# Shaping Visual Representations with Attributes for Few-Shot Recognition

Haoxing Chen, Huaxiong Li, Yaohui Li, Chunlin Chen

*Abstract*—**Few-shot recognition aims to recognize novel categories under low-data regimes. Some recent few-shot recognition methods introduce auxiliary semantic modality, i.e., category attribute information, into representation learning, which enhances the feature discrimination and improves the recognition performance. Most of these existing methods only consider the attribute information of support set while ignoring the query set, resulting in a potential loss of performance. In this letter, we propose a novel attribute-shaped learning (ASL) framework, which can jointly perform query attributes generation and discriminative visual representation learning for few-shot recognition. Specifically, a visual-attribute predictor (VAP) is constructed to predict the attributes of queries. By leveraging the attributes information, an attribute-visual attention module (AVAM) is designed, which can adaptively utilize attributes and visual representations to learn more discriminative features. Under the guidance of attribute modality, our method can learn enhanced semantic-aware representation for classification. Experiments demonstrate that our method can achieve competitive results on CUB and SUN benchmarks. Our source code is available at: https://github.com/chenhaoxing/ASL.**

*Index Terms*—**Attribute-shaped learning, few-shot learning, attribute-visual attention**

## I. INTRODUCTION

Deep learning has achieved outstanding performance in many visual tasks [1], [2], [3], [4]. However, training deep models often requires a lot of labeled data, which is not always accessible in real applications [5], [6], [7], [8]. Inspired by the ability of humans, few-shot learning aims to recognize new objects with few labeled training samples.

Recently, various successful few-shot recognition methods have been proposed, which can be roughly divided into three categories: meta-learning methods [9], [10], data augmentation methods [11] and metric-learning based methods [12], [13], [14]. In these methods, metric-learning based methods have attracted extensive attention due to their simplicity and effectiveness. Most of them mainly focus on enhancing the informativeness and discriminability of learned semantic visual representations to improve image recognition accuracy. However, these methods perform image classification in the context of uni-modal visual learning, i.e., only using the images. In the real-world, humans learn new concepts by leveraging multi-modal information rather than a single one [15], [16]. For
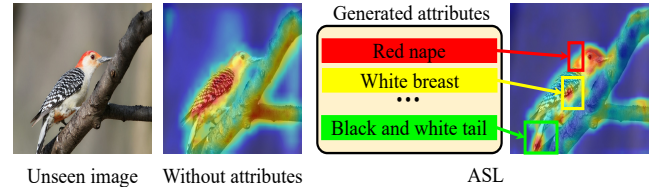
Fig. 1. An illustration of the effect of our proposed ASL method.

instance, one can learn about the *red billed blue magpie* by some images and attributes, such as *red bill* and *red nape*, instead of only seeing many pictures. *In other words, attribute information can help humans learn new visual objects* [17], [18], [19], [20].

To imitate this ability, some works introduce auxiliary semantic modalities to enhance feature learning in few-shot learning. Pavel *et al.* [17] took attribute features as supplementary information and enhanced the representation learning ability by adding regularization terms. AM3 [21] utilized convex combination to adaptively mix the semantic structures of visual representations and label semantics. Dual TriNet [22] is an auto-encoder network, which directly synthesizes instance features by leveraging semantics.

Although the above methods [17], [21], [22] have achieved impressive performance by leveraging semantic information, they only learn the feature representations of support set with the help of attribute information and ignore the query set that lacks the attribute information. To the best of our knowledge, there is no specific mechanism to explore the underlying attribute information of query samples, which is hopeful to improve the few-shot recognition performance.

