

EXPLOITING SEMANTIC ATTRIBUTES FOR TRANSDUCTIVE ZERO-SHOT LEARNING

Zhengbo Wang^{1,2}, Jian Liang^{2,*}, Zilei Wang¹, and Tieniu Tan^{2,3}

¹ University of Science and Technology of China, ² CRIPAC & MAIS, Institute of Automation, Chinese Academy of Sciences, ³ Nanjing University

ABSTRACT

Zero-shot learning (ZSL) aims to recognize unseen classes by generalizing the relation between visual features and semantic attributes learned from the seen classes. A recent paradigm called transductive zero-shot learning further leverages unlabeled unseen data during training and has obtained impressive results. These methods always synthesize unseen features from attributes through a generative adversarial network to mitigate the bias towards seen classes. However, they neglect the semantic information in the unlabeled unseen data and thus fail to generate high-fidelity attribute-consistent unseen features. To address this issue, we present a novel transductive ZSL method that produces semantic attributes of the unseen data and imposes them on the generative process. In particular, we first train an attribute decoder that learns the mapping from visual features to semantic attributes. Then, from the attribute decoder, we obtain pseudo-attributes of unlabeled data and integrate them into the generative model, which helps capture the detailed differences within unseen classes so as to synthesize more discriminative features. Experiments on five standard benchmarks show that our method yields state-of-the-art results for zero-shot learning.

Index Terms— Zero-shot learning, transductive learning

1. INTRODUCTION

Deep learning has achieved remarkable advances across a wide range of computer vision tasks. However, these improvements always rely on a fragile assumption under which annotated data is abundant and easy to obtain. In fact, millions of categories exist in the real world, and new concepts emerge every day. Therefore, collecting annotated data for all categories is unrealistic.

Zero-shot learning (ZSL) [1] resolves the issue by striving to recognize unseen objects for which no labeled data is available during training. To accomplish this, ZSL approaches leverage extra semantic information to bridge the gap between seen and unseen classes, which could be semantic attributes [2], word embeddings [3], or language descriptions [4] of classes. Hence, they can recognize the unseen classes by generalizing the visual-semantic relation learned

from seen classes. Typically, ZSL methods could be divided into two main categories: inductive ZSL where only labeled seen data is available during training, and transductive ZSL where both labeled seen data and unlabeled unseen data are available. Inductive ZSL methods have been found to yield biased predictions towards seen classes [5, 6, 7], which is unsatisfactory for real-world applications. Recent transductive ZSL methods [8, 9, 10] mitigate the bias by incorporating the unlabeled data into the popular generative modeling framework, generative adversarial network (GAN) [11]. Generally, there exists a generator, a conditional discriminator, and an unconditional discriminator, creating a Y-shape framework. The shared generator and two branched discriminators form a conditional and an unconditional GAN for seen and unseen classes, respectively. Taking seen images and matched attributes as input, the conditional discriminator discriminates whether the images come from the conditional distribution of the corresponding attributes. Contrarily, with unlabeled unseen images, the unconditional discriminator simply distinguishes whether the images belong to unseen classes distribution, serving as regularization to alleviate the bias problem. Since they neglect to model the conditional distribution of unseen classes like the conditional GAN, the generated features of different unseen classes may congregate.

To deal with the problem, we propose a pseudo-attribute-based method, which imposes class-wise constraints on the unseen discriminator by exploiting semantic attributes. First, we train a Y-shape generative model in company with an auxiliary attribute decoder, which learns the mapping from visual features to semantic attributes. With the pretrained attribute decoder, we can produce reliable pseudo-attributes to represent the class information of the unseen data. Taking the unseen data and the corresponding pseudo-attribute as input, the unseen discriminator can discriminate whether the data is real and consistent with the attribute, and therefore modeling the conditional distribution. Experiments on standard benchmarks validate the superior performance of our method over the state-of-the-art methods.

