SHAPING VISUAL REPRESENTATIONS WITH ATTRIBUTES FOR FEW-SHOT LEARNING

Haoxing Chen, Huaxiong Li, Yaohui Li, Chunlin Chen

Nanjing University, Nanjing, China

{haoxingchen, yaohuili}@smail.nju.edu.cn, {huaxiongli, clchen}@nju.edu.cn

ABSTRACT

Few-shot recognition aims to recognize novel categories under low-data regimes. Due to the scarcity of images, machines cannot obtain enough effective information, and the generalization ability of the model is extremely weak. By using auxiliary semantic modalities, recent metric-learning based few-shot learning methods have achieved promising performances. However, these methods only augment the representations of support classes, while query images have no semantic modalities information to enhance representations. Instead, we propose attribute-shaped learning (ASL), which can normalize visual representations to predict attributes for query images. And we further devise an attribute-visual attention module (AVAM), which utilizes attributes to generate more discriminative features. Our method enables visual representations to focus on important regions with attributes guidance. Experiments demonstrate that our method can achieve competitive results on CUB and SUN benchmarks. We provide a PyTorch implementation of ASL. Our code is available at https://github.com/chenhaoxing/ASL.

***Index Terms*—** Attribute-shaped learning, few-shot learning, multi modality

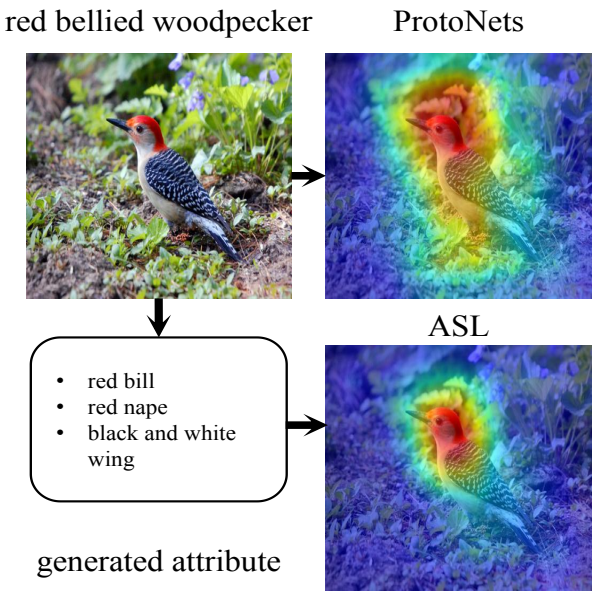

Fig. 1. An illustration of the effect of our proposed attribute-shaped learning method.

1. INTRODUCTION

Deep learning has achieved outstanding performance in many visual tasks [1, 2]. However, training deep models often require a lot of labeled data. Inspired by the ability of humans to recognize new objects with few labeled data, few-shot learning aims to imitate this ability. The main challenge in few-shot learning is to make the best use of labeled data to learn transferable knowledge.

Few-shot learning algorithms can be roughly divided into three categories: meta-learning methods [3], data augmentation methods [4] and metric-learning based methods [5, 6, 7, 8] . Recently, metric-learning based methods have attracted extensive attention due to their simplicity and effectiveness. These methods mainly focus on improving the informativeness and discriminability of the learned concept representations. However, these methods classify images in the context of unimodal learning. Recall the way that humans learn new concepts. For instance, we can learn about the red-billed blue magpie not by seeing thousands of images, but by being told that a red-billed blue magpie is a bird with a red beak and blue feathers, and learning from a few images. *In other words, language can help humans learn new visual objects.*

To imitate this ability, some works introduce auxiliary semantic modalities to enhance prototype learning in few-shot learning. FLS-MCS [9] proposed a model close to "infant learning", which uses a variety of complex semantic information to tackle few-shot learning. Pavel *et al.* [10] took attribute features as supplementary information and enhanced the representation ability of feature extractor by adding regularization terms in the loss function. AM3 [11] utilized convex combination to adaptively mix the semantic structures of the two modalities. Dual TriNet [12] is a novel auto-encoder network, which directly synthesizes instance features by leveraging semantics.

These existing methods [9, 10, 11] assume that auxiliary textual information is only available for the support set, not the query set. Following this realistic setup, these methods

only learn the feature representation of the supporting set with the help of textual information. There is no special mechanism designed for query images to achieve the same effect as textual information assistance, resulting in a potential loss of knowledge.