Towards this end, we propose a novel attribute-shaped learning (ASL) model that generates corresponding attributes through image features and learns more discriminative visual features combined with attributes. Specifically, since query images have no auxiliary attribute information, a visual-attribute predictor (VAP) is proposed to generate attribute features. Then, we propose an attribute-visual attention module (AVAM), which can adaptively utilize attributes and visual representations to learn more discriminative features. AVAM contains two sub-modules, i.e., channel attention module (CAM) and pyramid spatial attention module (PSAM). CAM finds important channels, and PSAM finds multi-scale spatial-visual representations. Fig. 1 shows the effectiveness of our ASL method and classical ProtoNets [12]. It can be observed that, by incorporating the generated attribute information, our method can focus on more discriminative local regions for vi-
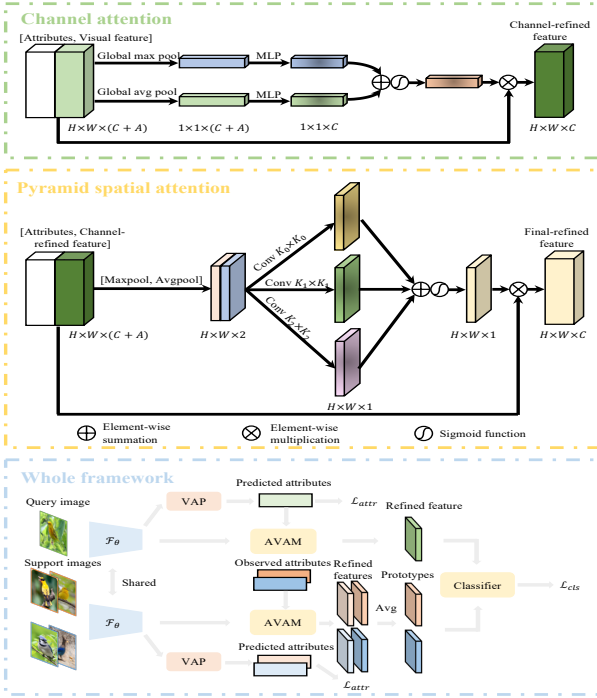
Fig. 2. Illustration of attribute-visual attention module

$\mathcal{F}_\theta$, and then compute prototype a $p_n$ for 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 image of support category $n$. For each query image $(q_j, y_j)$, ProtoNets calculate the Euclidean distance between $q$ and each prototype, and use softmax function to obtain the probability distribution:

$$p(\hat{y}_j = n|q_j) = \frac{\exp(-||\mathcal{F}_\theta(q_j) - p^n||^2)}{\sum_{k=1}^{N} \exp(-||\mathcal{F}_\theta(q_j), p^k||^2)}, \tag{2}$$

$\mathcal{F}_\theta$ is trained by minimizing the classification loss:

$$\mathcal{L}_{cls} = -\sum_{j=1}^{Q} \log p(\hat{y}_j = y_j|q_j). \tag{3}$$

### A. Visual-Attribute Predictor

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. We first use global-average pooling (GAP) to get pooled fetures for attribute prediction. Then define a visual-attribute predictor $g_\phi: \mathbb{R}^C \mapsto \mathbb{R}^A$, which can predict $A$ number of attributes. To validate the quality of the generated attribute vector, we define the loss function:

$$\mathcal{L}_{attr} = -\frac{1}{A} \sum_{i=1}^{N \times K + Q} \sum_{1}^{A} (\hat{a}_i^k - a_i^k), \tag{4}$$

where $a_i^k$ is the $i$-th observed attribute of the $k$-th sample and $\hat{a}_i^k$ is the predicted ones. The whole optimization objectives of the whole model are as follows:

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

where $\alpha$ is the weighting factor of $\mathcal{L}_{attr}$. The training strategy of ASL is illustrated in Algorithm 1.

### B. 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 sub-models, i.e., channel attention module (CAM) and pyramid spatial attention module (PSAM). CAM to blend cross-channel information and learn which channels to focus on. And PSAM can extract multi-scale spatial information at a more granular level by attributes guiding. Note that AVAM is a flexible module and can be easily added into any few-shot learning method.

Specifically, we use the attributes provided by the dataset for support images and the generated attributes for query images. Given a feature map $F^v \in \mathbb{R}^{H \times W \times C}$ and attributes vector $F^a \in \mathbb{R}^A$, we first broadcast $F^a$ along height and width dimension of $F^v$, and then concatenate $F^v$ and $F^a$ to get hybrid feature $F \in \mathbb{R}^{H \times W \times (C+A)}$. Through CAM, we can get channel-refined feature $F_c \in \mathbb{R}^{H \times W \times C}$. Then we broadcast

sual representation learning, contributing to better recognition performance. To summarize, our main contributions are as follows: 1) We propose a novel visual-attribute predictor (VAP), which can generate attributes for query images to assist the learning of visual representations. 2) We propose an attribute-visual attention module (AVAM), which can simultaneously leverage attribute and visual information to adaptively learn the discriminative features in images. 3) We show that our method achieves competitive results compared to other state-of-the-art methods.

