EMBEDDING

3.1 Notation and problem setting

We assume we have a dataset $\mathcal{D}^{train} = \left\{\left(v_i^{train}, y_i^{train}\right)\right\}_{i=1}^{N_{train}}$ with N_{train} images where each image feature $v_i^S \in V$ is annotated with the corresponding label $y_i^{train} \in \mathcal{S} = \{1, \ldots K\}$out of K (seen) categories. Image features are extracted the corresponding image x_i^S using a fixed pretrained feature extractor (a CNN). The objective in zero-shot learning is to learn a classifier $f_{ZSL} : V \mapsto \mathcal{U}$ that makes predictions over a test set with $\mathcal{D}^{test} = \left\{\left(v_i^{test}, y_i^{test}\right)\right\}_{i=1}^{N_{test}}$ with M unseen categories, that is $y_i^{test} \in \mathcal{U} = \{K+1, \ldots, K+M\}$. For the generalized zero-shot learning (GZSL) setting the classifier $f_{GZSL} : V \mapsto \mathcal{T}$ makes predictions over the combined set of seen and unseen categories, i.e. $y_i^{test} \in \mathcal{T} = \mathcal{S} \bigcup \mathcal{U} = \{1, \ldots, K+M\}$. In order to perform zero-shot inference, both seen and unseen categories are represented in a common semantic space, obtained through an embedding function. In our case, the input image can be associated with an attribute embedding $a_i = a(y_i)$, a label embedding $l_i = l(y_i)$ or a description embedding $d_i = d(y_i)$, depending on the particular source of semantic information: category-attribute matrices, word embeddings and language models, respectively.

In the next subsections, we first describe the feature generation framework for a single source, and then describe the case with multiple sources.

3.2 Image-level feature generation

We follow previous works using conditional generative adversarial networks (cGANs) as feature generators. A GAN consists of two neural networks, i.e. generator and discriminator, in an adversarial setting. The generator takes a semantic embedding a as input and generates a synthetic visual feature $\tilde{v} = G(z, a)$, where $z \sim \mathcal{N}(0, \mathbf{I})$is a random latent vector that allows diversity in the generated features. The task of the discriminator $D(v, a)$ is to predict whether a visual feature v (than can be real or generated) and the semantic embedding a (we use attributes for simplicity) are sampled from the real training dataset, while the task of the generator is to generate realistic features that can fool the discriminator. The adversarial problem is formulated as optimizing $\min_G \max_D L_{\text{GAN}}$ where:

$$\begin{aligned} L_{\text{GAN}} = {} & \mathbb{E}_{(v,y)\in\mathcal{D}^{train}}\left[\log D(v, a(y))\right] \\ & + \mathbb{E}_{z,y\in S}\left[\log\left(1 - D\left(G(z, a(y)), a(y)\right)\right)\right] \\ & + \alpha L_{\text{GP}} + \beta \mathbb{E}_{z,y\in S} \mathcal{CE}\left(f_{\text{AUX}}\left(G(z, a(y)), y\right)\right) \end{aligned} \quad (1)$$

and trained with the training data $\mathcal{D}^{train}$ and the gradient penalty loss L_{GP} from [10] that provides more stable training. The last term is an auxiliary classifier loss that corresponds to the classification loss ($\mathcal{CE}$denotes cross-entropy) of an auxiliary scene classifier loss that enforces that generated features are classified by the auxiliary classifier f_{AUX} consistently with the scene category used as input condition. The auxiliary classifier is learned jointly with the GAN.

3.3 Training the classifier on all categories

Once the generator is trained with seen data, it can generate synthetic features of unseen categories by conditioning on the corresponding semantic embedding. In this way, a synthetic dataset $\mathcal{D}^{synth} = \{(\hat{v_i}, y_i)\}_{i=1}^{N_{synth}}$ can be obtained for unseen categories, where $y_i \in \mathcal{U}$ and $\hat{v_i} = G(z_i, a(y_i))$ with random $z_i \sim \mathcal{N}(0, \mathbf{I})$. Then the classifier f_{GZSL} on all categories can be trained with the combined dataset $\mathcal{D}^{train} \bigcup \mathcal{D}^{synth}$.

.