Towards this end, we propose attributed-shaped learning (ASL), an end-to-end model that generates corresponding attributes through image features and learns more discriminative visual features combined with attributes. Specifically, since query images have no auxiliary textual information, a visual-attribute generator is proposed to generate attribute features. Then, we propose an attribute-visual attention module (AVAM) to utilize attribute as auxiliary knowledge and explore important visual information. AVAM contains two parallel branches, i.e., efficient channel-spatial attention module (EAM) and attribute attention module (A2M). EAM finds important channels and regions, and A2M utilizes attributes to find important visual representations. As shown in Fig. 1, our ASL can focus on more representative local features

To summarize, our main contributions are as follows: (i) we propose a novel attribute-shaped learning method, which can generate attributes and provide additional textual information for query images to assist the learning of visual representations. (ii) We propose an attribute-visual attention module (AVAM), which can combine attribute and visual information to highlight the important information in images. (iii) We show that our method achieves state-of-the-art results on the CUB and SUN in both 1-shot and 5-shot regimes.

2. METHODOLOGY

In this paper, we propose an attribute-visual multi-modal few-shot image classification method. To mimic the process of human learning new knowledge, we introduce auxiliary textual information to help learn transferable meta-knowledge.

Due to the scarcity of samples, we formulate the problem under episodic training paradigm [3, 5]. In general, the model is trained by a series of N-way M-shot episodes and each episode can be seen as an independent task. Each task τ is formed by randomly selecting N categories from training set $\mathcal{D}_{train}$, and then sampling support set $\mathcal{S} = \{(s_i, a_i, y_i)\}_{i=1}^{N\times M}$ and query set $\mathcal{Q} = \{(q_j, a_j, y_j)\}_{i=1}^{q}$. Here, x, $a \in \mathbb{R}^A$ and y represent the image, attribute vector and label respectively. Note that $\mathcal{Q}$ contains different examples from the same N categories and A is determined by the dataset. After learning on trainning set $\mathcal{D}_{train}$, a model is evaluated on unlabeled test set $\mathcal{D}_{test}$.

While the proposed method is applicable to nearly any metric-learning based framework, we use Prototypical Nets [5]. Prototypical Nets are widely used in few-shot learning because of their simplicity and effectiveness. Prototypical Nets aim to generate the prototype for each category and determine the category of query samples by calculating the distances to each prototype. Specifically, Prototypical Nets firstly obtain the feature representations of each image through a feature extractor $\mathcal{F}_\theta$. And then Prototypical Nets compute prototype p_n for each support class n:

$$p^n = \frac{1}{M}\sum_{i=1}^{M}\mathcal{F}_\theta(s_i^n) \tag{1}$$

where s_i^n is the i-th support image of support category n. For each query image (q, y_q), Prototypical Nets calculate the Euclidean distance between q and each prototype, and use softmax function to obtain the probability distribution:

$$p(\hat{y}_q = n|q) = \frac{\exp(-\mathrm{d}(\mathcal{F}_\theta(q), p^n))}{\sum_k \exp(-\mathrm{d}(\mathcal{F}_\theta(q), p^k))} \tag{2}$$

$\mathcal{F}_\theta$ is then trained to minimize the classification loss:

$$\mathcal{L}_{cls} = -\sum_{j=1}^{q}\log p(\hat{y}_q = y_q|q) \tag{3}$$

2.1. Attribute-Shaped Learning

To enable both the query images and support categories to have additional attribute descriptions, we generate attribute descriptions for each sample. Specifically, given an image x, through $\mathcal{F}_\theta$, we can get visual representation $\mathcal{F}_\theta(x) \in \mathbb{R}^{H\times W\times C}$, where W, H and C are width, height and channel dimension. Then define an attribute predict model g_ϕ (e.g., MLPs), which can predict A number of attributes. To validate the quality of the generated attribute vector, we define the loss function:

$$\mathcal{L}_{attr} = -\frac{1}{A}\sum_{k=1}^{q+NM}\sum_{i=1}^{A}(\hat{a}_i^k - a_i^k)^2 \tag{4}$$

where a_i^k is the observed attributes of the k-th samples and $\hat{a}_i^k$ is the predicted ones.

