En-Compactness: Self-Distillation Embedding & Contrastive Generation for Generalized Zero-Shot Learning

Xia Kong[1] Zuodong Gao[1] Xiaofan Li[1] Ming Hong[1]
Jun Liu[2] Chengjie Wang[2] Yuan Xie[3†] Yanyun Qu[1†]
[1]School of Informatics, Xiamen University, Fujian, China [2]Tencent Youtu Lab
[3]School of Computer Science and Technology, East China Normal University, Shanghai, China
kongxia@stu.xmu.edu.cn, {junsenselee, jasoncjwang}@tencent.com
yxie@cs.ecnu.edu.cn, yyqu@xmu.edu.cn

Abstract

Generalized zero-shot learning (GZSL) requires a classifier trained on seen classes that can recognize objects from both seen and unseen classes. Due to the absence of unseen training samples, the classifier tends to bias towards seen classes. To mitigate this problem, feature generation based models are proposed to synthesize visual features for unseen classes. However, these features are generated in the visual feature space which lacks of discriminative ability. Therefore, some methods turn to find a better embedding space for the classifier training. They emphasize the inter-class relationships of seen classes, leading the embedding space overfitted to seen classes and unfriendly to unseen classes. Instead, in this paper, we propose an Intra-Class Compactness Enhancement method (ICCE) for GZSL. Our ICCE promotes intra-class compactness with inter-class separability on both seen and unseen classes in the embedding space and visual feature space. By promoting the intra-class relationships but the inter-class structures, we can distinguish different classes with better generalization. Specifically, we propose a Self-Distillation Embedding (SDE) module and a Semantic-Visual Contrastive Generation (SVCG) module. The former promotes intra-class compactness in the embedding space, while the latter accomplishes it in the visual feature space. The experiments demonstrate that our ICCE outperforms the state-of-the-art methods on four datasets and achieves competitive results on the remaining dataset.

1. Introduction

Image classification tasks relying on large amounts of labeled data [6, 16, 23] have made tremendous progress due to the advancement of deep learning [13, 21, 55]. However, the data hunger nature of deep models leads them to perform unsatisfyingly when some categories have scarce or even no labeled data [47]. Zero-Shot Learning (ZSL) [24, 35] is proposed to tackle this data absence issue by recognizing objects from unseen classes. They first learn a classification model on the seen classes, of which the training samples are provided, then transfer the model to unseen classes using the class-level semantic descriptors [10, 24, 31, 32], such as visual attributes [10, 24] or word vectors [31, 32].

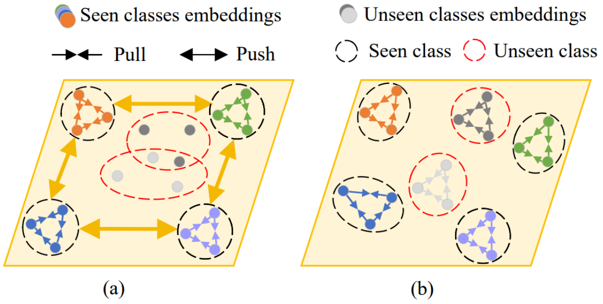


Figure 1. Motivation of this paper. (a) Existing methods such as CE-GZSL [14], produce discriminative embeddings for seen classes, but dispersed embeddings for unseen classes. (b) Our ICCE promotes intra-class compactness with inter-class separability for both seen and unseen classes in the embedding space.

Unlike ZSL, Generalized Zero-Shot Learning (GZSL) [7, 50] has been proposed to identify test samples from both seen and unseen classes, which is more challenging. Since the training set only contains seen classes samples, during testing, GZSL methods tend to misclassify unseen classes samples into seen classes, which is the widespread strong bias problem.

Recently, feature generation based GZSL methods [11, 14, 15, 26, 28, 38] have been proposed to mitigate the strong bias problem by synthesizing training samples for unseen classes conditioned on the semantic descriptors. Merging the real seen training features and the synthetic unseen

[†]Corresponding authors.

features, they obtain a fully-observed dataset to train a GZSL classification model, such as a softmax classifier. Early feature generation methods [11, 26, 28, 38] synthesize features in the visual features space which lacks of discriminative ability [8,14]. Lately, some methods [14,15] search for a new embedding space based on the inter-class relationships for GZSL classifier training. Specifically, RFF-GZSL [15] maps the visual features into a redundancy-free space and uses center loss [48] to strengthen seen classes relationships in that space. CE-GZSL [14] conducts instance-level and class-level contrastive supervision to improve the discrimination of the embedding space. However, in the above methods, the embedding space is strictly constrained by the relationships between seen classes, which is unfriendly to the synthetic unseen classes features. Moreover, the synthetic features of unseen classes have various distributions, as a consequence, mapping them into the embedding space will form confusing distributions. As depicted in Fig. 1 (a), the embeddings of seen classes have large inter-class distances, while the unseen classes embeddings are overlapping and lack of discrimination. Therefore, training the GZSL classifier in this kind of embedding space will end with inferior performance. Instead, as the intra-class relationships are class-independent, if we strengthen these relationships of seen classes, the embedding space can also separate different classes but with better generalization ability on unseen classes. As depicted in Fig. 1 (b), although the inter-class relationships are not highly restricted, a compact intra-class distribution can help all the classes (seen and unseen) distinguish from each other.

In this paper, we propose an Intra-Class Compactness Enhancement method (ICCE) for GZSL. Our ICCE promotes intra-class compactness with inter-class separability on both seen and unseen classes in the embedding space and visual feature space. By putting more emphasis on intra-class relationships but the inter-class structures, we can distinguish different classes with better generalization. Specifically, we produce compact intra-class distributions via a Self-Distillation Embedding (SDE) module and a Semantic-Visual Contrastive Generation (SVCG) module. The SDE module is built with a teacher-student structure, which aligns the representations and the predicted logits between two different samples from the same class. Using SDE, we can reduce the intra-class variations and obtain compact distribution for each class in the embedding space. The SVCG module is a conditional GAN, which synthesizes compact distributed features for unseen classes in the visual feature space with instance-wise semantic-visual contrastive loss. The experiments demonstrate that our ICCE outperforms the state-of-the-arts on four datasets and achieves competitive results on the remaining dataset.