## II. METHODOLOGY

In few-shot learning, 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 $\tau$ 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 $\mathcal{Q} = \{(q_j, a_j, y_j)\}_{j=1}^{Q}$ from these $N$ categories. 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 a test set $\mathcal{D}_{test}$.

The proposed method is applicable to nearly any metric-learning based framework. Since ProtoNets [12] is simple and effective, we choose it as our baseline framework. ProtoNets aim to generate the prototype for each category and determine the category of query samples by calculating the distances to each prototype. Specifically, ProtoNets firstly obtain the feature representations of each image through a feature extractor

**Algorithm 1** Training strategy of ASL

**Input:** training set $\mathcal{D}_{train}$;

1: **for all** iteration=1, ..., MaxIteration **do**
2:     sample $N$-way $M$-shot task $(\mathcal{Q}, \mathcal{S})$ from $\mathcal{D}_{train}$;
3:     compute support visual features $\{F_\theta(s_i)_{i=1}\}^{N \times M}$;
4:     compute refined support features Eq. (6)-(9);
5:     compute prototypes by Eq. (1);
6:     **for** $j$ in $\{1, ..., Q\}$ **do**
7:         compute query visual feature $F_\theta(q_j)$;
8:         compute refined query features Eq. (6)-(9);
9:         obtain the probability distribution by Eq. (2);
10:    **end for**
11:    obtain training loss by Eq. (3)-(5);
12:    update parametres in ASL by Adam;
13: **end for**

**Output:** trained ASL.

$F^a$ along height and width dimension of $F_c$, and concatenate $F_c$ and $F^a$ to get hybrid feature $F' \in \mathbb{R}^{H \times W \times (C+A)}$. Finally, we feed $F'$ into PSAM to get the final-refined feature $F_f \in \mathbb{R}^{H \times W \times C}$. The overall attribute-visual attention process can be summarized as:

$$F_c = \mathcal{M}_c(F) \otimes F, \tag{6}$$

$$F_f = \mathcal{M}_s(F') \otimes F_c, \tag{7}$$

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.

**Channel attention.** To compute the channel-wise attention efficiently, we first use global average pooling and global max pooling to aggregate channel information and get $F^{avg} \in \mathbb{R}^{1 \times 1 \times (C+A)}$ and $F^{max} \in \mathbb{R}^{1 \times 1 \times (C+A)}$. Then, channel attention generating network (i.e., one layer MLP) is adopted to generate a channel-wise attention map $\mathcal{M}_c$. To summarize, the channel attention is computed as:

$$\mathcal{M}_c(F) = \sigma(\text{MLP}(F^{avg}) + \text{MLP}(F^{max})). \tag{8}$$

**Pyramid spatial attention.** Since different visual representation of attributes correspond to different sizes in images [23], it is necessary to design a pyramid spatial attention module. Similar to the channel attention module, 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 $K$ different 2D convolution kernels. Finally, we sum all spatial-wise attention maps to generate the final spatial attention map. The pyramid spatial attention can be described as:

$$\mathcal{M}_s(F') = \sigma(\sum_{i=1}^{K} \text{Conv}_{K_i}([\text{AvgPool}(F'); \text{MaxPool}(F')])), \tag{9}$$

where $\text{Conv}_{K_i}$ represents a 2D convolution operation with the filter size of $K_i \times K_i$.

## III. EXPERIMENTS

In this section, the effectiveness of ASL is verified by various experiments. Two standard few-shot classification datasets