The whole optimization objectives of the whole model are as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \gamma\cdot\mathcal{L}_{attr} \tag{5}$$

where γ is the weighting factor of $\mathcal{L}_{attr}$.

2.2. Attribute-Visual Attention Module

To use attribute vectors to generate more discriminating features, we propose an attribute-visual attention module (AVAM). As shown in Fig. 2, AVAM consists of two parallel branches, i.e., efficient channel-spatial attention module (EAM) and attribute attention module (A2M).

2.2.1. Efficient channel-spatial attention module

Inspired by [13], we propose a novel efficient channel-spatial attention module (EAM) to learn which channels to focus and

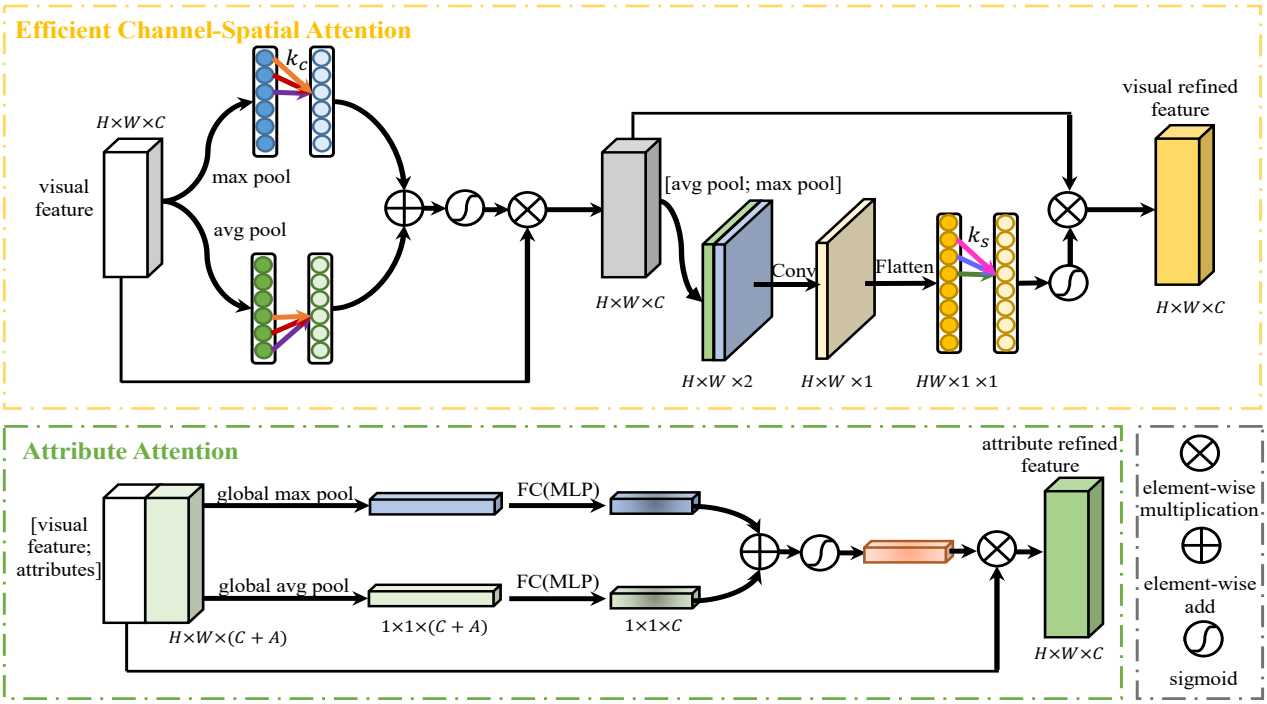

Fig. 2. Illustration of attribute-visual attention module

which areas to focus. As shown in Fig. 2, EAM has two sequential sub-modules: channel and spatial. Let the output of feature extractor $\mathcal{F}_\theta$ be $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$. The overall efficient attention process can be summarized as:

$$\mathcal{X}' = \mathcal{M}_c(\mathcal{X}) \otimes \mathcal{X} \tag{6}$$

$$\mathcal{X}_{eam} = \mathcal{M}_s(\mathcal{X}) \otimes \mathcal{X}' \tag{7}$$

where $\otimes$ denotes element-wise multiplication, $\mathcal{M}_c(\mathcal{X}) \in \mathbb{R}^{1\times 1\times C}$ denotes channel attention map and $\mathcal{M}_s(\mathcal{X}) \in \mathbb{R}^{H\times W\times 1}$ denotes spatial attention map.

