Global and Local Feature Interaction with Vision Transformer for Few-shot Image Classification

Mingze Sun
Department of Computer Science and Technology (DCS&T), Tsinghua University
Beijing, China
sunmingze02@gmail.com

Weizhi Ma*
Institute for AI Industry Research (AIR), Tsinghua University
Beijing, China
mawz@tsinghua.edu.cn

Yang Liu
DCS&T, Tsinghua University
AIR, Tsinghua University
Beijing, China
liuyang2011@tsinghua.edu.cn

ABSTRACT

Image classification is a classical machine learning task and has been widely used. Due to the high costs of annotation and data collection in real scenarios, few-shot learning has become a vital technique to improve image classification performances. However, most existing few-shot image classification methods only focus on modeling the global image feature or image local patches, which ignore the global-local interactions. In this study, we propose a new method, named GL-ViT, to integrate both global and local features to fully exploit the few-shot samples for image classification. Firstly, we design a feature extractor module to calculate the interactions between the global representation and local patch embeddings, where ViT is also adopted to achieve efficient and effective image representation. Then, Earth Mover's Distance is adopted to measure the similarity between two images. Abundant Experimental results on several widely-used open datasets show that GL-ViT outperforms state-of-the-art algorithms significantly, and our ablation studies also verify the effectiveness of both global-local features.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Few-shot Learning, Image Classification, Visual Transformer.

ACM Reference Format:
Mingze Sun, Weizhi Ma, and Yang Liu. 2022. Global and Local Feature Interaction with Vision Transformer for Few-shot Image Classification. In *Proceedings of the 31st ACM Int'l Conference on Information and Knowledge Management (CIKM '22), Oct. 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557604

1 INTRODUCTION

Recent years have witnessed rapid progress in deep learning, especially image classification. With an adequate amount of labeled data, numerous impressive methods have been proposed and achieved good performances [11, 22, 25]. However, in many practical scenarios, it is too costly, and sometimes even impossible to collect enough annotated data for training. In contrast, humans can classify items in a new class with only limited examples. Thus, many efforts have been made to enhance image classification methods with such a few-shot learning ability [1, 4, 5, 9, 15–17, 19, 24].

The main challenge in few-shot image classification is how to measure the similarity between the candidate image and labeled images. Most of the previous studies adopt a two-module strategy, namely the feature extractor module and similarity calculation module. And there are mainly two types of models: 1) Global-feature-based. These models generate a global feature vector for each image and then calculate similarities [4, 5, 16, 17]. 2) Local-feature-based. Instead of using only one feature vector, these methods split each image into several patches, and measure the similarity between two images based on patch-level feature interactions [1, 19, 24].

However, we find that most of previous studies fail to fully exploit global and local features of images, and especially ignore interactions between them. Global-feature-based methods focus on modeling global information, but are weak at modeling the fine-grained similarity between images. While local-feature-based methods, instead, may pay too much attention to calculating local features and fail to achieve optimal modeling of the whole image. Moreover, current works rely on using convolutional neural networks (CNNs) as their feature extractors [11, 22], which are unable to generate both global and local features at the same time.

To cope with these challenges, we propose a new algorithm, named GL-ViT, to model the global and local feature interactions for better few-shot classification performances. Firstly, vision transformer (ViT) [8], a pre-training model which can extract both global features and local features for images simultaneously, is applied as our backbone, and the ViT is fine-tuned in few-shot classification tasks. Then, we use a simple yet effective strategy to model the interactions between them, and adopt Earth Mover's Distance (EMD) to measure the similarity due to its outstanding performance in previous studies [24]. Our main contributions are as follows:

- To the best of our knowledge, our study first uses global & local feature interactions for few-shot image classifications;
- We design a new method, named GL-ViT, with the vision transformer to achieve efficient global & local feature generation and adopt EMD for similarity calculation;
- Experimental results on two public datasets show impressive improvements over the state-of-the-art methods. And ablation studies also verify the effectiveness of each module.