Our contributions are three-fold: (1) we propose an Intra-Class Compactness Enhancement method (ICCE) for GZSL. Our ICCE promotes intra-class compactness with inter-class separability on both seen and unseen classes in the embedding space and visual feature space; (2) we propose a Self-Distillation Embedding (SDE) module to learn an intra-class compact embedding space with representation distillation loss and normalized logits distillation loss; and (3) we propose a Semantic-Visual Contrastive Generation (SVCG) module to synthesize compact intra-class distributed features for unseen classes, with instance-wise semantic-visual contrastive loss.

2. Related Work

Generalized Zero-Shot Learning (GZSL). Zero-Shot Learning (ZSL) aims to train a classifier on seen classes to recognize objects from unseen classes absent in the training set. Provided with the semantic descriptors of both seen and unseen classes, earlier ZSL methods [24, 37, 49, 59] relate them with visual features in an embedding space. They recognize unseen samples by searching their nearest class-level semantic descriptor in this embedding space. Unlike ZSL, which only recognizes unseen classes samples in the test phase, the more challenging GZSL has been proposed to identify test samples from both seen and unseen classes. However, due to the imbalanced nature of ZSL, the early ZSL methods tend to bias towards seen classes under the GZSL scenario. To relieve the bias problem, some methods [3, 7, 29] design new loss functions to balance the predictions between seen and unseen classes, while others [9, 22, 30] solve the GZSL problem by regarding it as an out-of-distribution detection problem. Recently, feature generation based methods have been proposed to synthesize unseen classes features conditioned on the semantic descriptors [26, 28, 39, 51]. After that, they combine the generated unseen samples and the real seen samples to train a softmax classifier. Specifically, RFF-GZSL [15] and CE-GZSL [14] conjecture that the visual feature space lacks of discriminative ability and searches for a new embedding space for GZSL classifier training. However, these methods both construct the embedding space based on the class relevance of seen classes. As a result, the embedding spaces are overfitted to seen classes, leading to inferior generalization ability on unseen classes. Instead, we reinforce the intra-class relationships but the inter-class structures.

Knowledge Distillation. Knowledge distillation [17] aims to train a smaller student network by mimicking a pre-trained complex teacher network. The pioneering work [17] is proposed to optimize the student network by encouraging the student to mimic the teacher's output logits, while follow-up studies utilize other learning objectives, *i.e.*, consistency on feature maps [18] and maximizing the mutual information [42]. Recently, self-knowledge distillation has been proposed in some image classification

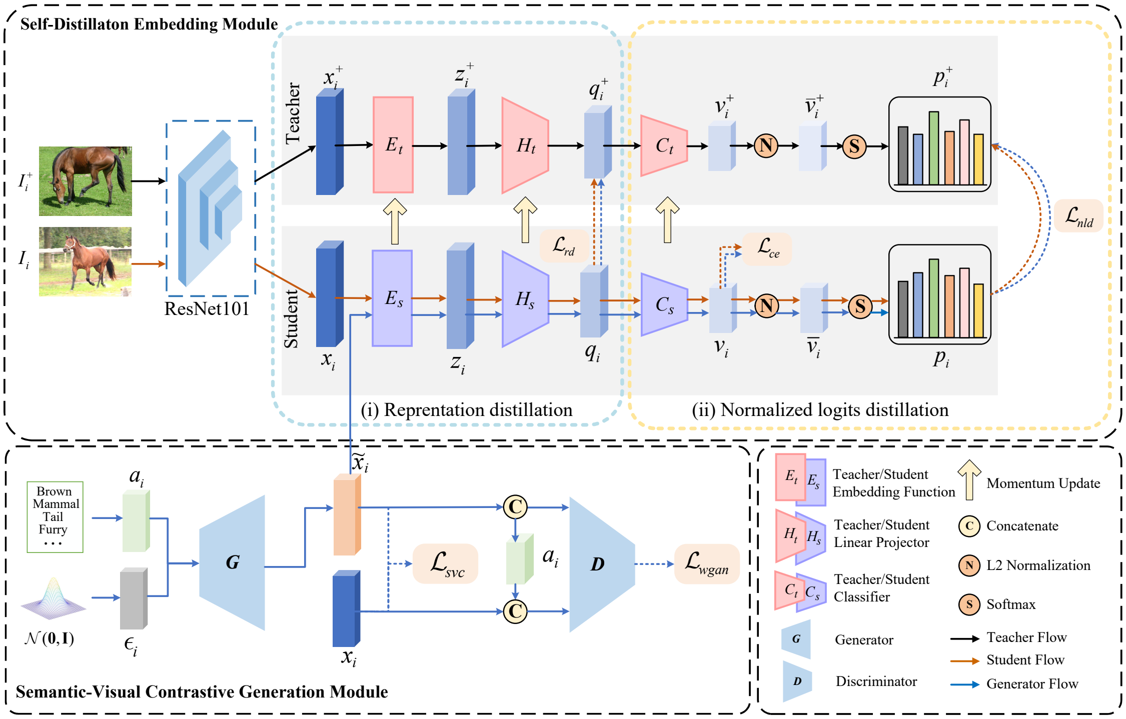


Figure 2. The architecture of the proposed ICCE. It contains a Self-Distillation Embedding (SDE) module and a Semantic-Visual Contrastive Generation (SVCG) module. SDE contains a teacher-student architecture that learns intra-class compact distributions using representation distillation loss ($\mathcal{L}_{rd}$) and normalized logits distillation loss ($\mathcal{L}_{nld}$). SVCG is a conditional GAN that synthesizes compact distributed visual features with a novel instance-wise semantic-visual contrastive loss ($\mathcal{L}_{svc}$).

works [5, 54, 58]. The self-distillation mechanism enhances the effectiveness of training a student network by utilizing its knowledge. For example, [54] transfers knowledge between different distorted versions of the same training data. [5] simplifies self-supervised training by predicting the output of a teacher network, which is built from past iterations of the student network. We follow the teacher updating strategy in [5] to build a self-distillation module and reduce intra-class variations by aligning the feature distribution and probability distribution between two samples from the same class.

3. Self-Distillation Embedding for GZSL

As shown in Fig. 2, our ICCE contains a Self-Distillation Embedding (SDE) module which learns a compact intra-class embedding space, and a Semantic-Visual Contrastive Generation (SVCG) module which synthesizes the compact distributed visual features. In this section, we define the GZSL problem and introduce the proposed SDE and SVCG of ICCE.