3.4 GZSL with Region-based Descriptions

Capturing the diversity in the visual feature distribution is critical to generate synthetic features capturing discriminative aspects that the classifier will in turn learn. This is particularly important in scene recognition since scenes are abstract concepts that are composed of intermediate ones such as objects distributed in diverse category-specific arrangements, and therefore the visual appearance of the scene can be very diverse and heterogeneous ways within the same category. While a GAN can capture some diversity in the distribution via the latent vector z, it is often limited to relatively global aspects of the scene while more subtle local aspects are more difficult to be captured. This is partly related with the fact that the semantic embeddings we have considered so far are global representations. In order to enrich the semantic representation of

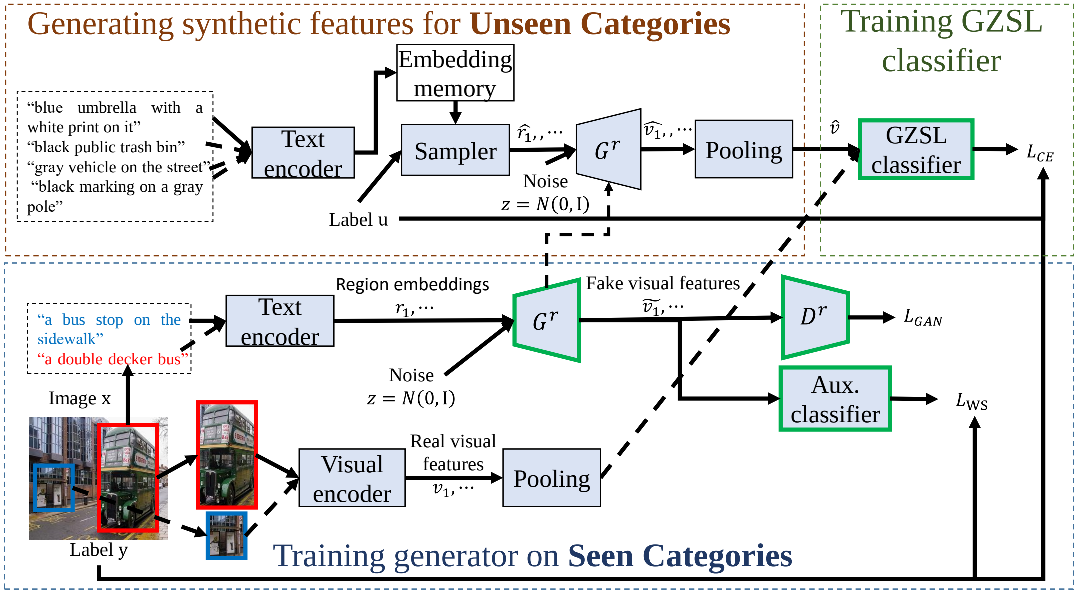

Figure 2: Framework of GZSL with region-based descriptions: 1) training generator with seen categories (bottom line); 2) generating synthetic features for unseen categories; 3) training GZSL joint classifier of seen and unseen categories.

the scene with more diversity and local information, we propose using local intermediate representations of the scene, in the form of captions of detected objects [14, 17]. Intermediate representations such as object banks [23], discriminative parts [16] and local CNN features [5, 8, 24] have been widely used in scene recognition. In our representation we generate instances of features at this local intermediate level, and then aggregate them into a global feature.

We first consider that a particular image x has associated a number of regions (could be annotated or extracted using an object detector). From each region we extract a visual feature v_k and also the corresponding description embedding r_k (obtained with a text encoder). Similarly, we aim at generating synthetic visual features in bottom-up fashion, first generating local description embeddings, then generating their corresponding local visual features, and finally aggregating them in a image-level visual feature. Compared to global semantic embedding, region semantic embedding can increase the diversity in the resulting synthetic scenes.

3.4.1 Region feature generation. The region feature generator is implemented as a GAN, where the region feature generator obtains local synthetic features as $\hat{v} = G^r(z, r)$ given a region description embedding r and a random latent vector z. Similar to (1), the loss to optimize in this case is

$$\begin{aligned} L^r_{\text{GAN}} = & \mathbb{E}_{(v,r)\in\mathcal{D}^{train}}\left[\log D^r(v,r)\right] \\ & + \mathbb{E}_{z,r\in\mathcal{D}^{train}}\left[\log\left(1 - D^r\left(G^r(z,r), r\right)\right)\right] \\ & + \alpha L_{\text{GP}} + \beta\mathbb{E}_{z,(y,r)\in\mathcal{D}^{train}}\mathcal{CE}\left(f_{\text{WS}}\left(G^r(z,r), y\right)\right) \end{aligned} \quad (2)$$

where the auxiliary classifier loss is now a form of weak supervision, since scene labels are used to supervise region predictions (which are at the object level rather than the scene level).

Prior to generating region features, region descriptors embeddings must be generated in the first place given the category. In this case we do not learn a parametric model but instead store region description embeddings in a memory, where we collect a basket region descriptions for each category from a disjoint set, which are then encoded with language model to construct the region embedding memory. The sampler is performed by randomly sampling among the embeddings stored of a particular category.