2. PROPOSED METHOD

Assuming that we have two disjoint sets of classes: seen classes \mathcal{Y}_s and unseen classes \mathcal{Y}_u , where $\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$. For

* corresponding author

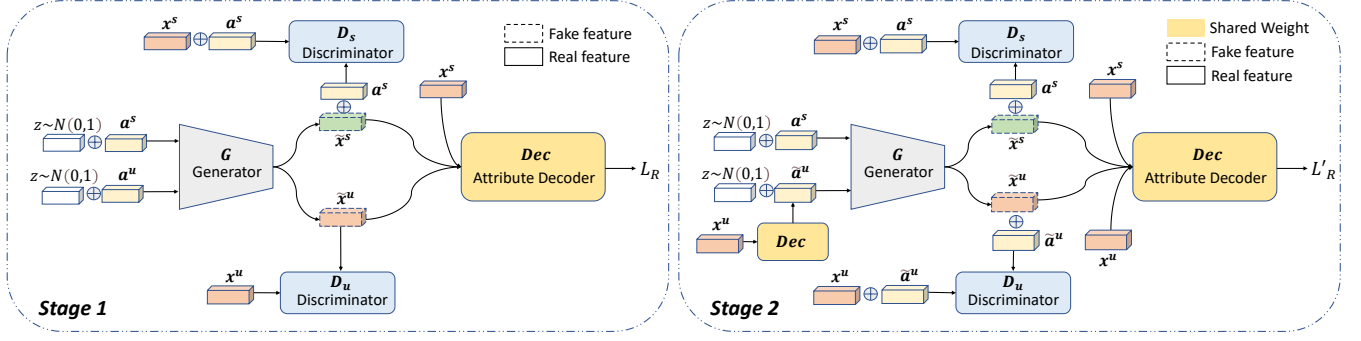


Fig. 1. The overall architecture of our method. It consists of one generator G and two discriminators, D_s and D_u , for seen and unseen classes, respectively. Moreover, an auxiliary attribute decoder Dec is introduced to learn to generate semantically consistent features. (1) In stage 1, we train a generative model with the unconditional unseen discriminator D_u in the absence of matched semantic attributes of unseen data. (2) In stage 2, we employ the pretrained attribute decoder Dec to exploit the class-wise information by generating pseudo-attributes for unseen data, which is then concatenated into the generative process. Finally, we retrain the modified generative model and obtain the generator for the ZSL classification.

the seen classes, we have labeled data: $\mathcal{S} = \{(x_i^s, y_i^s, a_i^s)\}_{i=1}^{N_s}$, where $x_i^s \in \mathcal{X}$ and $a_i^s \in \mathcal{A}_s$ denote its d -dimension visual feature and k -dimension semantic attribute for the i -th image with label $y_i^s \in \mathcal{Y}_s$. And for the unseen classes, an extra unlabeled unseen data is available: $\mathcal{U} = \{x_i^u\}_{i=1}^{N_u}$, where $x_i^u \in \mathcal{X}$ is i -th unseen visual feature. Moreover, we have a set of unseen attributes \mathcal{A}_u . We denote $\mathcal{A} = \mathcal{A}_s \cup \mathcal{A}_u$ as the set of attributes of all classes.

2.1. Pseudo-Attribute-Based Method

The overall architecture of our method is shown in Figure 1. We base our approach on a generative adversarial network, which has shown strong performance in recent ZSL methods [8, 5, 9, 10, 12]. As illustrated in Figure 1, our method contains two stages. In stage 1, an attribute decoder is trained with a generative model, learning to generate semantic attributes from visual features. Then, in stage 2, we produce pseudo-attributes for the unlabeled unseen data and integrate them into the generative process to exploit the semantic information in unseen classes, which helps synthesize semantically consistent unseen features. Finally, we utilize the generator to synthesize unseen features for ZSL classification.