3.1. Problem Definition

In ZSL, we have two sets of classes: S seen classes in Y_s and U unseen classes in Y_u, and $Y_s \cap Y_u = \varnothing$. We define a training set $\mathcal{D}_{tr} = \{(x_i, y_i)\}_{i=1}^{N}$, containing N labeled instances, where x_i is a feature vector and y_i is the corresponding label from the seen class Y_s. The test set $\mathcal{D}_{te} = \{x_i\}_{i=N+1}^{N+M}$ has M unlabeled instances. In conventional ZSL, the instances in $\mathcal{D}_{te}$ only come from unseen classes Y_u. In GZSL, test samples are drawn from both the seen and unseen classes. At the same time, the class-level semantic descriptors (attribute) $\mathcal{A} = \{a_i\}_{i=1}^{S+U}$ corresponding to S seen classes and U unseen classes are also provided. Under the two different settings (ZSL and GZSL), the attributes are provided during the whole training process as the bridge from seen classes to unseen classes.

3.2. Self-Distillation Embedding

Our Self-Distillation Embedding (SDE) module is based on the traditional semantic embedding model [2, 12, 37], which uses an embedding function to project the visual feature x_i into an embedding space. But differently, we aim to obtain compact intra-class distribution in this embedding space for both seen and unseen classes. As the intra-class relationships are independent of classes, a more compact distribution within class can also separate different classes but with better generalization ability. Therefore, we force different samples from a same class to be closer in the embedding space. As shown in Fig 2, we build SDE with a teacher-student architecture, which contains a teacher network f_{θ_t} and a student network f_{θ_s}, where θ_t and θ_s are their network parameters. The teacher and student have the same structure which comprises three parts: an embedding function E_t/E_s, a linear projector H_t/H_s, and a classifier C_t/C_s. A straightway is letting θ_t and θ_s also be the same, as traditional self-distillation methods [54, 57] do. However, we experimentally find that it performs poorly on the fine-grained datasets. Instead, we introduce a momentum teacher [41], and its parameters θ_t are updated with an exponential moving average of θ_s as follows:

$$\theta_t \leftarrow \xi\theta_t + (1-\xi)\theta_s, \tag{1}$$

where $\xi \in [0, 1]$ is the decay rate. The parameter of f_{θ_t} is an ensemble of the previous students' weights. Therefore, it can obtain a smooth representation and suppress large variations of the embeddings for better knowledge distillation. With the momentum teacher, we introduce the representation distillation loss and the normalized logits distillation loss to reduce the intra-class variations in the representation level and prediction level.

Representation distillation loss. To minimize the representation variations of two samples from the same class, we force the student to produce the same projections as the teachers. Given a seen class image I_i, we randomly select another image I_i^+ from the same class. Their visual features x_i and x_i^+ are extracted by a fixed ResNet101 [16] pre-trained on the ImageNet [23]. The student network takes x_i as input and produces an embedding $z_i = E_s(x_i)$ and a projection $q_i = H_s(z_i) = H_s(E_s(x_i))$. Similarly, the teacher network takes x_i^+ as input and produces an embedding $z_i^+ = E_t(x_i^+)$ and a projection $q_i^+ = H_t(z_i^+) = H_t(E_t(x_i^+))$. After that, we introduce the representation distillation loss to minimize the difference between $\bar{q_i}$ and $\bar{q_i}^+$, which is formulated as:

$$\mathcal{L}_{rd}(q_i, q_i^+) = 1 - \frac{{q_i}^T q_i^+}{||q_i||_2 \cdot ||q_i^+||_2}, \tag{2}$$

where $|| \cdot ||_2$ is the L_2 norm. By minimizing the representation distillation loss, we strengthen the intra-class compactness on the representation level.

Normalized logits distillation loss. To eliminate the intra-class variations on the prediction level, we reduce the divergences between teacher and student logits distributions (v_i^+ and v_i). Using the classifiers C_t and C_s, we obtain $v_i^+ = C_t(q_i^+)$ and $v_i = C_s(q_i)$, respectively. The traditional knowledge distillation methods [17, 57] usually use a softmax layer to produce the posterior distribution p_i, *e.g.*, given the input v_i, the posterior distribution is:

$$ {p_i}^{(k)} = \frac{\exp({v_i}^{(k)}/\tau)}{\sum_{j=1}^{K} \exp({v_i}^{(j)}/\tau)}, k = 1, 2, ..., K, \tag{3}$$

where K is the class number, k is the class index. ${v_i}^{(k)}$ and ${p_i}^{(k)}$ is the predicted logit value and the probability of the k_{th} class, respectively. $\tau > 0$ is a temperature scaling parameter that controls the sharpness of the output distribution. Previous works spend a tremendous effort to find a proper τ, e.g., DINO [5] sets a small τ to get a sharpen distribution, while CS-KD [57] uses a relatively larger τ to produce a softer distribution. In this paper, we find out that τ also greatly impacts the GZSL classification performance. Consequently, we investigate the effect of τ in Appendix 1.1 and find that τ can be considered compensation for the magnitude of teacher logits. Therefore, we give the following theorem to solve the hyperparameter searching issue:

Theorem 1. *If the magnitude of the teacher and student logits are normalized, the temperature in Eq.3 needs no more consideration (τ always equals to 1), that is:*

$$ {p_i}^{(k)} = \frac{\exp({\bar{v}_i}^{(k)})}{\sum_{j=1}^{K} \exp({\bar{v}_i}^{(j)})}, k = 1, 2, ..., K, \tag{4}$$

where $\bar{v}_i = \frac{v_i}{||v_i||_2}$ denotes the L_2 normalized logits. Detailed proof of theorem 1 can be found in Appendix 1.1. According to theorem 1, we obtain the soft probability distributions of teacher and student as follows:

$$ p_i^{+(k)} = \frac{\exp(\bar{v}_i^{+(k)})}{\sum_{j=1}^{K} \exp(\bar{v}_i^{+(j)})}, k = 1, 2, ..., K, \tag{5}$$

$$ {p_i}^{(k)} = \frac{\exp(\bar{v}_i^{(k)})}{\sum_{j=1}^{K} \exp(\bar{v}_i^{(j)})}, k = 1, 2, ..., K, \tag{6}$$