Efficient channel attention. We first get two different spatial context descriptors by global average pooling and global max pooling: $\mathcal{X}^c_{max}$ and $\mathcal{X}^c_{avg}$. Both descriptors are then forwarded to a shared 1D convolutional layer (with a kernel size of k_c) to generate the channel attention map $\mathcal{M}_c \in \mathbb{R}^{1\times 1\times C}$. In short, the efficient channel attention is computed as:

$$\begin{aligned}\mathcal{M}_c(\mathcal{X}) =& \sigma(\mathrm{C1D}_{k_c}(\mathrm{AvgPool}(\mathcal{X})) + \mathrm{C1D}_{k_c}(\mathrm{MaxPool}(\mathcal{X}))) \\ =& \sigma(\mathrm{C1D}_{k_c}(\mathcal{X}^c_{avg}) + \mathrm{C1D}_{k_c}(\mathcal{X}^c_{max}))\end{aligned} \tag{8}$$

where σ denotes the sigmoid function and C1D indicates 1D convolution.

Efficient spatial attention. Similarly, we first apply average-pooling and max-pooling operations along the channel dimension and concatenate the pooled features. Then we aggregate these two features via 2D convolution. Finally, we convert the size of aggregated features from $H \times W \times 1$ to $HW \times 1$ and use 1D convolutional layer (with kernel size of k_s) to generate the spatial attention map $\mathcal{M}_s \in \mathbb{R}^{H\times W\times 1}$. In short, the efficient spatial attention is computed as:

$$\mathcal{M}_s(\mathcal{X}') = \sigma(\mathrm{C1D}_{k_s}(f^{1\times 1}([\mathrm{AvgPool}(\mathcal{X}); \mathrm{MaxPool}(\mathcal{X})])) \tag{9}$$

where $f^{1\times 1}$ represents a 2D convolution operation with the filter size of 1×1.

2.2.2. *Attribute attention module*

A2M utilizes attribute vectors to generate features for specific attributes. Specifically, we use the attributes provided by the dataset for support images and the generated attributes for query images. Given a feature map $\mathcal{X}$ and attributes vector $a \in \mathbb{R}^A$, we first broadcast a long height and width dimension of $\mathcal{X}$, then concatenate $\mathcal{X}$ and a to get hybrid feature$\mathcal{X}^{v-a} \in \mathbb{R}^{H\times W\times(C+A)}$. Then we use global average pooling and global max pooling to aggregate channel information and get $\mathcal{X}^{v-a}_{avg} \in \mathbb{R}^{1\times 1\times(C+A)}$ and $\mathcal{X}^{v-a}_{max} \in \mathbb{R}^{1\times 1\times(C+A)}$. Finally, an attention generating network (i.e., one layer MLP) is adopted to generate attribute attention map $\mathcal{M}_a$. In short, the attribute attention is computed as:

$$\mathcal{M}_c(\mathcal{X}^{v-a}) = \sigma(\mathrm{MLP}(\mathcal{X}^{v-a}_{avg})) + \mathrm{MLP}(\mathcal{X}^{v-a}_{avg})) \tag{10}$$

$$\mathcal{X}_{a2m} = \mathcal{M}_c(\mathcal{X}^{v-a}) \otimes \mathcal{X} \tag{11}$$

where σ denotes the sigmoid function.