3.4.2 Global visual feature. After training the region feature generator, it is used to generate synthetic features of unseen categories to train the classifier. In contrast to previous works that generate different pseudo instance features with one category-level semantic embedding, our method generates global visual features with various semantic embeddings sampled from the embedding memory. We first randomly sample P embeddings from the region embedding memory, i.e. $\{r_1, \ldots r_P\}$, which can be regarded as the region-based semantic embeddings of a pseudo instances. Then we generate the corresponding set of synthetic visual features $\{v_1, \ldots \hat{v_P}\}$. Note that region embeddings are sampled within the same category. Finally, the visual features are pooled in a scene prototype $w = \frac{1}{P}\sum_{k=1}^{P}\hat{v_k}$,which is used as global synthetic visual feature for training the classifier.

3.5 Multi-source Aggregation

Due to the abstract and diverse nature of scenes, one type of semantic embeddings is often not sufficient to generate comprehensive visual features to represent scenes. Addressing this problem, we consider four different types of semantic embeddings, three global (i.e. attributes, label and scene descriptions) and the region descriptors described previously. We train separate GANs for each of these types. Since visual features are generated independently they are unlikely to represent the same instance, so merging them at the instance level (e.g. via concatenation) is not possible. Instead, we consider generated features by different generators as separated training instances and we aggregate them in a larger dataset which is used for training the joint classifier.

4 A NEW DATASET FOR ZERO-SHOT SCENE RECOGNITION

In this section, we introduce our dataset for zero-shot scene recognition, denoted as ZSSR. Our goal is to recognize unseen categories by generating features with multiple types of semantic embedding, including scene attributes, scene categories, scene descriptions, and region descriptions. However, there is no current dataset with such characteristics. We collect our datatset based on Visual Genome[18], which provides region-level description annotations for images, Since scene-level annotations require much less labor cost. Particularly, we organize the annotations of scene-level semantics, such as labels, attributes and descriptions. We invite more than 20 workers to take part in annotating all the 108077 images of Visual Genome, which costs more than 600 man-hours.

- Scene labels: annotated by the volunteers, labels are annotated within the vocabulary of 365 types of scenes in Places365.
- Scene attributes: with the obtained scene labels, attribute representation of each scene category is obtained from SUN Attribute [29], note that attributes are collected for each scene category instead of image.
- Scene descriptions: they are collected from Cambridge dictionary and Wikipedia entries of the scene labels.
- Region descriptions: for seen categories, region descriptions are obtained for each image; for unseen categories, a set of descriptions are collected without overlap to the test set of unseen categories.

We collect more than 44K images of 100 seen categories and more than 5K image of 100 unseen categories. Each seen image consists of all the above annotations, and the data distribution of seen categories is illustrated in 3. It can be observed that training data is in long-tailed distribution, where 60 images are randomly selected as training data for each category, and 15 images are selected as test. The rest images and their regions descriptions are used to train language models for region descriptions embedding, which is not sensitive to category distribution.

5 EXPERIMENTS

5.1 Setting

5.1.1 Feature extraction. For images,2048-dimensional visual features are extracted with a ResNet-101 model pretrained on Places365 [43]. For regions, all the training images of seen categories are used to train dense captioning models with the architecture described in [41]. During the training of the dense captioning model, all the images in seen categories (more than 44K images) are used as training and test. For the seen categories, region features are extracted from the annotated regions. For the unseen categories, region features are extracted inside the regions generated by a region proposal network (RPN) during test, and region-level features (combined with global features) are generated during training of the classifiers. Following the setting of [41], image-level features are also concatenated as contextual information.

5.1.2 Semantic embeddings. We consider the following types of semantic embeddings:

- Scene attribute embedding (SAE): obtained from the SUN Attribute, each category is represented with the distribution of 102 attributes.
- Scene label embedding (SLE): each category label is embedded with word2vec [27] in 300 dimension vectors.
- Scene description embedding (SDE): a text encoder combining word2vec and LSTM are is used to embed the scene descriptions, resulting in 300-dimensional vectors.
- Region description embedding (RDE): uses the same text encoder, resulting also in 300-dimensional vectors.

Particularly for RDE, scene labels are also used to supervise the training of LSTM models, inspired by joint model in [36].

- Multi-source Aggregated (MSA): features generated with different semantic embeddings are aggregated into a larger set.

5.1.3 Evaluation protocol. In conventional ZSL setting, the goal is to recognize unseen categories, i.e, $\mathcal{U}$. In GZSL setting, both seen and unseen categories can be involved in the evaluations, i.e., $\mathcal{T} = \mathcal{S} \bigcup \mathcal{U}$. Following the widely used evaluation protocol in [38], we compute the average per-class accuracy on seen (denoted as S) and unseen classes (denoted as U), and then compute their harmonic mean, i.e. $H = 2 \times \frac{acc_{\mathcal{S}} \times acc_{\mathcal{U}}}{acc_{\mathcal{S}} + acc_{\mathcal{U}}}$.