where $\bar{v}_i^+ = \frac{v_i^+}{||v_i^+||_2}$. We hope that projections from the same class have the same predicted probability, therefore, we introduce the normalized logits distillation loss:

$$\mathcal{L}_{nld}(p_i, p_i^+) = D_{KL}(p_i^+||p_i) = \sum_{k=1}^{K} p_i^{+(k)} \log(\frac{p_i^{+(k)}}{{p_i}^{(k)}}), \tag{7}$$

where $D_{KL}(p_i^+||p_i)$ denotes the KL divergence between p_i^+ and p_i. Through the normalized logits distillation, we pay more attention on the intra-class compactness and disregard the tuning process of τ. We also use the cross-entropy loss to supervise the classifier using the class labels:

$$\mathcal{L}_{ce}(v_i, y_i) = -\log(\frac{\exp(v_i^{(y_i)})}{\sum_{k=1}^{K} \exp({v_i}^{(k)})}). \tag{8}$$

Total loss of SDE. By integrating the losses in the representation level and prediction level, the final optimization objective of our SDE module is formulated as:

$$\begin{aligned} \mathcal{L}_{sd} = \mathbb{E}[\mathcal{L}_{rd}(q_i, q_i^+)] + \beta\mathbb{E}[\mathcal{L}_{nld}(p_i, p_i^+)] \\ +\gamma\mathbb{E}[\mathcal{L}_{ce}(v_i, y_i)], \end{aligned} \tag{9}$$

where β and γ denote the loss weights. We minimize $\mathcal{L}_{sd}$ with respect to student parameters, and the teacher parameters are updated according to Eq. 1.

3.3. Semantic-Visual Contrastive Generation

The goal of SVCG module is to synthesize unseen classes features with compact intra-class distributions. SVCG is a Conditional GAN (CGAN), which uses a generator G to synthesize training features $\widetilde{x} = G(a, \epsilon)$, conditioned on a Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a semantic descriptor a. Meanwhile, a Discriminator D is trained with G to distinguish a real pair (x, a) from $(\widetilde{x}, a)$. The G and D are optimized by the WGAN loss with gradient penalty:

$$\begin{aligned}\mathcal{L}_{wgan} = \mathbb{E}[D(x, a)] - \mathbb{E}[D(\widetilde{x}, a)] - \\ \lambda\mathbb{E}[(||\nabla_{\hat{x}} D(\hat{x}, a)||_2 - 1)^2],\end{aligned} \tag{10}$$

where $\hat{x} = \alpha x + (1 - \alpha)\widetilde{x}$ with $\alpha \sim U(0, 1)$ and λ is the penalty coefficient.

Using the CGAN, we can synthesize diverse and realistic unseen classes features to train a GZSL softmax classifier. However, the CGAN only considers the distribution relationships between synthetic and real pairs, and the pairwise relationships between features and semantic descriptors, *i.e.*, data-to-class relationships. It misses an additional opportunity to consider the relation information between instances, *i.e.*, data-to-data relationships. As a result, the synthetic features of an unseen class may be widely different with a loose intra-class distribution and many outliers, which is not conducive to train an unbiased GZSL classifier. To generate features with compact intra-class distributions without sacrificing their diversity, we try to maintain the data-to-data relationships and introduce an instance-wise Semantic-Visual Contrastive Loss ($\mathcal{L}_{svc}$), as shown in Fig. 3. In each training batch, we have B synthetic features $\{\widetilde{x}_i, y_i\}_{i=1}^{B}$ with their labels $y_i \in Y_s$, and B real features $\{x_j, y_j\}_{j=1}^{B}$ and $y_j \in Y_s$. To synthesize more general features, we sample multiple real features for one class in each training batch. Our objective is to maximize the similarity between synthetic and real features with the same label and minimize that when feature pairs have different labels. To achieve this, for a pair of samples $\widetilde{x}_i$ and x_j, we first calculate their similarity s_{ij} using the cosine distance: $s_{ij} = \frac{\widetilde{x}_i^T x_j}{\|\widetilde{x}_i\|_2 \|x_j\|_2}$. Then we treat our task as a binary classification problem: samples from the same class are classified to 1, otherwise 0. Lastly, we formulate our objective $\mathcal{L}_{svc}$ as:

$$\mathcal{L}_{svc} = \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} (\mathbb{1}_{y_i = y_j} log(\sigma(s_{ij})) + \mathbb{1}_{y_i \neq y_j} log(1 - \sigma(s_{ij}))), \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid function.

Except for the compactness constraint in the visual feature space, we also require the projected embeddings of synthetic features to have small intra-class variations. Note that the proposed self-distillation embedding module can force the embeddings of two samples to be similar. Hence, we borrow the teacher-student network in SDE to achieve this goal, by forcing the generated feature and real feature from the same class to be closer in the embedding space. Specifically, given a fake sample $\widetilde{x}_i$ generated by G, we send it into the student network to produce $\widetilde{q}_i = H_s(E_s(\widetilde{x}_i))$ and $\widetilde{v}_i = C_s(\widetilde{q}_i)$. After the L_2 normalization and softmax operation towards $\widetilde{v}_i$, we obtain the probability distribution $\widetilde{p}_i$. Analogously, a positive real feature x_i^+ from the same class serves as the reference of $\widetilde{x}_i$ and we expect the feature projections and probability distributions of them to be consistent. Hence, we conduct $\mathcal{L}_{rd}(\widetilde{q}_i, q_i^+)$, $\mathcal{L}_{nld}(\widetilde{p}_i, p_i^+)$, and $\mathcal{L}_{ce}(\widetilde{v}_i, y_i)$ to compose the synthetic self-distillation loss as the auxiliary task for training G:

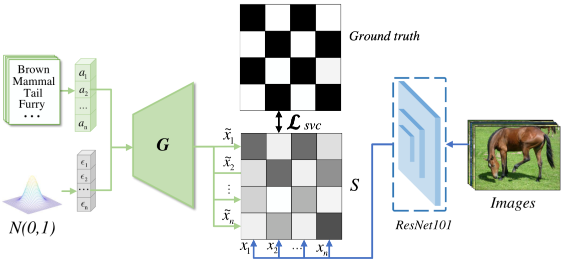


Figure 3. The diagram of our instance-wise Semantic-Visual Contrastive loss ($\mathcal{L}_{svc}$). The synthetic features are forced to be consistent with multiple real visual features having the same class label.

$$\begin{aligned}\mathcal{L}_{sd}^{syn} = \mathbb{E}[\mathcal{L}_{rd}(\widetilde{q}_i, q_i^+)] + \beta\mathbb{E}[\mathcal{L}_{nld}(\widetilde{p}_i, p_i^+)] + \\ \gamma\mathbb{E}[\mathcal{L}_{ce}(\widetilde{v}_i, y_i)].\end{aligned} \tag{12}$$

For stable training of G, we freeze the entire SDE and the loss only feeds back to G.