Stage 1. As shown in Figure 1, we train a three-branch GAN [13], which consists of a generator, an attribute decoder, and two discriminators. The generator $G(z, a)$ synthesizes a feature from random noise and an attribute. The conditional seen discriminator $D_s(x, a)$ takes a seen feature and its corresponding attribute as input and outputs a real value indicating the degree of realism or fakeness. The unconditional unseen discriminator $D_u(x)$ takes an unseen feature as input and outputs its realism/fakeness score for unseen classes. For seen class data $(x^s, y^s, a^s) \in \mathcal{S}$, since we have matched feature and attribute, the generator G and the seen discriminator D_s construct a conditional GAN, and it is optimized by the ad-

versarial loss:

$$L_S = \mathbb{E}[D_s(x^s, a^s)] - \mathbb{E}[D_s(\tilde{x}^s, a^s)] - \lambda \mathbb{E}[(\|\nabla D_s(\tilde{x}^s, a^s)\|_2 - 1)^2], \quad (1)$$

where $\tilde{x}^s = G(z, a^s)$, $z \sim \mathcal{N}(0, 1)$ is the synthesized feature, λ is the penalty coefficient, and $\hat{x}^s = \alpha x^s + (1 - \alpha)\tilde{x}^s$, $\alpha \sim U(0, 1)$. For the unseen feature $x^u \in \mathcal{U}$, we randomly sample an unseen attribute $a^u \in \mathcal{A}_u$. Since the feature and attribute are mismatched, the generator G and the unseen discriminator D_u construct an unconditional GAN, and it is optimized by:

$$L_U = \mathbb{E}[D_u(x^u)] - \mathbb{E}[D_u(\tilde{x}^u)] - \lambda \mathbb{E}[(\|\nabla D_u(\tilde{x}^u)\|_2 - 1)^2], \quad (2)$$

where $\tilde{x}^u = G(z, a^u)$, $z \sim \mathcal{N}(0, 1)$, λ is the penalty coefficient, and $\hat{x}^u = \beta x^u + (1 - \beta)\tilde{x}^u$, $\beta \sim U(0, 1)$. In order to generate semantically consistent features and generate pseudo-attributes for the next stage, we add an auxiliary attribute decoder $Dec: \mathcal{X} \rightarrow \mathcal{A}$. And the cycle-consistency of the attributes is achieved by using ℓ_1 reconstruction loss:

$$L_R = \mathbb{E}[\|Dec(x^s) - a^s\|_1] + \mathbb{E}[\|Dec(\tilde{x}^s) - a^s\|_1] + \mathbb{E}[\|Dec(\tilde{x}^u) - a^u\|_1]. \quad (3)$$

Hence, the overall model is optimized by combining adversarial losses and the reconstruction loss as follows:

$$\min_{D_s, D_u, Dec} \max_G L_S + L_U + w L_R, \quad (4)$$

where w is a hyper-parameter for weighting the reconstruction loss.

Stage 2. From the well-trained attribute decoder, we can obtain pseudo-attributes of the unlabeled data and utilize them as the conditional information of the unseen discriminator,

i.e., we transform the unconditional discriminator $D_u(x)$ to a conditional one $D_u(x, a)$ to exploit the class-wise information of the unseen data. After the transformation, we reinitialize the generator G and the two discriminators, D_s and D_u , to make the generator and the discriminators compatible. Only the attribute decoder Dec maintains the weights trained in stage 1. To optimize the discriminator $D_u(x, a)$, we first obtain the pseudo-attribute $\tilde{a}^u = Dec(x^u)$ and consider it to be the corresponding attribute of the unseen data. And we concatenate and input them into $D_u(x, a)$ as [14]. Thus the adversarial loss is modified as follows:

$$L'_U = \mathbb{E}[D_u(x^u, \tilde{a}^u)] - \mathbb{E}[D_u(\hat{x}^u, \tilde{a}^u)] - \lambda \mathbb{E}[(\|\nabla D_u(\hat{x}^u, \tilde{a}^u)\|_2 - 1)^2], \quad (5)$$