5.2 GZSL with Global Semantic Embedding

5.2.1 Number of generated synthetic features. We first evaluate the proposed method with global embeddings. One key factor of GZSL is the number of generated features. In this evaluation, we use 60 real images per category to train the GAN model, which is used to generate a number of synthetic features (denoted as #Syn). Table 1 compares the performance of the proposed method for different types of semantic embeddings and different amount of synthetic features. It can be observed that, with the increase of synthetic features, the accuracy on unseen categories improves significantly, meanwhile the accuracy of seen categories drops. The best performances are obtained with #Syn=1000, which is almost 17x times of the training images in each seen categories.

5.2.2 Multi-source aggregation. When comparing with different types of semantic embeddings, Table 1 shows that attribute embedding works better than label embedding, and better than scene-level description embedding. When merging all the generated features to train the classifier, the accuracy improves by a large margin (see Fig. 4). For fairness, all the comparisons generate the same number of synthetic features. For a single type of semantic embedding, we generate 300 synthetic features (this setting only used in this evaluation). For multi-source aggregation, each semantic embedding generates 100 synthetic features, with 300 in total. With the same number of synthetic features in classifier training, the proposed multi-semantic aggregation achieves a significant gain (over 11%) in unseen category, and a gain about 8% in H-mean.

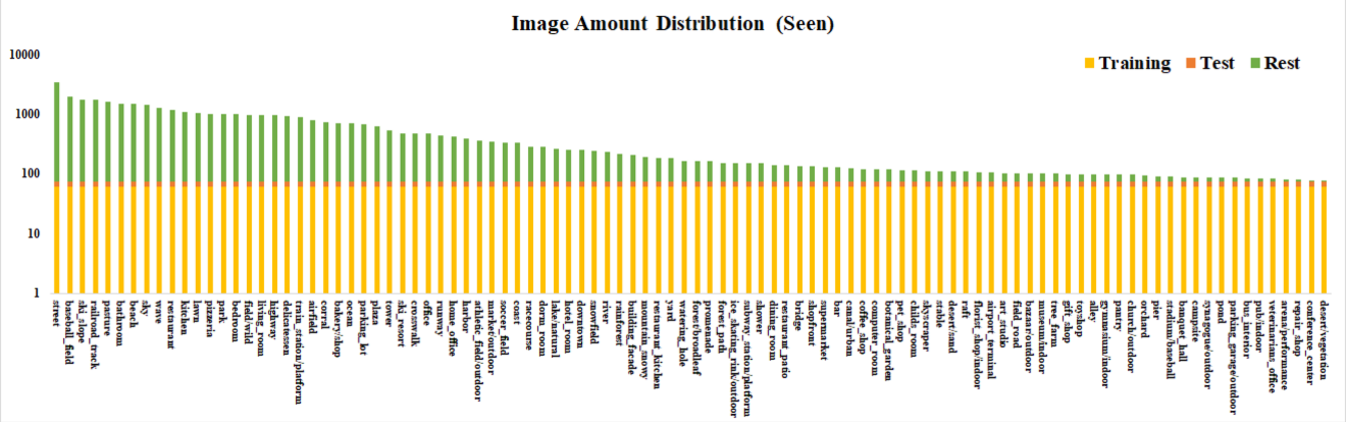

Figure 3: Data distribution of seen categories. The annotated images are in long-tailed category distribution.

Table 1: Accuracy (%) with different amount of generated features of unseen categories on ZSSR

#Unseen	SAE			SLE			SDE		
	U	S	H	U	S	H	U	S	H
100	4.53	**61.6** 0	8.45	3.80	**62.0**	7.16	2.07	**60.73**	4.00
300	14.4 0	48.07	22.16	12.27	51.13	19.79	6.00	51.67	10.75
500	19.01	38.44	25.44	15.64	40.39	22.55	8.98	40.40	**14.70**
1000	**22.58**	31.32	**26.24**	**19.03**	31.66	**23.77**	10.90	22.05	14.59
2000	17.64	24.93	20.66	15.79	27.60	20.08	9.80	21.42	13.45

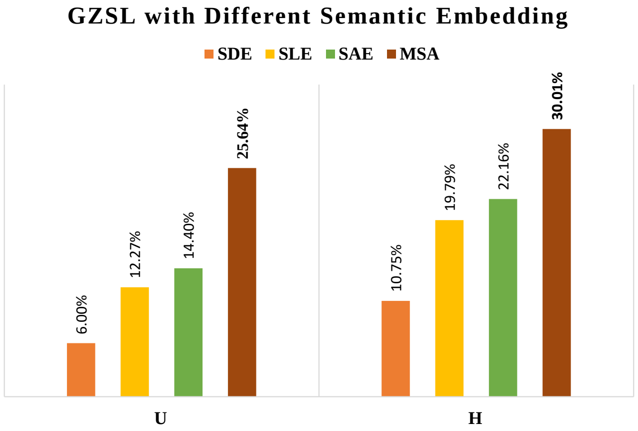

Figure 4: Comparison between different semantic embeddings and their aggregation on ZSSR.