*Corresponding Author.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557604

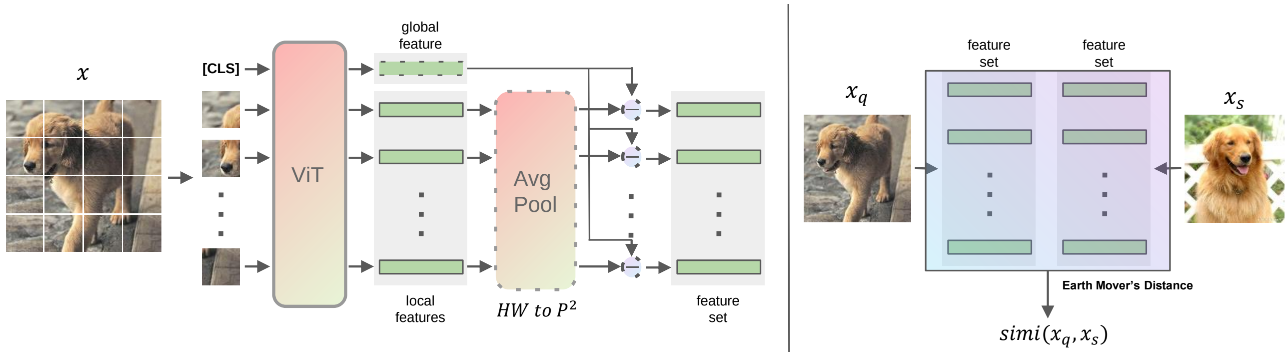


Figure 1: The left part shows the feature extractor module (generate the feature set of each image), and the right part shows the similarity calculation module (how to calculate similarities between images).

2 RELATED WORKS

2.1 Few-shot Image Classification

There are mainly two categories of few-shot image classification methods, optimization-based and metric-based. Optimization-based methods [9, 15] focus on efficient model training, i.e., rapidly adapting parameters to novel classes and avoiding over-fitting for classical image classification methods. While most recent studies are metric-based methods [1, 4, 5, 16, 17, 19, 24], which aims at learning a class-agnostic feature extractor, and measuring similarities between query images and support images using extracted features.

Metric-based methods can be further classified into global-feature-based ones and local-feature-based ones. Global-feature-based methods extract only one feature vector from an image [4, 5]. Matching Network [17] and Prototypical Network [16] propose an episodic paradigm for metric learning, where support features of each episode are averaged as a prototype for each class, and Euclidean distance between each query feature vector and class prototypes is regarded as logits for label prediction. Local-feature-based methods extract a feature set from an image, each of which captures a patch of the image [1, 19]. Zhang et al. [24] adopt the aforementioned episodic learning, and take the image similarity calculation as an Optimal Transport problem with EMD method [24].

In global-feature-based methods, the global feature vector captures the semantic information of the whole image, but features irrelevant to the classification task may be included. The feature vector is biased away from the cluster center, which may catastrophically impact the performance. While local-feature-based methods models patch-level information that are short of interaction with each other. Thus they may fail to capture the high-level features.

2.2 Backbone for Image Feature Extraction

The majority of few-shot learning studies adopts convolutional neural networks (CNNs) as their backbone for feature extractor [1, 4, 16, 17, 19, 20, 24]. Typically, backbones for common image classification are widely used after some modification, e.g. ResNet [11] or WRN [22]. Global-feature-based methods remove the fully-connected (FC) layers from them, while local-feature-based methods replace the FC layers with convolutional layers in the manner of fully convolutional networks (FCN). Although they have achieved competitive performance, CNN-based backbones fail to yield global features and local features simultaneously, which limits the exploration of the interaction between global and local features.

Recently studies show that self-supervised vision transformer can resolve the aforementioned difficulty. Vision Transformer (ViT) [8] adopts Transformer and achieve impressive performance in many computer vision tasks. It first crops the image into patches of fixed size, flattens them into an 1D sequence after linearly embedded, and adds a preceding classification token ($[CLS]$). Then the sequence is fed into an encoder composed of self-attention blocks. The final output of $[CLS]$ is regarded as the global feature vector, typically used for downstream tasks. Simultaneously, the outputs of other tokens are the local features. Various self-supervised pre-train tasks have been proposed for ViT and applied in downstream tasks [2, 6, 10]. Self-supervised ViT has been adopted in recent works on few-shot learning and dramatically outperformed CNN-based methods [5]. So we also adopt ViT as our backbone.

3 METHOD

In this section, we first introduce some preliminaries. Then, we will describe how to extract global & local features from a single image and conduct feature interactions. Further, we present how to measure the similarity between two images and the loss function. Our framework is shown in Fig. 1 [1].