where $\tilde{x}^u = G(z, \tilde{a}^u)$, $z \sim \mathcal{N}(0, 1)$, and $\hat{x}^u = \gamma x^u + (1 - \gamma)\tilde{x}^u$, $\gamma \sim U(0, 1)$. Furthermore, we modify the reconstruction loss by replacing a^u with \tilde{a}^u . Thus the attribute decoder is fine-tuned by:

$$L'_R = \mathbb{E}[\|Dec(x^s) - a^s\|_1] + \mathbb{E}[\|Dec(\tilde{x}^s) - a^s\|_1] + \mathbb{E}[\|Dec(\tilde{x}^u) - \tilde{a}^u\|_1]. \quad (6)$$

To this end, our full objective in stage 2 becomes:

$$\min_{D_s, D_u, Dec} \max_G L_S + L'_U + w' L'_R, \quad (7)$$

and w' is the loss coefficient.

2.2. Zero-Shot Classification

Once the generator is trained, we leverage it to synthesize unseen-class features $\tilde{\mathcal{U}}$ from corresponding unseen attributes in order to train the ZSL classifier. Typically, there are two kinds of test protocol in ZSL: conventional ZSL and generalized ZSL. For conventional ZSL, test samples only come from unseen categories. Thus, we use $\tilde{\mathcal{U}}$ to train a classifier by minimizing the cross-entropy loss: $\min_{\theta} -\frac{1}{|\mathcal{Y}|} \sum_{(x, y) \sim (\mathcal{X}, \mathcal{Y})} \log P(y|x; \theta)$, where $P(y|x; \theta) = \frac{\exp(\theta_y^T x)}{\sum_{j=1}^{|\mathcal{Y}|} \exp(\theta_j^T x)}$ is classification probability and θ denotes classifier parameters. For generalized ZSL, test samples come from both seen and unseen classes. To deal with the bias problem, we adopt the cascaded classifier mentioned in [15, 16]. We first train a binary classifier with \mathcal{U} and \mathcal{S} , which coarsely classify samples into seen or unseen classes. Then, we input the ‘seen-class’ data to the seen-class classifier and the ‘unseen-class’ data to the unseen-class classifier.

3. EXPERIMENTS

Datasets. We evaluate our method on five standard zero-shot object recognition datasets: CUB [21], FLO [22], SUN [23], AWA1 [1], and AWA2 [24], containing 200, 102, 717, 50, and 50 categories, respectively. We follow the same split setting,

Table 1. Conventional zero-shot learning performance. Both inductive (IN) and transductive (TR) results are shown in terms of average per-class top-1 accuracy. **Red** and **blue** denote the highest and second highest results.

		CUB	AWA1	AWA2	SUN	FLO
IN	DeViSE [17]	52.0	54.2	59.7	56.5	45.9
	DAZLE [7]	65.9	-	-	-	-
	CE-GZSL [6]	77.5	71.0	70.4	63.3	70.6
	LisGAN [18]	58.8	70.6	-	61.7	69.6
	IZF-Softmax [19]	-	-	-	-	-
	f-CLSGAN [5]	57.3	68.2	-	60.8	67.2
TR	GXE [20]	61.3	89.8	83.2	63.5	-
	GMN [12]	64.6	82.5	-	64.3	-
	f-VAEGAN-D2 [9]	71.1	-	89.8	70.1	89.1
	SGDN [10]	74.9	92.3	93.4	68.4	81.8
	STHS-GAN [16]	77.4	-	94.9	67.5	-
	TF-VAEGAN [8]	74.7	-	92.1	70.9	92.6
	ours	79.7	94.2	94.2	73.8	94.9

evaluation metric, and semantic embedding mentioned in [24] for a fair comparison.

