A NEW TRANSDUCTIVE FRAMEWORK FOR FEW-SHOT BIOACOUSTIC EVENT DETECTION TASK

Technical Report

Yizhou Tan [1], Lifan Xu [2], Chenyang Zhu [3], Shengchen Li[2], Haojun Ai [1], Xi Shao [4]

[1] Wuhan University, School of Cyber Science and Engineering, Wuhan, China,
[2] Xi'an Jiaotong-Liverpool University, Department of Intelligent Science School of Advanced Engineering, Suzhou, China,
[3] Jiangnan University, School of Artificial Intelligence and Computer Science,Wuxi, China
[4] Nanjing University of Posts and Telecommunications,School of Communication and Information Engineering, Nanjing, China
Yizhou.Tan@ieee.org, Lifan.Xu17@student.xjtlu.edu.cn, chenyangzhu2018@163.com, Shengchen.Li@xjtlu.edu.cn, aihj@whu.edu.cn, shaoxi@njupt.edu.cn

ABSTRACT

Few-shot learning is introduced to reduce the requirements of data availability in machine learning, especially when the labelling is labour expensive. Few-shot learning algorithms usually suffer from the extraordinary feature distribution of the query class, especially in few-shot bioacoustic event detection task. In this work, Knowledge transfer technique is introduced into the transductive inference process to restrict the feature distribution of newly appeared class to a dedicated sub-space, while adapts the feature distribution for existing classes. The proposed system outperforms the traditional few-shot learning system according to the development dataset provided by bioacoustics event detection (Task 5) in DCASE data challenge 2022. The f-measure score of the validation in development dataset successfully reaches 57.40.

Index Terms— Few-shot Learning, Sound Event Detection

1. INTRODUCTION

As the development of deep learning, neural networks have been applied to solve various problem in reality. Due to the limited manual labour in data annotation, Few-shot learning [1, 2] has become a promising paradigm to satisfy the task needs under the limited labelled dataset, such as few-shot bioacoustic event detection task. Few-shot learning is expected the model to acquire the ability to predict the class of queried sample from few labelled samples. Common few-shot learning task can be represented as K-way N-shot task which means there are K and N labelled samples for each class provided in support set. The K classes are usually newly appeared to the classes from training set and the prediction of query set are limited to these K classes. Few-shot bioacoustic event detection task can be considered as a unique One Class Classification (OOC) task, which provides few labelled positive samples and several labelled negative noise samples.

The main barriers of few-shot learning are how to avoid the extraordinary feature distribution of newly appeared classes and the feature bias of support set caused by sparse and limited data. The extraordinary feature distribution means that features of these newly appeared classes samples are hard to follow a stable and predictable distribution, as the classes of support set are brand new to the trained model. Especially in few-shot bioacoustic event detection task, the negative noise samples do not have the stable natural pattern which lead to a worse extraordinary feature distribution. At the same time, the few labelled samples in support further increase the bias between the posterior distribution and real distribution of these newly appeared classes. Prior methods based on inductive inference methods [3–9] are expected to generalise a robust model through some training design, such as meta learning. Unfortunately, since inductive inference methods are limited to scope of training dataset, they can not well solve the two problems above. Recently, transductive inference methods [10–13] are proposed to take the prediction of all queried samples as an integral process instead of one sample at a time in inductive inference. This idea can utilize the extra data in query set to modify the original model to optimize the posterior distribution of newly appeared classes through some regulariers. However, existing transductive inference methods still facing the fitting problem of the extraordinary feature distribution of newly appeared classes. Large-scale parameters updating [11] during transductive inference tend to fall into the overfitting problem, while limited parameter updating [10] or graph clustering methods [12] are hard to well fit the extraordinary feature distribution of newly appeared classes.

In this work, the knowledge transfer idea is introduced to solve the extraordinary feature distribution problem in few-shot learning. Knowledge transfer idea is aimed at transferring extra information into a new task to refine the model. The proposed knowledge transfer method construct a dedicate feature sub-space to restrict the feature distribution of newly appeared classes through transferring the pre-trained classes distributions into transductive inference process as the anchor points to preserve the pre-trained knowledge.

We use the Prototype Network [4] as the pre-trained model and construct a new task adaptive feature extractor to replace part of neural layers in Prototype Network for prediction of each few-shot task. During the re-training process, we load the prototypes of training dataset as the prior knowledge of original space, and encourage the output of task adaptive feature extractor to keep the same weight of original model in order to align the original feature space of the pre-trained model and our new task adaptive feature extractor. The

parameters updating of task adaptive feature extractor will tend to preserve the original knowledge from pre-trained model and avoid the overfit problem. Furthermore, we also introduce the regularier of maximizing mutual information [10] and cross-entropy classification loss to optimize the task adaptive feature extractor. The experiments are conducted on the development dataset of DCASE 2022 Task5. Each audio file is considered as an individual few-shot bioacoustic event detection task. The general result of f-measure (57.40) is in line with our expectations which shows a significant improvement on prior methods as well as the official baseline.