3.4. Optimization

Our ICCE simultaneously enhances intra-class compactness in the embedding space and visual feature space via a self-distillation embedding module and a semantic-visual contrastive generation module. The overall objective function of ICCE is formulated as:

$$\min_{G, E_s, H_s, C_s} \max_{D} \mathcal{L}_{wgan} + \mathcal{L}_{sd} + \eta\mathcal{L}_{sd}^{syn} + \varphi\mathcal{L}_{svc}, \tag{13}$$

where η and φ are the hyper-parameters indicating the effect of $\mathcal{L}_{sd}^{syn}$ and $\mathcal{L}_{svc}$ towards the generator.

In the end, we use generator G to synthesize features for unseen classes and map them with real features from seen classes to the embedding space using the student embedding function E_s. After that, we train a softmax classifier as our final GZSL classifier.

	Method	Venue	AWA1			AWA2			CUB			FLO			APY		
			U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
Non-generative	DEVISE [12]	NIPS'13	13.4	68.7	22.4	17.1	74.7	27.8	11.5	70.9	19.8	9.9	44.2	16.2	3.5	**78.4**	6.7
	ESZSL [37]	ICML'15	6.6	75.6	12.1	5.9	77.8	11.0	12.6	63.8	21.0	11.4	56.8	19.0	2.4	70.1	4.6
	ALE [1]	TPAMI'16	16.8	76.1	27.5	14.0	81.8	23.9	23.7	62.8	34.4	34.4	13.3	21.9	4.6	73.7	8.7
	COSMO [4]	CVPR'19	52.8	80.0	63.6	-	-	-	44.4	57.8	50.2	59.6	81.4	68.8	-	-	-
	GXE [27]	CVPR'19	62.7	77.0	69.1	56.4	81.4	66.7	47.4	47.6	47.7	-	-	-	-	-	-
	DAZLE [19]	CVPR'20	-	-	-	60.3	75.7	67.1	56.7	59.6	58.1	-	-	-	-	-	-
	RGEN [53]	ECCV'20	-	-	-	**67.1**	76.5	71.5	60.0	**73.5**	66.1	-	-	-	41.8	30.4	37.2
	CN-GZSL [20]	ICLR'21	63.1	73.4	67.8	60.2	77.1	67.6	49.9	50.7	50.3	-	-	-	-	-	-
	HSVA [40]	NIPS'21	59.3	76.6	66.8	56.7	79.8	66.3	52.7	58.3	55.3	-	-	-	-	-	-
Generative	f-CLSGAN [51]	CVPR'18	57.9	61.4	59.6	-	-	-	43.7	57.7	49.7	59.0	73.8	65.6	32.9	61.7	42.9
	cycle-CLSWGAN [11]	ECCV'18	56.9	64.0	60.2	-	-	-	45.7	61.0	52.3	59.2	72.5	65.1	-	-	-
	f-VAEGAN-D2 [52]	CVPR'19	-	-	-	57.6	70.6	63.5	48.4	60.1	53.6	56.8	74.9	64.6	-	-	-
	LisGAN [26]	CVPR'19	52.6	56.3	62.3	-	-	-	46.5	57.9	51.6	57.7	83.8	68.3	33.2	56.9	41.9
	ZSML [45]	AAAI'20	57.4	71.1	63.5	58.9	74.6	65.8	60.0	52.1	55.7	-	-	-	36.3	46.6	40.9
	OCD-CVAE [22]	CVPR'20	-	-	-	59.5	73.4	65.7	44.8	59.9	51.3	-	-	-	-	-	-
	RFF-GZSL [15]	CVPR'20	59.8	75.1	66.5	-	-	-	52.6	56.6	54.6	65.2	78.2	71.1	-	-	-
	TF-VAEGAN [33]	ECCV'20	-	-	-	59.8	75.1	66.6	52.8	64.7	58.1	62.5	84.1	71.1	-	-	-
	GCM-CF [56]	CVPR'21	-	-	-	60.4	75.1	67.0	61.0	59.7	60.3	-	-	-	37.1	56.8	44.9
	CE-GZSL [14]	CVPR'21	65.3	73.4	69.1	63.1	78.6	70.0	63.9	66.8	65.3	**69.0**	78.7	73.5	-	-	-
	FREE [8]	ICCV'21	62.9	69.4	66.0	60.4	75.4	67.1	55.7	59.9	57.7	67.4	84.5	**75.0**	-	-	-
	ICCE	Ours	**67.4**	**81.2**	**73.6**	65.3	**82.3**	**72.8**	**67.3**	65.5	**66.4**	66.1	**86.5**	74.9	**45.2**	46.3	**45.7**

Table 1. Comparisons with state-of-the-art GZSL methods. U and S are the Top-1 accuracy of the unseen and seen classes, respectively. H is the harmonic mean of U and S. The first nine methods are Non-Generative methods, and the following eleven methods are Generative methods. The best and second best results are marked in **bold** and underline, respectively.

4. Experiments

Datasets. We conduct experiments on five widely used ZSL datasets: Animals with Attributes 1&2 (**AWA1** [25] & **AWA2** [50]), USCD Birds-200-2011 (**CUB**) [46], Oxford Flowers (**FLO**) [34], and Attributes Pascal and Yahoo (**APY**). AWA1 and AWA2 share the same 50 animal classes with 85-dimensions attributes. AWA1 includes 30,475 images and AWA2 consists of 37,322 images; CUB contains 11,788 images of 200 bird species; FLO consists of 8189 images from 102 flower classes and APY comprises 12,051 images of 32 diverse classes, e.g., buildings and animals. We use hand-engineering attribute vectors in AWA1, AWA2, and APY, and use the 1024-dimensional attributes generated from textual descriptions [36] in CUB and FLO. Note that AWA1, AWA2, and APY are coarse-grained datasets, while CUB and FLO are fine-grained datasets. We follow the setting of Proposed Split (PS) [50] to split all classes on each dataset into seen and unseen classes.