Model	Venue	Modality	Backbone	CUB	
				1-shot	5-shot
Prototypical Nets [5]	NeurIPS 2017	V	Conv-64F	51.31±0.91	70.77±0.69
Relation Nets [6]	CVPR 2018	V	Conv-64F	62.45±0.98	76.11±0.69
AM3 [11]	NeurIPS 2019	V&T	Conv-64F	**73.78±0.28**	81.39±0.26
Comp. [10]	ICCV 2019	V&T	Conv-64F	53.60±0.00	74.60±0.00
LRPABN [14]	T-MM 2021	V	Conv-64F	67.97±0.44	78.26±0.22
IEPT [15]	ICLR 2021	V	Conv-64F	69.97±0.49	**84.33±0.33**
AGAM [16]	AAAI 2021	V&T	Conv-64F	75.87±0.29	**81.66±0.25**
ASL	Ours	V&T	Conv-64F	**74.28±0.27**	79.83±0.13
Prototypical Nets [5]	NeurIPS 2017	V	ResNet-12	68.80±0.00	76.40±0.00
Relation Nets [6]	CVPR 2018	V	ResNet-12	62.45±0.98	76.11±0.69
AM3 [7]	NeurIPS 2019	V&T	ResNet-12	73.60±0.00	79.90±0.00
Dual TriNet [12]	T-IP 2019	V&T	ResNet-18	69.61±0.46	84.10±0.35
FEAT [7]	CVPR 2020	V	ResNet-12	68.87±0.22	82.90±0.15
DeepEMD [17]	CVPR 2020	V	ResNet-12	75.65±0.83	**88.69±0.50**
AGAM [16]	AAAI 2021	V&T	ResNet-12	**79.58±0.25**	87.17±0.23
ASL	Ours	V&T	ResNet-12	**76.89±0.25**	**90.41±0.10**

Model	Venue	Modality	Backbone	SUN	
				1-shot	5-shot
Prototypical Nets [5]	NeurIPS 2017	V	Conv-64F	57.76±0.29	79.27±0.19
Relation Nets [6]	CVPR 2018	V	Conv-64F	49.58±0.35	76.21±0.19
Comp. [10]	ICCV 2019	V&T	ResNet-10	45.90±0.00	67.10±0.00
AM3 [7]	NeurIPS 2019	V&T	Conv-64F	**62.79±0.32**	79.69 ±0.23
AGAM [16]	AAAI 2021	V&T	Conv-64F	**65.15±0.31**	**80.08±0.21**
ASL	Ours	V&T	Conv-64F	61.71±0.28	**80.15±0.16**

Table 1. Comparison with other state-of-the-art methods with 95% confidence intervals on CUB-200-2011 and SUN. (Top two performances are in bold font.)

After feature $\mathcal{X}$ passes through EAM and A2M respectively, we obtain the final feature by dynamic weighting:

$$\mathcal{X}_att = \alpha \cdot \mathcal{X}_{eam} + (1-\alpha) \cdot \mathcal{X}_{a2m} \tag{12}$$

where α is a learnable parameter.

3. EXPERIMENTS

3.1. Datasets

Caltech-UCSD Birds-200-2011 (CUB) [18] is a fine-grained dataset of bird species. CUB consists of 11,788 images and 312 attributes from 200 bird classes. Following [19], we use 100/50/50 categories for training, validation, and evaluation respectively.

SUN Attribute Database (SUN) is a fine-grained scene recognition dataset that contains 14,340 images of 717 different categories and 102 attributes. We use 580/65/72 categories for training, validation, and evaluation respectively.

3.2. Experimental Settings

Backbone Networks. Following [7, 15], we conduct experiments with both shallow four layer convolutional Conv-64F [20] and ResNet-12 [21].

Implementation Details. We train our model from scratch by Adam optimizer with an initial learning rate 1×10^{-3}. For ASL, we set superparameter $\gamma = 1$, $k_c = 3$ and $k_a = 3$ for all experiments. We train our model for 60,000 iterations. During the test stage, we report the top-1 mean accuracy over 10,000 tasks.

3.3. Results and Analysis

Comparisons with the state-of-the-arts. Table 1 shows that ASL achieve new state-of-the-art for 5-shot tasks and have competitive performances for 1-shot tasks. To be more specific, our model is around 4.4%/4.0% and 4.4%/4.0% better than FEAT [5] and DeepEMD [17] on CUB with *ResNet12* for 1-shot and 5-shot tasks.

Comparisons with single visual modality based methods.

Table 2. Ablation study on our model. The experiments are conducted with *ResNet12* on CUB-200-2011.

Method	5-way 1-shot	5-way 5-shot
w/o ECA	71.86±0.27	86.13±0.13
w/o ESA	73.48±0.26	86.64±0.12
w/o EAM	69.67±0.29	85.18±0.13
w/o A2M	67.75±0.30	88.95±0.12
ASL	**76.89±0.25**	**90.41±0.10**

Table 3. Ablation test results of different attribute-shaped learning loss. The experiments are conducted with *ResNet12* on CUB-200-2011.