3.1 Preliminaries

Few-shot image classification follows the N-way K-shot settings, namely classifying on N classes with K example images in each class. Neural models are first trained on $D_{base} = \{(x_i, y_i)|y_i \in C_{base}\}$, where x_i is an image and y_i is the corresponding label. Then, in testing phase, it predicts the classes of unlabeled query set $Q_{novel} = \{x_i\}$ given a labeled support set $S_{novel} = \{(x_i, y_i)|y_i \in C_{novel}, C_{base} \cap C_{novel} = \emptyset\}$, where S_{novel} has N classes and K images per class and the classes of Q_{novel} are identical to S_{novel}.

In training phase, following previous study [16, 17], we adopt a meta-learning paradigm named episodic learning. Simulating the testing phase, in each episode, we sample N classes from C_{base}, K images per class as support set S_{base}, and multiple images as the

[1]Our code is available at: https://github.com/waltsun/GL-ViT.

query set Q_{base} from D_{base}, and then update the model using the prediction on Q_{base}. In episodic learning, the pipelines in the training phase and testing phase are highly similar, which minimizes task settings difference between the two phases.

3.2 Feature Extractor Module

3.2.1 Feature Generation. To yield global and local features simultaneously, we use self-supervised vision transformers (ViT) [8], an pretraining backbone in computer vision, as our feature extractor. We use not only $[CLS]$ token output vector as a global feature capturing the whole picture, but also patch-level outputs vectors as local features. Extracted features are denoted as:

$$[f^{global}; f_1^{local}; f_2^{local}; \cdots ; f_{HW}^{local}] = ViT(\cdot) \tag{1}$$

where HW is the patch number.

Considering HW is typically large, we concentrate local features to a smaller number. Local features are weighted average pooled into (P, P) numbers, the self-attention weights of $[CLS]$ token on other ones in the last layer are used as average weight (the attention weight on $[CLS]$ token itself is deprecated here). The attention weights are also average pooled for subsequent use:

$$[\alpha^{global}; \alpha_1^{local}; \alpha_2^{local}; \cdots ; \alpha_{HW}^{local}] = LastAttn(\cdot) \tag{2}$$

$$\hat{f}_i^{local} = \alpha_i^{local} * f_i^{local}, \quad i = 1, 2, \cdots, HW \tag{3}$$

$$\begin{bmatrix} \widetilde{f}_1^{local} & \cdots & \widetilde{f}_P^{local} \\ \vdots & \ddots & \vdots \\ \widetilde{f}_{P^2-P+1}^{local} & \cdots & \widetilde{f}_{P^2}^{local} \end{bmatrix} = AvgPool(\begin{bmatrix} \hat{f}_1^{local} & \cdots & \hat{f}_W^{local} \\ \vdots & \ddots & \vdots \\ \hat{f}_{HW-W+1}^{local} & \cdots & \hat{f}_{HW}^{local} \end{bmatrix}) \tag{4}$$

$$\begin{bmatrix} \widetilde{\alpha}_1 & \cdots & \widetilde{\alpha}_P \\ \vdots & \ddots & \vdots \\ \widetilde{\alpha}_{P^2-P+1} & \cdots & \widetilde{\alpha}_{P^2} \end{bmatrix} = AvgPool(\begin{bmatrix} \alpha_1^{local} & \cdots & \alpha_W^{local} \\ \vdots & \ddots & \vdots \\ \alpha_{HW-W+1}^{local} & \cdots & \alpha_{HW}^{local} \end{bmatrix}) \tag{5}$$

3.2.2 Feature Interactions. For a single image, we have a global feature vector capturing the semantic information of the whole image, and local feature vectors capturing the semantic information of respective patches. We aim to make them interact and integrate the information of two different grains. We choose a simple yet effective method to achieve that:

$$\widetilde{f}_i = \widetilde{f}_i^{local} - \alpha^{global} * f^{global}, \quad i = 1, 2, \cdots, P^2 \tag{6}$$

3.3 Similarity Calculating Module

With a feature set for each image, we need to calculate the similarity between two feature sets of images. We adopt EMD due to its outstanding performance in previous studies [24]. EMD is a algorithm for the Optimal Transport problem, which assumes that n source depots need to transport goods to m target depots. Given the cost of transportation per unit of goods $c_{i,j}$ between two depots of i and j, the amount of supplied goods from each source depot s_i, and the amount of needed goods from each target depot t_j, it can measure the minimal cost of transporting all goods and the corresponding goods flow. Mathematically, the problem can be described as:

$$\begin{aligned} \widetilde{flow} = \underset{flow}{\arg\min} \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} c_{i,j} \cdot flow_{i,j} \\ \text{s.t.} \quad & flow_{i,j} \geq 0, \quad i = 1, \cdots, n, \quad j = 1, \cdots, m \\ & \sum_{j=1}^{m} flow_{i,j} = s_i, \quad i = 1, \cdots, n \\ & \sum_{i=1}^{n} flow_{i,j} = t_j, \quad j = 1, \cdots, m \end{aligned} \tag{7}$$