### TABLE I
COMPARISON WITH OTHER STATE-OF-THE-ART METHODS WITH 95% CONFIDENCE INTERVALS ON CUB AND SUN.

| Model | Backbone | CUB | |
| --- | --- | --- | --- |
| | | 1-shot | 5-shot |
| Matching Nets [24] | Conv-64F | 61.16±0.89 | 72.86±0.70 |
| ProtoNets [12] | Conv-64F | 51.31±0.91 | 70.77±0.69 |
| Relation Nets [25] | Conv-64F | 62.45±0.98 | 76.11±0.69 |
| Comp. [17] | Conv-64F | 53.60±0.00 | 74.60±0.00 |
| CovaMNet [26] | Conv-64F | 60.58±0.69 | 74.24±0.68 |
| LRPABN [27] | Conv-64F | 67.97±0.44 | 78.26±0.22 |
| **ASL** | Conv-64F | **74.82±0.17** | **80.15±0.12** |
| ProtoNets [12] | ResNet-12 | 68.80±0.00 | 76.40±0.00 |
| Relation Nets [25] | ResNet-12 | 62.45±0.98 | 76.11±0.69 |
| AM3 [28] | ResNet-12 | 73.60±0.00 | 79.90±0.00 |
| FEAT [28] | ResNet-12 | 68.87±0.22 | 82.90±0.15 |
| Mul. ProtoNets [29] | ResNet-12 | 75.01±0.81 | 85.30±0.54 |
| DeepEMD [30] | ResNet-12 | 75.65±0.83 | 88.69±0.50 |
| AGAM [31] | ResNet-12 | 79.58±0.25 | 87.17±0.23 |
| Dual TriNet [22] | ResNet-18 | 69.61±0.46 | 84.10±0.35 |
| Baseline [32] | ResNet-18 | 65.51±0.87 | 82.85±0.55 |
| Baseline++ [32] | ResNet-18 | 67.02±0.90 | 83.58±0.54 |
| **ASL** | ResNet-12 | **82.12±0.14** | **89.65±0.11** |

| Model | Backbone | SUN | |
| --- | --- | --- | --- |
| | | 1-shot | 5-shot |
| MatchingNet [24] | Conv-64F | 55.72±0.40 | 76.59±0.21 |
| ProtoNets [12] | Conv-64F | 57.76±0.29 | 79.27±0.19 |
| Relation Nets [25] | Conv-64F | 49.58±0.35 | 76.21±0.19 |
| Comp. [17] | ResNet-10 | 45.90±0.00 | 67.10±0.00 |
| AM3 [28] | Conv-64F | 62.79±0.32 | 79.69±0.23 |
| AGAM [31] | Conv-64F | 65.15±0.31 | 80.08±0.21 |
| **ASL** | Conv-64F | **66.17±0.17** | **80.91±0.15** |

### TABLE II
AVERAGE ACCURACY COMPARISON BEFORE AND AFTER INCORPORATING ASL INTO EXISTING METHODS. BACKBONE: CONV-64F.

| Method | 5-way 1-shot | 5-way 5-shot |
| --- | --- | --- |
| Matching Nets [24] | 61.16±0.89 | 72.86±0.70 |
| **Matching Nets + ASL** | **69.13±0.18 (+7.97)** | **74.94±0.11 (+2.08)** |
| ProtoNets [12] | 51.31±0.91 | 70.77±0.69 |
| **ProtoNets + ASL** | **74.82±0.17 (+23.51)** | **80.15±0.12 (+9.38)** |
| Relation Nets [25] | 62.45±0.98 | 76.11±0.69 |
| **Relation Nets + ASL** | **67.02±0.19 (+4.57)** | **79.73±0.13 (+3.62)** |

CUB [33] and SUN [34] are selected to compare the performance of our approach with previous few-shot recognition methods.

### A. Experimental Settings

Following [12], [25], we use both shallow four layer convolutional Conv-64F [24] and ResNet-12 [35] as our backbone network. We train our model from scratch by Adam optimizer [36] with an initial learning rate $1 \times 10^{-3}$. For ASL, we set weighting factor $\alpha = 1.0$ in all experiments. For VAP, we only use the categories of training set for training. We train our model for 60,000 iterations. For both CUB and SUN, the images are resized to $84 \times 84$, and no data augmentations are adopted. During the test stage, we report the top-1 mean accuracy over 10,000 tasks. Note that all experiments in Sec. Ex

| TABLE III | | |
|-----------|---|---|
| ABLATION STUDY ON OUR MODEL. BACKBONE: RESNET-12. | | |
| Method | 5-way 1-shot | 5-way 5-shot |
| w/o VAP | 73.23±0.17 | 82.59±0.13 |
| w/o CAM | 69.86±0.19 | 86.64±0.13 |
| w/o PSAM | 79.21±0.17 | 87.87±0.13 |
| w/o VAP & AVAM | 69.60±0.19 | 79.83±0.23 |
| Not using attributes | 69.98±0.19 | 86.48±0.11 |
| Using all-0 attributes | 69.76±0.17 | 84.54±0.11 |
| ASL | **82.12±0.14** | **89.65±0.11** |

| TABLE IV | | |
|----------|---|---|
| INFLUENCE OF KERNEL COMBINATION. BACKBONE: RESNET-12. | | |
| Kernel Size | 5-way 1-shot | 5-way 5-shot |
| 3 | 80.65±0.16 | 86.61±0.12 |
| 5 | 80.49±0.15 | 87.05±0.12 |
| 7 | 78.30±0.17 | 86.69±0.11 |
| 9 | 76.37±0.17 | 86.03±0.12 |
| (3, 5, 7) | 80.78±0.17 | 87.99±0.13 |
| (5, 7, 9) | 80.89±0.18 | 88.33±0.10 |
| (3, 5, 7, 9) | **82.12±0.14** | **89.65±0.11** |