Without considering the same number of synthetic features, the direct gain of multi-source aggregation (over independent semantic embeddings) is 20% in unseen and H-mean, which can be observed between the results in Fig. 4 and Table 1 (first row), since the aggregated model is merged with #syn=100 (corresponded to the first row in Table 1) per category. Another possible comparison is the gain of multi-semantic over the best single semantic embedding (i.e., attribute with #syn=1000 in Table 1), the gains of unseen and H-mean are still over 3%.

5.2.3 *Number of Seen features .*

We also analyze the effects of the amount of seen features (see Table 2). ZSSR has at least 60 training images for each seen category, which are all used in the training of the GAN model. For this evaluation we use the best single semantic embedding, i.e., attribute. We also consider to use different amounts of seen features to train the classifier, which are combined with a fixed number of synthetic features (1000), which obtained the best results in Table 1).

When the number of training features of each seen category is less than 60, we randomly select a subset. Although the best results are still obtained with 60 samples, we found that 30 achieve close performance (drops 2.4%).

Another interesting application of feature generation is to rebalance the imbalanced datasets. This is clearly the case in ZSSR (see Fig. 3) when the number of training features is higher than 60 per category. We evaluate the performance with larger amounts of training features, ranging from 80 to 600 (see Table 2, column #MaxSeen). We consider two cases: imbalanced, where some categories have more training features than others, and re-balanced, where categories with insufficient number of training features are complemented with additional synthetic features. The best H-mean is achieved for 100 training features, with re-balancing achieving slightly better results. We can observe that re-balancing is very effective in improving the accuracy on seen categories, while penalizing the accuracy on unseen categories, and the overall H-mean drops as the number of maximum is larger. The larger #MaxSeen

Table 2: Seen feature analysis on ZSSR.

#Seen	Accuracy (%)					
	U		S		H	
10	12.33		30.40		17.55	
20	14.13		44.40		21.44	
30	22.40		25.40		23.81	
40	20.27		35.87		25.90	
50	20.53		34.93		25.86	
60	22.58		31.32		**26.24**	
Synthetic Features for Seen						
#MaxSeen	U		S		H	
	Reb	Imb	Reb	Imb	Reb	Imb
80	19.53	20.53	37.73	33.73	25.74	25.53
100	20.13	20.80	39.40	36.40	**26.65**	26.47
200	17.93	20.27	41.20	29.53	24.99	24.04
400	15.20	19.67	44.33	27.47	22.64	22.92
600	11.80	16.53	46.73	51.40	18.84	20.83
800	10.40	16.07	46.33	28.13	16.99	20.45
1000	9.40	15.13	46.53	28.73	15.64	19.83

Imb: imbalanced, Reb: re-balanced with synthetic features

requires more synthetic images to re-balance the training data distribution, and more synthetic images also may result in bias for real features.

5.3 GZSL with Region-based Semantic Embedding

We analyze the performance of region-based semantic embeddings for two key factors: the number of visual features of seen categories (i.e., #seen) and the number of synthetic of unseen categories (i.e., #unseen). We first evaluated different combinations of these two values, setting the number of parameters, with. The number of region descriptions sampled for every synthetic feature is empirically set to $P = 5$. As in previous dense captioning works [36, 41], image-level CNN features are concatenated to the region-level CNN[36] we also integrate scenes labels into region-based descriptions to train dense captioning models.

Table 1, when training with a small number of synthetic features generating synthetic features from local embeddings is significantly better than using category-embedding (for instance, see Table 3, when #syn=100, the best H is 26.58%, with a gain over 18%). Even with #syn=40 synthetic features generated from local embeddings, it can achieve H=25.8%, which is better than the results using any of the global semantic embeddings when less than 500 synthetic features (see Table 1).

The comparison to global semantic embeddings suggests that the proposed local embedding works typically better when generating a small number of synthetic features. The effect of the proposed region-level GZSL mainly benefits from the diversity of region-level descriptions. Conventional methods use a single global embedding and one random noise as the condition to generate unseen features, where the diversity of generated features is very limited and uncertain. In contrast, generating region-level features from local embeddings leads to a better representation of the diversity found in real scenes, which may be the main reason to the gain in performance.

5.4 Comparison on ZSSR

We also compare with other related works on the proposed ZSSR dataset, particularly, LisGAN [21], CAVA [46] and ABP [34], which we re-implemented ourselves. The results are shown in Table 4. Compared with baseline methods that generate features with global embedding, our method RDE obtains the best H-mean (26.60%) when using a single semantic embedding.

The aggregation of synthetic features generated from global semantic embeddings (MSA (SAE+SLE+SDE) in Table 4), obtains and accuracy of 25.64% on unseen categories and and an H-mean of 30.01% , which outperforms the best results of single semantic embedding by a large margin of over 3%. This important gain suggests that synthetic features generating from different sources of semantic information are indeed complementary.