Costly depot pairs are always assigned with no or few good flow, which in this context, avoids irrelevant visual patches disturbing similarity calculation and focuses on semantic information relevant to the class labels. Thus, we adopt EMD as our feature set matching function. In our settings, cosine similarity is used to measure unit cost and the attention weight for the transported amount from or to a depot. Using Equation 7, the similarity between two images x_1, x_2 can be determined (here $n = m = P^2$):

$$S_{i,j} = cosine(\widetilde{f}_i^{x_1}, \widetilde{f}_j^{x_2}), \quad i = 1, \cdots, n, \quad j = 1, \cdots, m \tag{8}$$

$$c_{i,j} = 1 - S_{i,j}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, m \tag{9}$$

$$s_i = \widetilde{\alpha}_i^{x_1}, \quad i = 1, \cdots, n \tag{10}$$

$$t_j = \widetilde{\alpha}_j^{x_2}, \quad j = 1, \cdots, m \tag{11}$$

$$simi(\cdot, \cdot) = \sum_{i=1}^{n}\sum_{j=1}^{m} S_{i,j} \cdot \widetilde{flow}_{i,j}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, m \tag{12}$$

3.4 Loss Function

After introducing how to measure the similarity between two images, and we need to update our model with such metric. Fuse Score [19] is adopted to calculate probability distribution over N classes. Similarity scores with the support images are averaged after softmax for each class. Negative log-likelihood loss is used.

$$p_i^{x_q} = \frac{\sum_{j=1}^{K} \exp(simi(x_q, x_{s,j}^{c_i}))/K}{\sum_{l=1}^{N}\sum_{j=1}^{K} \exp(simi(x_q, x_{s,j}^{c_l}))}, i = 1, 2, \cdots, N \tag{13}$$

$$loss = \sum_{x_q}\sum_{i=1}^{N} \mathbb{I}(y_q = c_i) \cdot \log(p_i^{x_q}) \tag{14}$$

In the case of $K = 1$, the procedure described above is the same with cross entropy loss.

4 EXPERIMENTS

4.1 Experimental Settings

Experiments are conducted on two few-shot image classification datasets, mini-ImageNet [17] and Caltech-UCSD Birds-200-2011 (CUB) [18]. Mini-ImageNet is a subset of ImageNet [14], which is a popular benchmark in few-shot image classification and has 100 classes with 600 images in each class. CUB, a bird classification dataset, contains 11,788 images from 200 classes.

We adopt vision transformer with patch resolution of 16×16 [8] as our backbone, and the parameters are self-supervised pre-trained in DINO framework [6]. The implement details and hyper-parameters

Methods	Backbone	Feature Type	Self-supervised	mini-ImageNet		CUB	
				5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
SM-112 [19]	ResNet-12	Local	No	69.48 ± 0.46	84.51 ± 0.30	84.11 ± 0.39	93.62 ± 0.19
DeepEMD [24]	ResNet-12	Local	No	65.91 ± 0.82	82.41 ± 0.56	75.65 ± 0.83	88.69 ± 0.50
Sum-min [1]	SF-12	Local	No	68.32 ± 0.62	82.71 ± 0.46	79.60 ± 0.80	90.48 ± 0.44
DeepEMD v2 [23]	ResNet-12	Local	Yes	68.77 ± 0.29	84.13 ± 0.53	79.27 ± 0.29	89.80 ± 0.51
Prototypical Net [16]	Conv-4	Global	No	49.42 ± 0.78	68.20 ± 0.66	-	-
Prototypical Net [16]	ResNet-12	Global	No	60.37 ± 0.83	78.02 ± 0.57	66.09 ± 0.92	82.50 ± 0.58
Distribution Calibration [21]	WRN-28-10	Global	No	68.57 ± 0.55	82.88 ± 0.42	79.56 ± 0.87	90.67 ± 0.35
EASY [4]	3 × ResNet-12	Global	Yes	71.75 ± 0.19	87.15 ± 0.12	78.56 ± 0.19	91.93 ± 0.10
Image900-SSL [7]	AmdimNet	Global	Yes	76.82 ± 0.19	90.98 ± 0.10	77.09 ± 0.21	89.18 ± 0.13
Simple CNAPS [3]	ResNet-18	Global	Yes	82.16	89.80	-	-
HCTransformer [12]	ViT-S/8	Global	Yes	74.62 ± 0.20	89.19 ± 0.13	-	-
SSL-ViT-16 [5]	ViT-S/16	Global	Yes	86.50 ± 0.17	96.22 ± 0.06	89.94 ± 0.15	96.98 ± 0.05
GL-ViT(Ours)	ViT-S/16	Global+Local	Yes	**88.04 ± 0.59**	**96.45 ± 0.20**	**92.81 ± 0.53**	**97.80 ± 0.23**