Evaluation Metric. In conventional ZSL, where test images only come from unseen classes, we measure it by the average per-class top-1 accuracy. In generalized ZSL, test images come from either seen or unseen classes. Thus it is measured by the average per-class top-1 accuracy on seen and unseen classes, denoted as S and U , respectively. Furthermore, to measure the comprehensive performance, we use the harmonic mean as the final metrics: $H = \frac{2*U*S}{U+S}$.

Implementation Details. Following [8, 9], we adopt VAE-GAN as our generative model. The generator G , discriminators D_s, D_u , and attribute decoder Dec are implemented as two-layer fully-connected networks with 4,096 hidden dimensions. The classifiers at the test stage input the transformed features mentioned in [8]. The dimension of random noise z is equal to the attribute dimension ($\mathbb{R}^{d_z} = \mathbb{R}^{d_a}$). We use Adam optimizer with 10^{-4} learning rate. We set hyperparameters $\lambda = 5$, $w = 0.1$, and $w' = 0.1$ following [8].

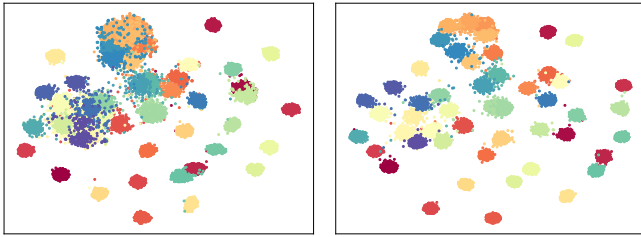
3.1. State-of-the-art Comparison

We select several advanced inductive ZSL and transductive ZSL methods for comparison. (1) For inductive ZSL, we choose six competitive methods, including DeVISE [17], DAZLE [7], CE-GZSL [6], LisGAN [18], IZF-Softmax [19], and f-CLSGAN [5]. (2) For transductive ZSL, we choose six state-of-the-art methods, including GXE [20], GMN [12], f-VAEGAN-D2 [9], SGDN [10], STHS-GAN [16], and TF-VAEGAN [8]. Our method falls into the transductive type.

Results of conventional zero-shot learning. As shown in Table 1, our method achieves state-of-the-art results on five ZSL datasets in competition with both inductive methods and transductive methods. Our method acquires excellent performance and establishes new benchmarks on four datasets. Especially on CUB, SUN, and FLO, the improvements are

Table 2. Generalized zero-shot learning performance. Both inductive (IN) and transductive (TR) results are shown. U = unseen acc, S = seen acc, H = harmonic mean of seen and unseen accs.

Method		CUB			AWA1			AWA2			SUN			FLO		
		U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
IN	DeViSE [17]	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	16.9	27.4	20.9	9.9	44.2	16.2
	DAZLE [7]	56.7	59.6	58.1	-	-	-	60.3	75.7	67.1	52.3	24.3	33.2	-	-	-
	CE-GZSL [6]	63.9	66.8	65.3	65.3	73.4	69.1	63.1	78.6	70.0	48.8	38.6	43.1	69.0	78.7	73.5
	LisGAN [18]	46.5	57.9	51.6	52.6	76.3	62.3	-	-	-	42.9	37.8	40.2	57.7	83.8	68.3
	IZF-Softmax [19]	52.7	68.0	59.4	61.3	80.5	69.6	60.6	77.5	68.0	52.7	57.0	54.8	-	-	-
	f-CLSGAN [5]	43.7	57.7	49.7	57.9	61.4	59.6	52.1	68.9	59.4	42.6	36.6	39.4	59.0	73.8	65.6
TR	GXE [20]	57.0	68.7	62.3	87.7	89.0	88.4	80.2	90.0	84.8	45.4	58.1	51.0	-	-	-
	GMN [12]	60.2	70.6	65.0	70.8	79.2	74.8	-	-	-	57.1	40.7	47.5	-	-	-
	f-VAEGAN-D2 [9]	61.4	65.1	63.2	-	-	-	84.8	88.6	86.7	60.6	41.9	49.6	78.7	87.2	82.7
	SGDN [10]	69.9	70.2	70.1	87.3	88.1	87.7	88.8	89.3	89.1	62.0	46.0	52.8	78.3	91.4	84.4
	STHS-GAN [16]	77.4	74.5	72.8	-	-	-	94.9	92.3	93.6	67.5	44.8	53.9	-	-	-
	TF-VAEGAN [8]	69.9	72.1	71.0	-	-	-	87.3	89.6	88.4	62.4	47.1	53.7	91.8	93.2	92.5
	ours	79.3	73.2	76.1	94.1	90.0	92.0	93.6	91.7	92.6	72.4	48.9	58.4	94.6	93.9	94.3