We also replace SDE by RDE for aggregation, in order to include both global and local semantic information (i.e., MSA (SAE+SLE+RDE) in Table 4). This variant achieves the highest H-mean(i.e., 30.92%), which is over 4% better than the best single global semantic embedding (i.e., SAE). Compared to MSA (SSS), the gain in H-mean is around 1%. The gain mainly benefits from RDE (over SDE), suggesting that there may be certain discriminative features that are not captured in global descriptions, and therefore combining both global and local semantic descriptions can generate more complete and discriminative visual representation of unseen scenes.

MSA (SAE+SLE+RDE) also outperforms the related works by over 3% in both unseen accuracy (U) and H-mean, which is mainly due to the aggregation of features generated with both global and local semantic embeddings. Note that in CAVA and ABP the feature generators are implemented as a variational autoencoder (VAE). LisGAN also generates features with GAN, but the result H-mean is 5.48% lower than our MSA (SAE+SLE+RDE) variant.

5.5 Comparison on SUN Attribute

In addition to the proposed dataset ZSSR, we evaluate the proposed method on the SUN Attribute dataset [29], which consists of 14340 images annotated with 102 scene attributes, and categorized with 717 scenes. The 200 categories in ZSSR are also categories in SUN Attribute. For these experiments, we adapt the model pretrained on ZSSR to SUN Attribute. In particular, we propose a new seen\unseen split for SUN Attribute that follows the split in ZSSR, i.e. in this evaluation, unseen categories are as same as ZSSR, while the remaining 617 categories of SUN Attribute are regarded as the seen categories. Compared to SUN Attribute, which consists of 20 images in each category (16 of them are used for training in seen categories), ZSSR contains 60 training images in each category. Test data is the same as in SUN Attribute (each unseen category contains 20 images, and each seen category contains 4 images), while we have a different setting for training data.

SDE can be regarded as a particular case of RDE (i.e., considering the whole image as a single large region). However, the results of RDE are much better than those of SDE, we mainly report the results of RDE in this comparison, and MSA (all) means the aggregation of SAE+SLE+RDE. Comparison results on SUN Attribute

Table 3: Region feature analysis on ZSSR

#Seen	Unseen #syn																	
	S						U						H					
	20	40	60	80	100	200	20	40	60	80	100	200	20	40	60	80	100	200
10	47.87	31.13	28.67	26.87	26.40	26.67	12.13	22.07	**23.40**	22.20	20.13	16.00	19.36	25.83	25.77	24.31	22.84	20.00
20	59.07	37.80	36.33	36.40	34.00	28.07	4.60	19.47	20.07	20.93	20.73	20.93	8.54	25.70	25.85	**26.58**	25.76	23.98
30	61.46	44.00	38.47	39.07	39.27	32.13	3.67	16.73	19.73	19.87	20.06	21.93	6.92	24.25	26.09	26.34	26.56	26.07
40	61.20	52.00	38.27	37.00	38.27	36.27	2.60	14.40	19.00	18.87	19.60	20.80	4.99	22.55	25.39	24.99	25.92	26.44
50	64.13	51.73	39.80	39.27	42.80	37.47	2.07	13.73	18.13	19.27	18.27	20.53	4.00	21.70	24.92	25.85	25.61	26.53
60	63.93	52.67	43.87	39.33	38.33	39.73	1.47	11.93	17.00	18.73	19.33	19.60	2.87	19.46	24.50	25.38	25.70	26.25

Table 4: Comparison on ZSSR in accuracy (%)

	Semantic	U	S	H
Baseline	SAE	22.58	31.32	26.24
	SLE	19.03	31.66	23.77
	SDE	8.98	40.40	14.70
Ours	**RDE**	20.90	36.40	**26.60**
	MSA (SSS)	**25.64**	36.19	30.01
	MSA (SSR)	25.20	40.00	**30.92**
Related works	LisGAN [21]	18.67	39.87	25.44
	CAVA [46]	18.07	45.93	25.93
	ABP [34]	23.33	34.27	27.76

MSA(SSS): MSA(SAE+SLE+SDE)
MSA(SSR): MSA(SAE+SLE+RDE)

Table 5: Comparison to SUN Attribute

	Seen	Unseen	Accuracy (%)		
			U	S	H
SAE	SUN	SUN	46.70	39.75	42.94
	ZSSR	SUN	60.00	35.25	44.10
	SUN+ZSSR	SUN	61.25	50.75	55.51
SLE	SUN	SUN	27.50	39.75	32.51
	ZSSR	SUN	38.00	32.75	35.18
	SUN+ZSSR	SUN	36.25	49.50	41.85
RDE (ours)	ZSSR	SUN	33.65	40.50	36.76
	SUN+ZSSR	SUN	36.55	47.00	41.12
MSA (SSR) (ours)	**ZSSR**	**SUN**	61.75	45.75	**52.56**
	SUN+ZSSR	**SUN**	60.75	59.50	**60.12**
Related Works					
LisGAN [21]	SUN	SUN	49.20	44.33	46.64
CAVA [46]	SUN	SUN	49.65	49.47	49.56
ABP [34]	SUN	SUN	49.45	43.11	46.06