Evaluation Protocols. During testing, we measure the average per-class Top-1 accuracy [50] of unseen class for the conventional ZSL. Under the GZSL scenario, we evaluate the top-1 accuracy on seen classes (S) and unseen classes (U), as well as their harmonic mean (defined as $H = 2 \times S \times U/(S + U)$).

Implementation Details. In the pre-processing step, we normalize the visual and semantic features into $[0, 1]$ as suggested in [26]. We design the embedding function E_t/E_s as a 2048×2048 Linear layer with LeakyReLU activation. The projector H_t/H_s maps the embeddings into 512 dimension using a 2048×512 fully connected (FC) layer. The classifier C_t/C_s outputs logits on all classes (S and U). The architecture of our generator G and discriminator D are both multilayer perceptrons (MLPs) containing a 4096-unit hidden layer with LeakyReLu activation. We set mini-batch to 512 for AWA1 and AWA2, 64 for CUB, FLO, and APY. The input noise in the generator has the same dimension as the corresponding attributes. All networks are optimized by an Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and an initial learning rate 0.0001. The penalty coefficient λ is set to 10. We empirically set the loss weights η and φ to 0.001.

Method	AWA1	AWA2	CUB	FLO	APY
LATEM [49]	55.1	55.8	49.3	40.4	35.2
DEVISE [12]	54.2	59.7	52.0	45.9	39.8
SJE [2]	65.6	61.9	53.9	53.4	32.9
ALE [1]	59.9	62.5	54.9	48.5	39.7
ESZSL [37]	58.2	58.6	53.9	51.0	38.3
cycle-CLSWGAN [11]	66.3	-	58.4	70.1	-
DLFZRL [43]	71.3	70.3	61.8	-	46.7
GXE [27]	70.9	71.1	54.4	-	38.0
f-CLSWGAN [51]	68.2	-	57.3	67.2	-
f-VAEGAN-D2 [52]	-	71.1	61	67.7	-
TF-VAEGAN [33]	-	72.2	64.9	70.8	-
CE-GZSL [14]	71.0	70.4	77.5	70.6	-
HSVA [40]	70.6	-	62.8	-	-
Our ICCE	**74.2**	**72.7**	**78.4**	**71.6**	**49.5**

Table 2. Results of conventional ZSL. The first five methods are early conventional ZSL methods and following eight are recently proposed GZSL methods. The best and second best results are respectively marked in **bold** and underline.

Case	SDE		SVCG		AWA1			AWA2			CUB			FLO			APY		
	$\mathcal{L}_{rd}$	$\mathcal{L}_{nld}$	$\mathcal{L}_{sd}^{syn}$	$\mathcal{L}_{svc}$	U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
baseline					57.7	81.9	67.7	56.5	81.1	66.6	70.7	58.5	64.0	62.4	80.5	70.3	16.6	**74.8**	27.1
(a)	✓				63.0	79.7	70.4	63.0	76.8	69.3	**71.1**	59.0	64.5	62.9	81.7	71.7	14.9	63.3	24.1
(b)		✓			60.8	80.6	69.3	49.8	84.1	62.6	70.6	56.2	62.6	60.7	81.9	69.7	18.1	68.2	28.6
(c)	✓	✓			64.5	**82.8**	72.5	62.0	**85.3**	71.8	69.3	61.0	64.9	66.8	79.3	72.5	33.5	51.2	40.5
(d)	✓	✓	✓		66.7	80.8	73.1	64.1	82.3	72.1	68.4	63.5	65.9	**67.3**	83.4	74.5	40.9	42.2	41.5
(e)	✓	✓	✓	✓	**67.4**	81.2	**73.6**	**65.3**	82.3	**72.8**	67.3	**65.5**	**66.4**	66.1	**86.5**	**74.9**	**45.2**	46.3	**45.7**

Table 3. Ablation study on the effectiveness of our proposed loss functions on five datasets, the best results are marked in **bold**.

Case	Teacher	AWA1			AWA2			CUB			FLO			APY		
		U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
(1)	Constant random	63.0	80.1	70.5	60.8	79.5	68.9	71.8	57.2	63.7	62.6	82.8	71.3	17.0	62.6	26.8
(2)	Student copy	65.3	79.8	71.8	62.9	**82.9**	71.5	71.3	57.2	63.5	63.3	82.3	71.6	43.2	38.8	40.9
(3)	Previous epoch	66.6	77.2	71.5	63.1	75.7	68.8	68.5	57.2	62.3	62.0	83.5	71.1	12.0	**75.1**	20.7
(4)	Previous iter	65.3	79.0	71.5	62.9	81.4	70.9	**73.0**	56.3	63.6	63.3	84.5	72.4	31.2	29.7	30.5
(5)	Momentum	**67.4**	**81.2**	**73.6**	**65.3**	82.3	**72.8**	67.3	**65.5**	**66.4**	**66.1**	**86.5**	**74.9**	**45.2**	46.3	**45.7**

Table 4. Evaluations of different teachers in SDE module. Our ICCE adopts the momentum teacher, the best results are marked in **bold**.

For AWA1 and AWA2, we set $\beta = \gamma = 0.01$. For CUB, FLO, and APY, we set $\beta = \gamma = 0.001$.

4.1. Comparison with State-of-the-art Methods

In Table 1, we compare our ICCE with the state-of-the-art GZSL methods, including non-generative methods and generative methods. Compared to other generative methods, our ICCE yields further improvements of 4.5%, 1.3%, 0.3%, 0.8% for harmonic mean on AWA1, AWA2, CUB, and APY. Our ICCE achieves the second best H on FLO. Notably, our method achieves the best results on the unseen classes of AWA1, CUB, APY, and the second best results on AWA2 and FLO. Meanwhile, for seen classes, we achieve the best performance on AWA1, AWA2, and FLO. It indicates that our ICCE performs well on seen classes and can also generalize to unseen classes. Moreover, we present the results of our method under the conventional ZSL setting, as reported in Table 2. Our ICCE outperforms the state-of-the-arts by at least 2.9%, 0.5%, 0.9%, 0.8% and 2.8% on AWA1, AWA2, CUB, FLO and APY. These results consistently demonstrate that our ICCE is still effective in conventional ZSL.