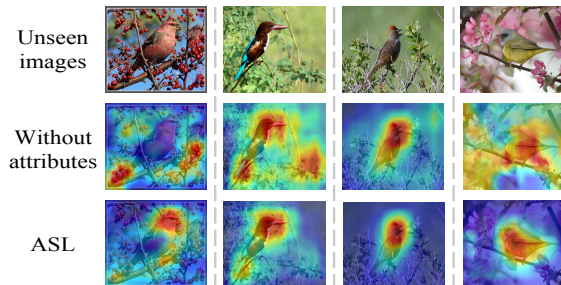| TABLE V | | |
|---------|---|---|
| INFLUENCE OF WEIGHTING FACTOR $\alpha$. BACKBONE: RESNET-12. | | |
| $\alpha$ | 5-way 1-shot | 5-way 5-shot |
| 0 | 76.93±0.15 | 83.60±0.11 |
| 0.001 | 77.85±0.14 | 85.91±0.12 |
| 0.01 | 78.25±0.18 | 86.61±0.13 |
| 0.1 | 80.24±0.18 | 87.69±0.12 |
| 0.5 | 80.45±0.16 | 87.15±0.13 |
| 1.0 | **82.12±0.14** | **89.65±0.11** |
| 2.0 | 79.01±0.17 | 86.62±0.12 |



Fig. 3. Grad-CAM visualization of four unseen images.

## B. Results and Analysis

TABLE I lists the recognition results of different methods on CUB and SUN datasets. From the experimental results, we have the following observations:

1) It can be observed that ASL achieves the best performance among all approaches. To be more specific, our model is around 19.2%/8.1% better than FEAT [28] on CUB with ResNet-12 under 1-shot and 5-shot settings.

2) Multi-modal based methods (e.g., Mul. ProtoNets, AM3, Dual TriNet and Comp.) generally outperform classic meta-learning based methods that rely on the single visual modality by a large margin, which validates the effectiveness of using auxiliary semantic information.

3) Among these methods, AM3, Dual TriNet, Comp., and AGAM only augment the representations of support classes, while query images have no semantic modalities information to enhance representations. ASL generally achieves better performance than these methods, demonstrating that generating and leveraging the attribute information of query images is beneficial to improve representation learning.

4) The performance improvement of the multi-modal based methods on CUB is greater than that on SUN. This may be because there are more attribute categories in CUB, which can better guide visual attention.

## C. Discussion

**Pluggability of ASL.** To verify the effectiveness of our proposed ASL, we embed it into three classic metric-based approaches. Table II shows the gains obtained by incorporating ASL into each approach on CUB, and for all three approaches, incorporating ASL leads to a significant improvement. These experimental results verify that our model is a flexible plug-and-play method.

**Framework Design.** As shown in Table III, each module in ASL has a significant contribution to performance improvement. 1) Removing the VAP drops the performance by 10.8%/7.9% on CUB under 1-shot/5-shot settings respectively, which attests that the attribute generation task brings extra useful information and significant improvement. 2) Removing either of the CAM and the PSAM leads to 32.0%/3.5% performance drops on the 1-shot setting. This demonstrates that using both attention modules simultaneously can captures more useful information. 3) To prove the effectiveness of our whole model ASL, we also remove both VAP and AVAM. Removing these two modules drops the performance by 15.2%/11.0% on 1-shot/5-shot settings respectively. 4) If we only use visual features, the improvement brought by pure visual attention mechanism is limited.

**Influence of kernel combination.** We select various kernels and their combinations to conduct experiments on CUB dataset. As shown in Table IV, it is better to use a combination of different kernels. A possible explanation is that different attribute features of objects correspond to different sizes in images, and using a single size cannot precisely locate attribute features.

**Influence of weighting factor $\alpha$.** $\alpha$ is a pre-defined parameter that influences the recognition performance. Experimental results on CUB dataset are shown in Table V. It can be seen that when we set $\alpha$ to 1, ASL can achieve the best performance.