MSA(SAE+SLE+RDE)

are illustrated in Table 4. Compared to SUN data, training with our ZSSR obtains significant gains on SAE, and SLE, since ZSSR contains more data in training, suggesting that larger number of training is also more effective even for GZSL model. Since SUN Attribute does not contain region descriptions, RDE is not reported with training data of SUN. In addition, we also try to merge of SUN Attribute and ZSSR (denoted as SUN+ZSSR) for training, where the common 100 seen categories are merged with both dataset, thus, these 100 categories contains 60+16 images for training. SUN+ZSSR obtains the best result on H-mean, with a gain more than 7% over ZSSR, and outperforms the best related work CAVA [46] with a gain over 10%, outperforms the most related work LisGAN (also use GAN) over 13%.

6 CONCLUSION

Previous GZSL methods mainly generate synthetic features from global semantic representations obtained by embedding seen or unseen categories. Generative models such as VAE and GAN can capture the visual diversity in such categories by learning the class-conditional distribution (encouraged and controlled during via a random latent vector). In our experiments we show that global descriptions fall short in describing the subtle and discriminative characteristics of the different scene categories, which are largely localized in regions, limiting the ability of the generator to capture and subsequently reproduce their diversity with scene categories. We demonstrate that combining multiple sources of semantic descriptions can significantly alleviate this problem, since different semantic embeddings can be complementary. However, global descriptions still do not cover many important local object-related information. This limitations motivates us to propose a two-step generative framework to generate synthetic scene prototypes from a number of region description embeddings in turn sampled from a embedding memory given the category. Our experiments demonstrate the potential of region-based semantic descriptions for zero-shot scene recognition. By further aggregating the proposed scene prototypes with the other synthetic features generated by global embeddings, we obtain significant improvements on both the proposed ZSSR and SUN Attribute datasets. To the best of our knowledge, this is the first GZSL work to generate features with both region-based and global semantic embeddings.

ACKNOWLEDGMENT

This work was supported by National Key Research and Development Project of New Generation Artificial Intelligence of China, under Grant 2018AAA0102500, in part by the National Natural Science Foundation of China under Grant 61532018, 61902378 and U1936203, in part by Beijing Natural Science Foundation under Grant L182054 and Z190020, in part by the Lenovo Outstanding Young Scientists Program, in part by the National Postdoctoral Program for Innovative Talents under Grant BX201700255, in part by the EU Marie Sklodowska-Curie grant No.665919 and the Spanish project RTI2018-102285-A-I00.

REFERENCES

[1] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. 2016. Label-Embedding for Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 7 (2016), 1425–1438. DOI: https://doi.org/10.1109/TPAMI.2015.2487986

[2] S. Changpinyo, W. Chao, B. Gong, and F. Sha. 2016. Synthesized Classifiers for Zero-Shot Learning. In *CVPR*. 5327–5336.

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

[4] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568* (2014).

[5] Mandar Dixit, Si Chen, Dashan Gao, Nikhil Rasiwasia, and Nuno Vasconcelos. 2015. Scene Classification with Semantic Fisher Vectors. In *CVPR*.

[6] Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed M. Elgammal. 2017. Link the Head to the "Beak": Zero Shot Learning from Noisy Text Description at Part Precision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 6288–6297. DOI: https://doi.org/10.1109/CVPR.2017.666

[7] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 2121–2129. http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model

[8] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. 2014. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*. 5767–5777.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[12] L. Herranz, S. Jiang, and X. Li. 2016. Scene Recognition with CNNs: Objects, Scales and Dataset Bias. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 571–579. DOI: https://doi.org/10.1109/CVPR.2016.68

[13] He Huang, Changhu Wang, Philip S. Yu, and Chang-Dong Wang. 2019. Generative Dual Adversarial Network for Generalized Zero-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 801–810. DOI: https://doi.org/10.1109/CVPR.2019.00089

[14] A. Karpathy J. Johnson and L. Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Zhong Ji, Yanwei Fu, Jichang Guo, Yanwei Pang, Zhongfei Mark Zhang, and others. 2018. Stacked semantics-guided attention model for fine-grained zero-shot learning. In *Advances in Neural Information Processing Systems*. 5995–6004.

[16] Mayank Juneja, Andrea Vedaldi, C. V. Jawahar, and Andrew Zisserman. 2013. Blocks that Shout: Distinctive Parts for Scene Classification. In *CVPR*.