Loss	5-way 1-shot	5-way 5-shot
L1	62.89±0.22	85.93±0.12
smoothL1	74.67±0.29	86.87±0.12
soft margin	70.30±0.27	89.06±0.10
MSE	**76.89±0.25**	**90.41±0.10**

Multi-modality based methods (e.g., AM3, Dual TriNet and Comp.) generally outperform classic meta-learning based methods that rely on the single visual modality (e.g., Prototypical Nets, Relation Nets, and FEAT) by a large margin, which validates the effectiveness of using auxiliary semantic modalities.

Comparisons with multi-modality based methods. Among those methods, AM3 uses the convex combination to mix the semantic structures of the two modalities; Dual TriNet directly synthesizes instance features by leveraging semantics; Comp. adds regularization terms in loss function to enhance the representation ability of feature extractor. They all only augment the representations of support classes, while query images have no semantic modalities information to enhance representations. Our ASL outperforms other model variants on almost all tasks as we think it is equally important to generate attribute descriptions for each query image.

3.4. Ablation Study

Framework Design. As shown in Table 2, we conduct sufficient experiments to prove the effectiveness of our model design. The results show that every component in ASL has a significant contribution. For example, without the attribute attention module, the performance of the model will drop by 3.73% on 1-shot tasks.

Attribute-shaped learning loss. To prove that the MSE loss is an ideal choice for the attribute-shaped learning, we compare the results on CUB with four different candidate losses. As shown in Table 4, choosing the MSE loss in the attribute-shaped learning brings in better performance.

Effectiveness of EAM. As shown in Table 4, we compare the trainable parameters to prove that our EAM is both simple and effective. We replaced EAM with SE [22] and CBAM [23], respectively. Following [16], the reduction factors in both SE and CBAM are set to 4. Our EAM achieved better results with fewer trainable parameters. For example, our EAM obtains 10.0%/5.2% improvements over CBAM under the 5-way 1-shot/5-shot few-shot learning setting.

Table 4. Ablation test results of different visual attention methods. The experiments are conducted with *ResNet12* on CUB-200-2011.

Method	Params.	5-way 1-shot	5-way 5-shot
SE [22]	131.072K	73.82±0.27	86.34±0.11
CBAM [23]	131.170K	69.88±0.27	85.97±0.13
EAM	0.008K	**76.89±0.25**	**90.41±0.10**

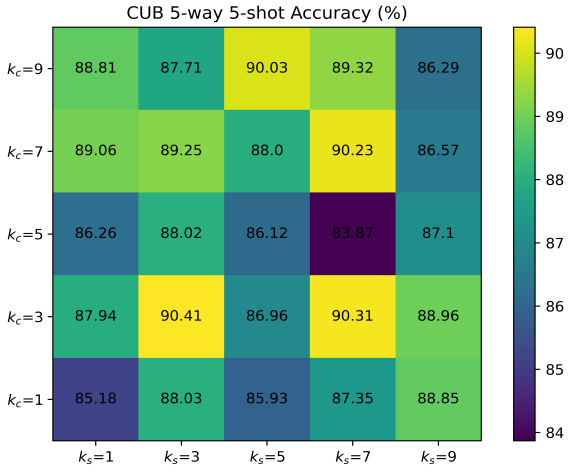

Fig. 3. Influence of superparameters k_c and k_s.

Influence of superparameter k_c and k_s. In the efficient, we need to choose suitable k_c and k_s. For this purpose, we conduct a contrast experiment on CUB under 5-way 5-shot settings by varying the value of $k_c \in \{1, 3, 5, 7, 9\}$ and the value of $k_s \in \{1, 3, 5, 7, 9\}$. Experimental results are shown in Fig. 3. It can be seen that when $k_c = 3$ and $k_s = 3$, the experimental result of ASL is the best.

3.5. Grad-CAM Visualization Analysis

To get a deeper understanding of our ASL, Fig. 4 visualizes the Grad-CAM [24] from Prototypical Network and ASL. It can be seen that Prototypical Networks fail to identify the most relevant regions of interest in the task. ASL improves the recognition performance by introducing additional attribute vectors to help the model focus on more representative local features.