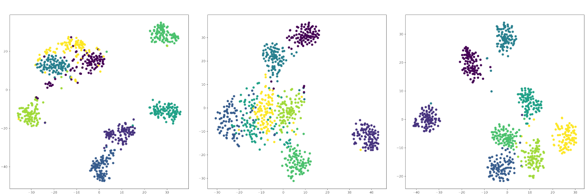
Table 1: 5-way 1-shot and 5-way 5-shot classification accuracy (%) with 95% confidence intervals on mini-ImageNet and CUB. Due to the experimental settings are the same as SSL-ViT-16 [5], some reported experimental results are adopted.

No. Patch	Accuracy
2 × 2	86.31 ± 0.64
3 × 3	87.40 ± 0.68
4 × 4	**88.04 ± 0.59**
5 × 5	87.32 ± 0.61
6 × 6	86.47 ± 0.67

Table 2: 5-way 1-shot accuracy on mini-ImageNet in different patch number.

Model	Accuracy
Local-Only	81.21 ± 0.69
Global-Only	86.15 ± 0.61
Global-Local (+)	87.87 ± 0.62
Global-Local (-)	**88.04 ± 0.59**

Table 3: 5-way 1-shot accuracy on mini-ImageNet with different grained feature.



(a) Global Only **(b) Local Only** **(c) Global&Local**

Figure 2: Visualizing feature vectors using t-SNE. 800 images are sampled from 8 classes in mini-ImageNet. Points are colored according to different classes.

of vision transformer follow ViT-small from [5]. Output local features are pooled into the number of 4×4 ($P = 4$). Besides, we use accuracy as the evaluation metric as previous studies [16, 19, 20, 24].

4.2 Analysis of Experimental Results

Main Result. The overall performances are reported in Table 1. Firstly, our method GL-ViT outperforms all baseline methods, especially in the 5-way 1-shot scenario. The results verify the effectiveness of GL-ViT. Secondly, we can see methods with ViT backbone perform better than models with CNN-based backbones, which showing the usefulness of the pre-training strategy. Thirdly, most global-feature based methods are stronger than local-feature based, and global+local feature based method GL-ViT achieves the best.

Hyper-parameter study on patch number. To find the best patch number for GL-ViT, we conduct experiments on 5-way 1-shot mini-ImageNet with patch number from 2×2 to 6×6. Due to the limit of space, we only show the 5-way 1-shot accuracy results on mini-ImageNet in Table 2. Results show that too fewer or too many patches result in worse performances, and 4×4 is the best.

Ablation Studies. To verify the effectiveness of the multi-grained feature interaction module, we compare our method with local-feature-only variant and global-feature-only variant. Moreover, we change our interaction approach from subtraction to addition. Experimental results are shown in Table 3, which demonstrate multi-grained feature interaction consistently outperform local-only and global-only variants.

Visualization on feature vectors. To intuitively illustrate the effectiveness of our method, we visualize the feature vectors by t-SNE [13]. We sample 800 images on 8 different classes from mini-ImageNet, and get respective outputs in global&local, global-only, and local-only settings. The feature set in global&local or local-only is averaged to a single vector. The vectors are visualized in Fig. 2, with different classes marked with different colors. The distribution of our Global&Local method is more reasonable than the other two.