(a) TF-VAEGAN synthesized features (b) Our method synthesized features

Fig. 2. The t-SNE visualization of unseen features synthesized by (a) TF-VAEGAN [8] and (b) our method on CUB.

more than 2%. The reason is that they are very fine-grained datasets. In this case, features in different unseen classes are close to each other. Therefore, with the aid of pseudo-attributes, we synthesize more discriminative features and generate more clear boundaries for unseen classes, which improves ZSL performance. Moreover, our method consistently outperforms SGDN [10], f-VAEGAN-D2 [9], and TF-VAEGAN [8] on all datasets, because they all neglect the semantic information in unseen classes.

Results of generalized zero-shot learning. We also conduct experiments on generalized ZSL. As shown in Table 2, our method achieves either the best or second best results among most of the results of seen accuracy, unseen accuracy, and the harmonic mean. Our method obtains 3.3%, 3.6%, 3.6%, and 1.8% improvement in harmonic mean for CUB, AWA1, SUN, and FLO datasets, respectively. We attribute this gain to the pseudo-attribute-based method to make clear decision boundaries for different unseen classes.

3.2. Ablation Study

t-SNE Visualization. To further verify our method, we conduct t-SNE [25] visualization for synthesized unseen features

of our method and TF-VAEGAN [8] on CUB. As shown in Figure 2, compared to TF-VAEGAN [8], the synthesized features of our method are more compact and more distinct, especially in the top-left corner of the figure.

Table 3. Top-1 ZSL accuracy of different pretrained epochs at stage 1 on the CUB dataset.

epochs	0	100	200	300	400	500
acc	74.5	77.9	78.4	79.2	79.3	79.7

Importance of stage 1. We take experiments to verify the importance of pretraining at stage 1. Table 3 shows the conventional ZSL results of our method of CUB dataset with different training epochs in stage 1. Table 3 demonstrates that (1) without stage 1 (epoch=0), our method performs even worse than TF-VAEGAN [8], and (2) including stage 1 in training has a positive effect, and the more epochs we pretrain at stage 1, the better results we finally achieve.

4. CONCLUSION

This paper proposes a pseudo-attribute-based method for transductive ZSL. It exploits semantic information of unseen classes by generating and incorporating the pseudo-attributes into the generative process, which helps synthesize discriminative unseen features. We conduct extensive experiments to show the superiority of our method over prior methods.

5. ACKNOWLEDGEMENTS

This work was in part funded by National Natural Science Foundation of China (Grant No. 62276256), Beijing Nova Program under Grant Z211100002121108 and CAAI-Huawei MindSpore Open Fund. Our work is partially based on the MindSpore framework.