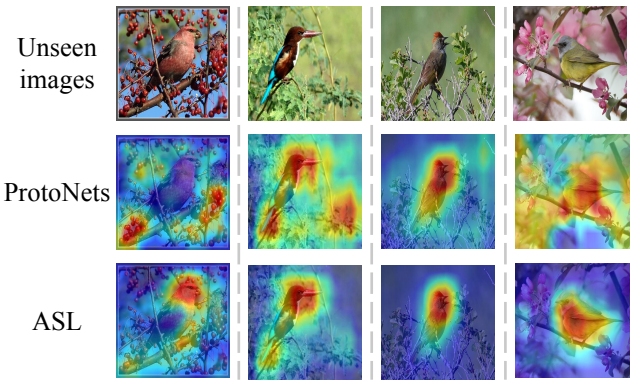

Fig. 4. Grad-Class Activation Mapping (Grad-CAM) visualization of four unseen images.

4. CONCLUSION

In this paper, we argue that auxiliary semantic modalities are necessary for query images. We propose an attribute-shaped learning method to generate corresponding attributes through visual features and learn more discriminative visual features. Experimental results on popular datasets illustrate the encouraging performance of our ASL, which has achieved competitive results with other state-of-the-art few-shot learning methods.

5. REFERENCES

[1] Chao Zhang, Huaxiong Li, Yuhua Qian, Chunlin Chen, and Yang Gao, "Pairwise relations oriented discriminative regression," *TCSVT*, vol. 31, pp. 2646–2660, 2021.

[2] Ge Gao, Pei You, Rong Pan, Shunyuan Han, Yuanyuan Zhang, Yuchao Dai, and Hojae Lee, "Neural image compression via attentional multi-scale back projection and frequency decomposition," in *ICCV*, 2021, pp. 14677–14686.

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017, vol. 70, pp. 1126–1135.

[4] Mandar Dixit, Roland Kwitt, Marc Niethammer, and Nuno Vasconcelos, "AGA: attribute-guided augmentation," in *CVPR*, 2017, pp. 3328–3336.

[5] Jake Snell, Kevin Swersky, and Richard S. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017, pp. 4077–4087.

[6] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018, pp. 1199–1208.

[7] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *CVPR*, 2020, pp. 8808–8817.

[8] Haoxing Chen, Huaxiong Li, Yaohui Li, and Chunlin Chen, "Sparse spatial transformers for few-shot learning," *arXiv preprint arXiv:2109.12932*, 2021.

[9] Eli Schwartz, Leonid Karlinsky, Rogério Schmidt Feris, Raja Giryes, and Alexander M. Bronstein, "Baby steps towards few-shot learning with multiple semantics," *arXiv preprint arXiv:1906.01905*, 2019.

[10] Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert, "Learning compositional representations for few-shot recognition," in *ICCV*, 2019, pp. 6371–6380.

[11] Chen Xing, Negar Rostamzadeh, Boris N. Oreshkin, and Pedro O. Pinheiro, "Adaptive cross-modal few-shot learning," in *NeurIPS 2019*, 2019, pp. 4848–4858.

[12] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal, "Multi-level semantic feature augmentation for one-shot learning," *TIP*, vol. 28, no. 9, pp. 4594–4605, 2019.

[13] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *CVPR*, 2020, pp. 11531–11539.

[14] Huaxi Huang, Junjie Zhang, Jian Zhang, Jingsong Xu, and Qiang Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *TMM*, vol. 23, pp. 1666–1680, 2021.

[15] Manli Zhang, Jianhong Zhang, Zhiwu Lu, Tao Xiang, Mingyu Ding, and Songfang Huang, "IEPT: instance-level and episode-level pretext tasks for few-shot learning," in *ICLR*, 2021.

[16] Siteng Huang, Min Zhang, Yachen Kang, and Donglin Wang, "Attributes-guided and pure-visual attention alignment for few-shot recognition," in *AAAI*, 2021, pp. 7840–7847.

[17] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *CVPR*, 2020, pp. 12200–12210.

[18] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona, "Caltech-ucsd birds 200," 2010.

[19] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *CVPR*, 2019, pp. 7260–7268.

[20] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, "Matching networks for one shot learning," in *NeurIPS*, 2016, pp. 3630–3638.

[21] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele, "Meta-transfer learning for few-shot learning," in *CVPR*, 2019, pp. 403–412.

[22] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[23] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "CBAM: convolutional block attention module," in *ECCV*, 2018, vol. 11211, pp. 3–19.

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.