5 CONCLUSIONS

In this work, we point out the weaknesses of methods that use only global-feature or local-feature. We propose a few-shot image classification method, named GL-ViT, with multi-grained feature interaction and visual transformer backbone. The feature interaction part of our method is simple and effective. Experimental results on several datasets shows that GL-VIT outperforms all SOTA baseline methods, and further analysis verified the effectiveness of our method. In the future, we plan to further explore more effectiveness feature interaction modules for few-shot image classification.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (No. 6200219) and Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

[1] Arman Afrasiyabi, Hugo Larochelle, Jean-François Lalonde, and Christian Gagné. 2022. Matching Feature Sets for Few-Shot Image Classification. *CoRR* abs/2204.00949 (2022). https://doi.org/10.48550/arXiv.2204.00949 arXiv:2204.00949

[2] Hangbo Bao, Li Dong, and Furu Wei. 2021. BEiT: BERT Pre-Training of Image Transformers. *CoRR* abs/2106.08254 (2021). arXiv:2106.08254 https://arxiv.org/abs/2106.08254

[3] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. 2020. Improved Few-Shot Visual Classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 14481–14490. https://doi.org/10.1109/CVPR42600.2020.01450

[4] Yassir Bendou, Yuqing Hu, Raphaël Lafargue, Giulia Lioi, Bastien Pasdeloup, Stéphane Pateux, and Vincent Gripon. 2022. EASY: Ensemble Augmented-Shot Y-shaped Learning: State-Of-The-Art Few-Shot Classification with Simple Ingredients. *CoRR* abs/2201.09699 (2022). arXiv:2201.09699 https://arxiv.org/abs/2201.09699

[5] Prarthana Bhattacharyya, Chenge Li, Xiaonan Zhao, István Fehérvári, and Jason Sun. 2022. Visual Representation Learning with Self-Supervised Attention for Low-Label High-data Regime. *CoRR* abs/2201.08951 (2022). arXiv:2201.08951 https://arxiv.org/abs/2201.08951

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 9630–9640. https://doi.org/10.1109/ICCV48922.2021.00951

[7] Da Chen, Yuefeng Chen, Yuhong Li, Feng Mao, Yuan He, and Hui Xue. 2021. Self-Supervised Learning for Few-Shot Image Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 1745–1749. https://doi.org/10.1109/ICASSP39728.2021.9413783

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1126–1135. http://proceedings.mlr.press/v70/finn17a.html

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. *CoRR* abs/2111.06377 (2021). arXiv:2111.06377 https://arxiv.org/abs/2111.06377

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[12] Yangji He, Weihan Liang, Dongyang Zhao, Hong-Yu Zhou, Weifeng Ge, Yizhou Yu, and Wenqiang Zhang. 2022. Attribute Surrogates Learning and Spectral Tokens Pooling in Transformers for Few-shot Learning. *CoRR* abs/2203.09064 (2022). https://doi.org/10.48550/arXiv.2203.09064 arXiv:2203.09064

[13] Laurens van der Maaten. 2008. Visualizing datausing t-sne. *Journal of machine learning research* 9 (2008), 2579–2605.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[15] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2019. Meta-Learning with Latent Embedding Optimization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=BJgklhAcK7

[16] Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4077–4087. https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html

[17] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 3630–3638. https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html

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

[19] Duo Wang, Qianxia Ma, Qingyuan Zheng, Yu Cheng, and Tao Zhang. 2022. Improved local-feature-based few-shot learning with Sinkhorn metrics. *Int. J. Mach. Learn. Cybern.* 13, 4 (2022), 1099–1114. https://doi.org/10.1007/s13042-021-01437-y

[20] Wanqi Xue and Wei Wang. 2020. One-Shot Image Classification by Learning to Restore Prototypes. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 6558–6565. https://ojs.aaai.org/index.php/AAAI/article/view/6130

[21] Shuo Yang, Lu Liu, and Min Xu. 2021. Free Lunch for Few-shot Learning: Distribution Calibration. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=JWOiYxMG92s

[22] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide Residual Networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith (Eds.). BMVA Press. http://www.bmva.org/bmvc/2016/papers/paper087/index.html

[23] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. 2020. Deepemd: Differentiable earth mover's distance for few-shot learning. *arXiv preprint arXiv:2003.06777* (2020).

[24] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. 2020. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 12200–12210. https://doi.org/10.1109/CVPR42600.2020.01222

[25] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander J. Smola. 2020. ResNeSt: Split-Attention Networks. *CoRR* abs/2004.08955 (2020). arXiv:2004.08955 https://arxiv.org/abs/2004.08955