[17] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. 3128–3137.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (01 May 2017), 32–73. DOI: https://doi.org/10.1007/s11263-016-0981-7

[19] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 3 (2014), 453–465. DOI: https://doi.org/10.1109/TPAMI.2013.140

[20] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Trans. on Image Process.* 36, 3 (2014), 453–465.

[21] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. 2019. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7402–7411.

[22] Kai Li, Martin Renqiang Min, and Yun Fu. 2019. Rethinking Zero-Shot Learning: A Conditional Visual Classification Perspective. In *Proceedings of the IEEE International Conference on Computer Vision*. 3583–3592.

[23] Li-Jia Li, Hao Su, Yongwhan Lim, and Li Fei-Fei. 2014. Object Bank: An Object-Level Image Representation for High-Level Visual Recognition. *Int. J. Comput. Vision* 107, 1 (2014), 20–39. DOI: https://doi.org/10.1007/s11263-013-0660-x

[24] Yunsheng Li, Mandar Dixit, and Nuno Vasconcelos. 2017. Deep Scene Image Classification with the MFAFVNet. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 5757–5765. DOI: https://doi.org/10.1109/ICCV.2017.613

[25] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, and Yueting Zhuang. 2017. Zero-Shot Recognition Using Dual Visual-Semantic Mapping Paths. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 5207–5215. DOI: https://doi.org/10.1109/CVPR.2017.553

[26] Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han. 2017. From Zero-Shot Learning to Conventional Supervised Classification: Unseen Visual Data Synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 6165–6174. DOI: https://doi.org/10.1109/CVPR.2017.653

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). arXiv:1301.3781 http://arxiv.org/abs/1301.3781

[28] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. 2014. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *ICLR*.

[29] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *International Journal of Computer Vision* 108, 1-2 (2014), 59–81.

[30] Guo-Jun Qi, Liheng Zhang, Hao Hu, Marzieh Edraki, Jingdong Wang, and Xian-Sheng Hua. 2018. Global Versus Localized Generative Adversarial Nets. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 1517–1525. DOI: https://doi.org/10.1109/CVPR.2018.00164

[31] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *J. Mach. Learn. Res.* 11 (2010), 2487–2531. http://portal.acm.org/citation.cfm?id=1953015

[32] Bernardino Romera-Paredes and Philip H. S. Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings)*, Francis R. Bach and David M. Blei (Eds.), Vol. 37. JMLR.org, 2152–2161. http://proceedings.mlr.press/v37/romera-paredes15.html

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, AlexanderC. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* (2015), 1–42. DOI: https://doi.org/10.1007/s11263-015-0816-y

[34] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8247–8255.

[35] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero-Shot Learning Through Cross-Modal Transfer. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 935–943. http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer

[36] Xinhang Song, Bohan Wang, Gongwei Chen, and Shuqiang Jiang. 2019. MUCH: Mutual Coupling Enhancement of Scene Recognition and Dense Captioning. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 793–801. DOI: https://doi.org/10.1145/3343031.3350913

[37] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. 2018. Generalized Zero-Shot Learning via Synthesized Examples. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 4281–4289. DOI: https://doi.org/10.1109/CVPR.2018.00450

[38] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. 2018. Feature Generating Networks for Zero-Shot Learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 5542–5551. DOI: https://doi.org/10.1109/CVPR.2018.00581

[39] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. 2019. F-VAEGAN-D2: A Feature Generating Framework for Any-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 10275–10284.

DOI: https://doi.org/10.1109/CVPR.2019.01052

[40] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. 2019. Attentive Region Embedding Network for Zero-Shot Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[41] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li). 2017. Dense Captioning with Joint Inference and Visual Context). In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)).*

[42] Li Zhang, Tao Xiang, and Shaogang Gong. 2017. Learning a Deep Embedding Model for Zero-Shot Learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017.* IEEE Computer Society, 3010–3019. DOI: https://doi.org/10.1109/CVPR.2017.321

[43] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2018), 1452–1464. DOI: https://doi.org/10.1109/TPAMI.2017.2723009

[44] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. In *NIPS*, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger (Eds.). 487–495.

[45] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. 2018. A Generative Adversarial Approach for Zero-Shot Learning From Noisy Texts. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018.* IEEE Computer Society, 1004–1013. DOI: https://doi.org/10.1109/CVPR.2018.00111

[46] Yizhe Zhu, Jianwen Xie, Bingchen Liu, and Ahmed Elgammal. 2019. Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision.* 9844–9854.

[47] Yizhe Zhu, Jianwen Xie, Zhiqiang Tang, Xi Peng, and Ahmed Elgammal. 2019. Semantic-Guided Multi-Attention Localization for Zero-Shot Learning. In *Advances in Neural Information Processing Systems.* 14917–14927.