**More discussions.** We further inserted AVAM after each ResBlock and achieve 83.42%/90.32% accuracy on CUB under 1-shot/5-shot settings with ResNet-12 respectively, indicating that using AVAM at different layers can provide more useful auxiliary information.

**Visualization.** To make an intuitive understanding of our ASL, Fig. 3 visualizes the Grad-CAM [37]. It can be seen that ASL improves the recognition performance by introducing additional attribute vectors to help the model focus on more representative local features.

## IV. CONCLUSION

In this letter, 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 illustrate the encouraging performance of our ASL, which has achieved competitive results with other state-of-the-art few-shot learning methods.

## REFERENCES

[1] Huaxiong Li, Chao Zhang, Xiuyi Jia, Yang Gao, and Chunlin Chen, "Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval," *IEEE Trans. Knowl. Data Eng.*, 2021, to be published, doi:10.1109/TKDE.2021.3102119.

[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] Tao Li, Zheng Zhang, Lishen Pei, and Yan Gan, "Hashformer: Vision transformer based deep hashing for image retrieval," *IEEE Signal Process. Lett.*, to be published, doi:10.1109/LSP.2022.3157517.

[4] Lei Zhong, Feng-Heng Li, Hao-Zhi Huang, Yong Zhang, Shao-Ping Lu, and Jue Wang, "Aesthetic-guided outward image cropping," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 211:1–211:13, 2021.

[5] Nikita Dvornik, Cordelia Schmid, and Julien Mairal, "Selecting relevant features from a multi-domain representation for few-shot classification," in *ECCV*, 2020, vol. 12355, pp. 769–786.

[6] Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, and Yang Gao, "Lofgan: Fusing local representations for few-shot image generation," in *ICCV*, 2021, pp. 8443–8451.

[7] Chunhao Cai, Minglei Yuan, and Tong Lu, "IFSM: an iterative feature selection mechanism for few-shot image classification," in *ICPR*, 2020, pp. 9429–9436.

[8] Debasmit Das, Sungrack Yun, and Fatih Porikli, "ConfeSS: A framework for single source cross-domain few-shot learning," in *ICLR*, 2022.

[9] 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.

[10] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel, "Meta-learning for semi-supervised few-shot classification," in *ICLR*, 2018.

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

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

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

[14] Wen Fu, Li Zhou, and Jie Chen, "Bidirectional matching prototypical network for few-shot image classification," *IEEE Signal Process. Lett.*, to be published, doi:10.1109/LSP.2022.3152686.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, vol. 139, pp. 8748–8763.

[16] Xingyu Chen, Jin Li, Xuguang Lan, and Nanning Zheng, "Generalized zero-shot learning via multi-modal aggregated posterior aligning neural network," *IEEE Trans. Multim.*, vol. 24, pp. 177–187, 2022.

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

[18] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You, "Transzero: Attribute-guided transformer for zero-shot learning," in *AAAI*, 2021.

[19] Min Meng, Jie Wei, and Jungang Wu, "Learning multi-part attention neural network for zero-shot classification," *IEEE Trans. Cogn. Dev. Syst.*, vol. PP, no. 99, pp. 1–1, 2020.

[20] Min Meng and Jun Yu, "Zero-shot learning via robust latent representation and manifold regularization," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1824–1836, 2019.

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

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

[23] Yifeng Ding, Zhanyu Ma, Shaoguo Wen, Jiyang Xie, Dongliang Chang, Zhongwei Si, Ming Wu, and Haibin Ling, "AP-CNN: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2826–2836, 2021.

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

[25] 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.

[26] Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao, and Jiebo Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *AAAI*, 2019, pp. 8642–8649.

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

[28] 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.

[29] Frederik Pahde, Mihai Marian Puscas, Tassilo Klein, and Moin Nabi, "Multimodal prototypical networks for few-shot learning," in *WACV*, 2021, pp. 2643–2652.

[30] 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.

[31] 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.

[32] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang, "A closer look at few-shot classification," in *ICLR*, 2019.

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

[34] Genevieve Patterson, Chen Xu, Hang Su, and James Hays, "The SUN attribute database: Beyond categories for deeper scene understanding," *Int. J. Comput. Vis.*, vol. 108, no. 1-2, pp. 59–81, 2014.

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

[36] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[37] 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.