4.2. Ablation Study and Discussion

Importance of different components. Here we present ablation experiments to demonstrate the impact of each component in our ICCE. The baseline model is the same as our ICCE but without the teacher branch in the SDE module. We only use the classification loss $\mathcal{L}_{ce}$ and WGAN loss $\mathcal{L}_{wgan}$ to train our baseline model. Totally, we conduct five other experiments using the entire architecture of ICCE but with different loss functions: (**a**) only using $\mathcal{L}_{rd}$ for representation distillation; (**b**) only adopting $\mathcal{L}_{nld}$ for normalized logits distillation; (**c**) applying both $\mathcal{L}_{rd}$ and $\mathcal{L}_{nld}$ for knowledge distillation; (**d**) adding $\mathcal{L}_{sd}^{syn}$ to (**c**) for the generation module training; (**e**) adding $\mathcal{L}_{svc}$ to (**d**) for the semantic-visual contrastive generation module training. According to the results reported in Table 3, we have the following observations:

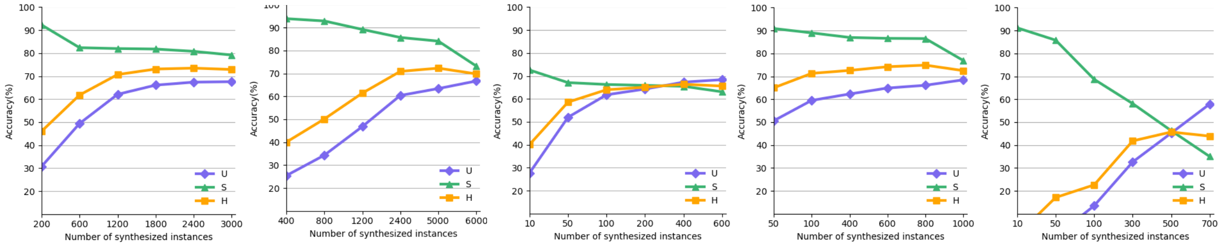
(1) Our SDE module and the proposed knowledge distillation losses can bring obvious performance improvements (comparing baseline with (**a**), (**b**), and (**c**)), and using the combination of $\mathcal{L}_{rd}$ and $\mathcal{L}_{nld}$ can achieve better results. It indicates that reducing the intra-class variations in both representation level and prediction level is more effective.

(2) With $\mathcal{L}_{sd}^{syn}$ and $\mathcal{L}_{svc}$, we can further improve the classification results on all datasets (comparing (**c**) with (**d**) and (**e**)). It shows that using our SVCG module can synthesize better features for unseen classes to train an unbiased GZSL classifier.

(3) Our ICCE benefits from the combination of SDE and SVCG. Therefore, learning compact intra-class distributions in both embedding space and visual feature space is necessary for GZSL.

Effectiveness of momentum teacher. In Table 4, we compare five different strategies to build the teacher from previous instances of the student. (**1**) Constant random: the weights of the teacher network are randomly initialized and fixed during training; (**2**) Student copy: the weights of the teacher are the same as that of the student; (**3**) Previous epoch: use student network from previous epoch as the teacher; (**4**) Previous iteration: use student network from previous iteration as the teacher; (**5**) Our momentum teacher. As illustrated in Table 4, we observe that teacher updating strategies effect the results of GZSL classification, while updating too often or too slowly yields inferior performance. Our momentum teacher ensembles historical students' parameters and is more effective than other versions of teachers.

Influence of number of synthetic instances. In Fig.4, we study the impact of different numbers of synthetic instances per unseen class. The accuracy of unseen classes on five datasets increases along with the number of synthetic



(a) AWA1 (b) AWA2 (c) CUB (d) FLO (e) APY

Figure 4. The GZSL results with respect to different numbers of the synthesized samples for each unseen class.

	AWA1			AWA2			CUB			FLO			APY		
	U-R@1	S-R@1	H-R@1	U-R@1	S-R@1	H-R@1	U-R@1	S-R@1	H-R@1	U-R@1	S-R@1	H-R@1	U-R@1	S-R@1	H-R@1
CE-GZSL	75.1	83.8	79.2	88.7	86.1	87.4	65.3	40	49.6	85.2	81.5	83.3	71.0	87.9	78.6
ICCE(Ours)	**83.9**	**84.2**	**84.0**	**91.4**	**86.3**	**88.8**	**66.4**	**41.3**	**50.9**	**86.4**	**81.7**	**84.0**	**73.5**	**88.3**	**84.0**

Table 5. The Recall at 1 (R@1) rates (%) of CE-GZSL and our method on five datasets. U-R@1 and S-R@1 denote R@1 rates of unseen and seen class, respectively. H-R@1 is the harmonic mean of U-R@1 and S-R@1.

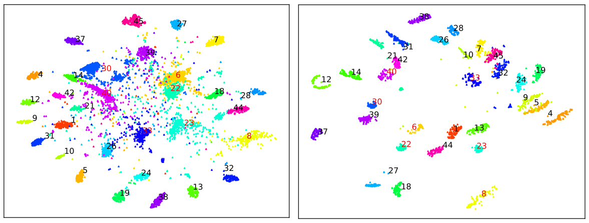
samples, which shows that our ICCE has compensated for the absence of unseen classes features. Our method achieves the best results when synthesizing 24,00, 5000, 400, 800, and 500 samples per unseen classes for AWA1, AWA2, CUB, FLO, and APY, respectively.

Effectiveness of Intra-class Compactness. We first present a visualization analysis of the embeddings in Fig. 5. Specifically, we project the embeddings obtained by CE-GZSL and our ICCE onto two principal components using t-SNE [44]. As illustrated in Fig. 5(a), in CE-GZSL, the embeddings of seen classes are discriminative, while the unseen classes obtain confusion distributions, which indicates that the embedding space cannot effectively generalize to unseen classes. On the contrary, our ICCE promotes intra-class compactness with inter-class separability on both seen and unseen classes in the embedding space.