6. REFERENCES

- [1] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in *CVPR*, 2009, pp. 951–958. 1, 3
- [2] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth, “Describing objects by their attributes,” in *CVPR*, 2009, pp. 1778–1785. 1
- [3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, “Efficient estimation of word representations in vector space,” *arXiv*, 2013. 1
- [4] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele, “Learning deep representations of fine-grained visual descriptions,” in *CVPR*, 2016, pp. 49–58. 1
- [5] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata, “Feature generating networks for zero-shot learning,” in *CVPR*, 2018, pp. 5542–5551. 1, 2.1, 1, 3.1, 2
- [6] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang, “Contrastive embedding for generalized zero-shot learning,” in *CVPR*, 2021, pp. 2371–2381. 1, 1, 3.1, 2
- [7] D. Huynh and E. Elhamifar, “Fine-grained generalized zero-shot learning via dense attribute-based attention,” in *CVPR*, 2020, pp. 4482–4492. 1, 1, 3.1, 2
- [8] 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, pp. 479–495. 1, 2.1, 1, 3, 3.1, 2, 2, 3.1, 3.2, 3.2
- [9] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata, “f-vaegan-d2: A feature generating framework for any-shot learning,” in *CVPR*, 2019, pp. 10275–10284. 1, 2.1, 1, 3, 3.1, 2, 3.1
- [10] Jiamin Wu, Tianzhu Zhang, Zheng-Jun Zha, Jiebo Luo, Yongdong Zhang, and Feng Wu, “Self-supervised domain-aware generative network for generalized zero-shot learning,” in *CVPR*, 2020, pp. 12767–12776. 1, 2.1, 1, 3.1, 2, 3.1
- [11] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio, “Generative adversarial nets,” in *NeurIPS*, 2014, pp. 2672–2680. 1
- [12] Mert Bulent Sariyildiz and Ramazan Gokberk Cinbis, “Gradient matching generative networks for zero-shot learning,” in *CVPR*, 2019, pp. 2168–2178. 2.1, 1, 3.1, 2
- [13] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville, “Improved training of wasserstein gans,” in *NeurIPS*, 2017, pp. 5767–5777. 2.1
- [14] Dapeng Hu, Jian Liang, Qibin Hou, Hanshu Yan, and Yunpeng Chen, “Adversarial domain adaptation with prototype-based normalized output conditioner,” *TIP*, vol. 30, pp. 9359–9371, 2021. 2.1
- [15] Bo Liu, Lihua Hu, Zhanyi Hu, and Qiulei Dong, “Hardboost: Boosting zero-shot learning with hard classes,” *arXiv*, 2022. 2.2
- [16] Liu Bo, Qiulei Dong, and Zhanyi Hu, “Hardness sampling for self-training based transductive zero-shot learning,” in *CVPR*, 2021, pp. 16499–16508. 2.2, 1, 3.1, 2
- [17] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov, “Devise: A deep visual-semantic embedding model,” in *NeurIPS*, 2013, pp. 2121–2129. 1, 3.1, 2
- [18] 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, pp. 7402–7411. 1, 3.1, 2
- [19] Yuming Shen, Jie Qin, Lei Huang, Li Liu, Fan Zhu, and Ling Shao, “Invertible zero-shot recognition flows,” in *ECCV*, 2020, pp. 614–631. 1, 3.1, 2
- [20] Kai Li, Martin Renqiang Min, and Yun Fu, “Rethinking zero-shot learning: A conditional visual classification perspective,” in *ICCV*, 2019, pp. 3583–3592. 1, 3.1, 2
- [21] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie, “The caltech-ucsd birds-200-2011 dataset,” Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011. 3
- [22] Maria-Elena Nilsback and Andrew Zisserman, “Automated flower classification over a large number of classes,” in *ICVGIP*, 2008, pp. 722–729. 3
- [23] Genevieve Patterson and James Hays, “Sun attribute database: Discovering, annotating, and recognizing scene attributes,” in *CVPR*, 2012, pp. 2751–2758. 3
- [24] Yongqin Xian, H. Christoph Lampert, Bernt Schiele, and Zeynep Akata, “Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly,” *TPAMI*, vol. 41, no. 9, pp. 2251–2265, 2018. 3
- [25] Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-sne,” *JMLR*, vol. 9, no. 86, pp. 2579–2605, 2008. 3.2