Moreover, we use the Recall at k (R@k) [60] metric to quantitatively analyze intra-class compactness. Recall at k is the percentage of test samples with at least one from the same class in k nearest neighbors in the embedding space. We utilize the Euclidean distance here and adopt $k = 1$, *i.e.*, R@1. The results of seen, unseen classes and their harmonic mean are denoted as S-R@1, U-R@1, and H-R@1, respectively. Note that the larger value of R@1 implies smaller intra-class variations. As illustrated in Table 5, the R@1 values of our ICCE are all larger than CE-GZSL. It verifies that our ICCE reduces the intra-class variations for both seen and unseen classes. Combining the analysis of the GZSL classification performances, we conclude that enhancing intra-class compactness is better for GZSL.

5. Conclusion

In this paper, we propose an Intra-Class Compactness Enhancement method (ICCE) for GZSL. Our ICCE enhances intra-class compactness with inter-class separability on both seen and unseen classes in the embedding space and visual feature space. Specifically, we have proposed a self-distillation embedding module to reduce the intra-class variations in the representation level and prediction level. Moreover, we have introduced a semantic-visual contrastive generation module to synthesize intra-class compact features for unseen classes. By enhancing the intra-class relationships but the inter-class structures, we can distinguish different classes with better generalization. The experiments on five benchmarks show that our ICCE has outperformed the state-of-the-arts on four datasets and achieved the second best result on the remaining dataset.



(a) CE-GZSL [14] (b) Our ICCE

Figure 5. 2D Projection of the embeddings by using t-SNE. The results are obtained from CE-GZSL [14] and our ICCE on AWA1 testing set. Red numbers denote the unseen classes, black numbers denote the seen classes. Please zoom in for a better view.

Acknowledgement

This work was supported by the National Key Research and Development Programme of China (2020AAA0108301), National Natural Science Foundation of China (61876161, 62176224, 62176092), Shanghai Science and Technology Commission (21511100700), and CAAI-Huawei MindSpore Open Fund.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*. 6

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 3, 6

[3] Yashas Annadani and Soma Biswas. Preserving semantic relations for zero-shot learning. In *CVPR*, 2018. 2

[4] Yuval Atzmon and Gal Chechik. Adaptive confidence smoothing for generalized zero-shot learning. In *CVPR*, 2019. 6

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv:2104.14294*, 2021. 3, 4

[6] Dongliang Chang, Yifeng Ding, Jiyang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi-Zhe Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *TIP*, 2020. 1

[7] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 2016. 1, 2

[8] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. *ArXiv*, 2021. 2, 6

[9] Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *ECCV*, 2020. 2

[10] Ali Farhadi, Ian Endres, Derek Hoiem, and David Alexander Forsyth. Describing objects by their attributes. *CVPR*, 2009. 1

[11] Rafael Felix, Ian Reid, Gustavo Carneiro, et al. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, 2018. 1, 2, 6

[12] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, MarcAurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. 2013. 3, 6

[13] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. Deep learning. *Nature*, 521:436–444, 2015. 1

[14] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *CVPR*, 2021. 1, 2, 6, 8

[15] Zongyan Han, Zhenyong Fu, and Jian Yang. Learning the redundancy-free features for generalized zero-shot object recognition. In *CVPR*, 2020. 1, 2, 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 4

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 2, 4

[18] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv:1707.01219*, 2017. 2

[19] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, 2020. 6

[20] Skorokhodov Ivan and Elhoseiny Mohamed. Class normalization for (continual) generalized zero-shot learning. *ICLR*, 2021. 6

[21] Yakun Ju, Kin-Man Lam, Yang Chen, Lin Qi, and Junyu Dong. Pay attention to devils: A photometric stereo network for better details. In *IJCAI*, 2020. 1

[22] Rohit Keshari, Richa Singh, and Mayank Vatsa. Generalized zero-shot learning via over-complete distribution. In *CVPR*, 2020. 2, 6

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 4

[24] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. *CVPR*, 2009. 1, 2

[25] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 2013. 6

[26] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *CVPR*, 2019. 1, 2, 6

[27] Kai Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *ICCV*, 2019. 6

[28] Bo Liu, Qiulei Dong, and Zhanyi Hu. Zero-shot learning from adversarial feature residual to compact visual feature. In *AAAI*, 2020. 1, 2

[29] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. In *NIPS*, 2018. 2

[30] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *CVPR*, 2019. 2

[31] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013. 1

[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 1

[33] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *ECCV*, 2020. 6

[34] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. 6

[35] Mark Palatucci, Dean A. Pomerleau, Geoffrey E. Hinton, and Tom Michael Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 1

[36] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 6

[37] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. 2, 3, 6

[38] Mert Bulent Sariyildiz and Ramazan Gokberk Cinbis. Gradient matching generative networks for zero-shot learning. In *CVPR*, 2019. 1, 2

[39] 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

[40] Chen Shiming, Xie Guo-Sen, Peng Qinmu, Liu Yang, Sun Baigui, Li Hao, You Xinge, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *NIPS*, 2021. 6

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv:1703.01780*, 2017. 4

[42] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv:1910.10699*, 2019. 2

[43] Bin Tong, Chao Wang, Martin Klinkigt, Yoshiyuki Kobayashi, and Yuuichi Nonaka. Hierarchical disentanglement of discriminative latent features for zero-shot learning. In *CVPR*, 2019. 6

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JLMR*, 2008. 8

[45] Vinay Kumar Verma, Dhanajit Brahma, and Piyush Rai. Meta-learning for generalized zero-shot learning. In *AAAI*, 2020. 6

[46] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[47] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.*, 2019. 1

[48] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 2

[49] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016. 2, 6

[50] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learninga comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018. 1, 6

[51] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 2, 6

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

[53] Guo-Sen Xie, Li Liu, Fan Zhu, Fang Zhao, Zheng Zhang, Yazhou Yao, Jie Qin, and Ling Shao. Region graph embedding network for zero-shot learning. In *ECCV*, 2020. 6

[54] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *AAAI*, 2019. 3, 4

[55] Xu Yang, Cheng Deng, Kun-Juan Wei, Junchi Yan, and Wei Liu. Adversarial learning for robust deep clustering. In *NeurIPS*, 2020. 1

[56] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 2021. 6

[57] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *CVPR*, 2020. 4

[58] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *CVPR*, 2019. 3

[59] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 2

[60] Wengang Zhou, Houqiang Li, and Qi Tian. Recent advance in content-based image retrieval: A literature survey. *arXiv:1706.06064*, 2017. 8