Section 2 is the motivation of our design and the problems of other transductive inference methods. Section 3 is the general introduction of the proposed method. Section 4 contains all the details and results of our experiments.

2. MOTIVATION

The applications of Few-shot Bioacoustic Event Detection tend to deal with one specific biology detection as the positive class, while all other irrelevant acoustic event and noise are considered as negative class. This one class classification scenario bring a huge barrier to few-shot learning. Labelled negative samples in support set can not be guaranteed to belong to the same acoustic event, and even the negative samples in support set and query set are belong to an identical data distribution. The inconsistent data distribution between support set and query set make it hard to generalize the model from the support set to query set through meta-learning methods in inductive inference.

In the premise of accuracy first, the introduction of transductive inference is necessary to enable model to adapt different few-shot tasks. However, existing transductive inference methods still can not well solve the one class classification in few-shot learning. The mainstream idea of transductive inference is use a regularizer to update the model in a limited latent space. Fine-tuning the whole pre-trained network [11] is a well performance method as a baseline in transductive learning although it requires a long run-time. The balance sample amount of each class in support set and the significant feature pattern of different classes are the keys to prevents the fine-tuning method from overfit problem. Both of them are not satisfied in our task so that the pre-trained network is fragile and will happily overfit during fine-tuning. TIM [10] has also found that the updating of whole pre-trained network may easily go overfit and drop the model performance, which leads TIM [10] only to update the classifier by maximizing the mutual information as the regularizer. Although updating the linear classifier prevents the overfit problem during transductive inference, the improvement of these methods in our task is limited by the linear classifier. Due to the diversity of negative samples, the boundary between positive and negative class tend to be a non-linear plane as the left sides showing in figure 1. The the upper bound of accuracy of one class classification is limited to the feature distribution of the query set, which has been determined by the feature extractor of pre-trained model. Based on the above issues, we propose a new transductive inference method to construct a new task adaptive feature extractor that can effectively solve the overfit problem. We utilize the pre-trained knowledge to align the feature space of new task adaptive feature extractor and original pre-trained feature extractor, which successfully prevent the overfit problem due to the limitation of the latent space of the task adaptive feature extractor updating. As a result, our model only need to update the non-linear task adaptive feature extractor to mapping the queried samples into the proper positions

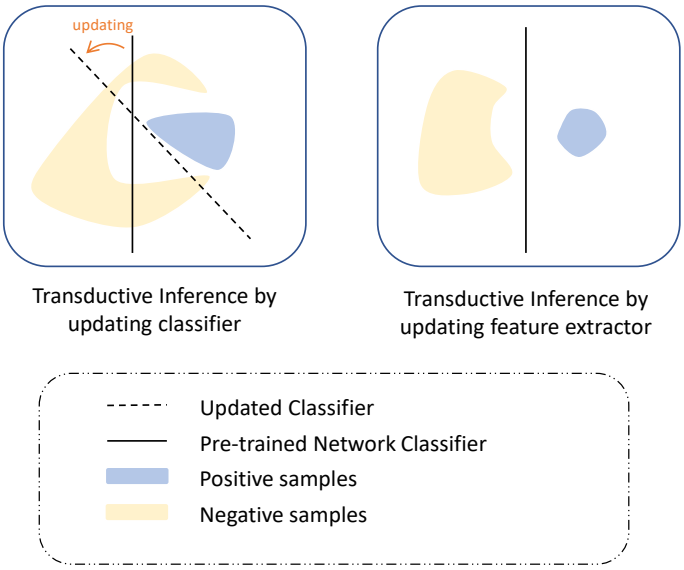

Figure 1: The feature space of different methods

in feature space instead of adjusting the linear classifier, as the right side showing in figure 1.

3. METHOD

3.1. Few-shot Scenario

Starting from the common few-shot setting, the training dataset $X_{train} = \{(x_i, y_i)|y_i \in Y_{train}\}$ is a large scale of labelled dataset, where x_i is a audio clip, y_i is the corresponding event category and Y_{train} is set of all categories in training dataset. Through splitting the whole audio files into several clips, the event detection task can be convert to a classification task. The testing set is consisting of a support set $X_s = \{(x_i, y_i)|y_i \in Y_s\}$ and a query set $X_q = \{x_i\}$, where Y_s is the set of categories in testing set. It is worth to note that the categories of training set and testing set do not overlap with each other, $Y_{train} \cap Y_s = \emptyset$. In common few-shot learning scenario, the support set is randomly sampled from the whole testing set and only contains few samples with labels acting as reference to the prediction of unlabelled query set. Specifically, a few-shot task can be referred to K-way N-shot task indicating that the support set consists of N labelled samples randomly sampled from each of K classes. The data amount of support set is $N \cdot K$ for a K-way N-shot task. Few-shot learning techniques will use the training set X_{train} to training a model with can introduce support set into prediction process to adapt the query set at hand.

Different from common few-shot setting, in our Few-shot Bioacoustic Event Detection task, the support set X_s consists of 5-shot positive samples, which is the target bioacoustic event audio clips, and several negative samples. In another word, this scenario can be viewed as a 1-way 5-shot one class classification task with several negative samples in support set.

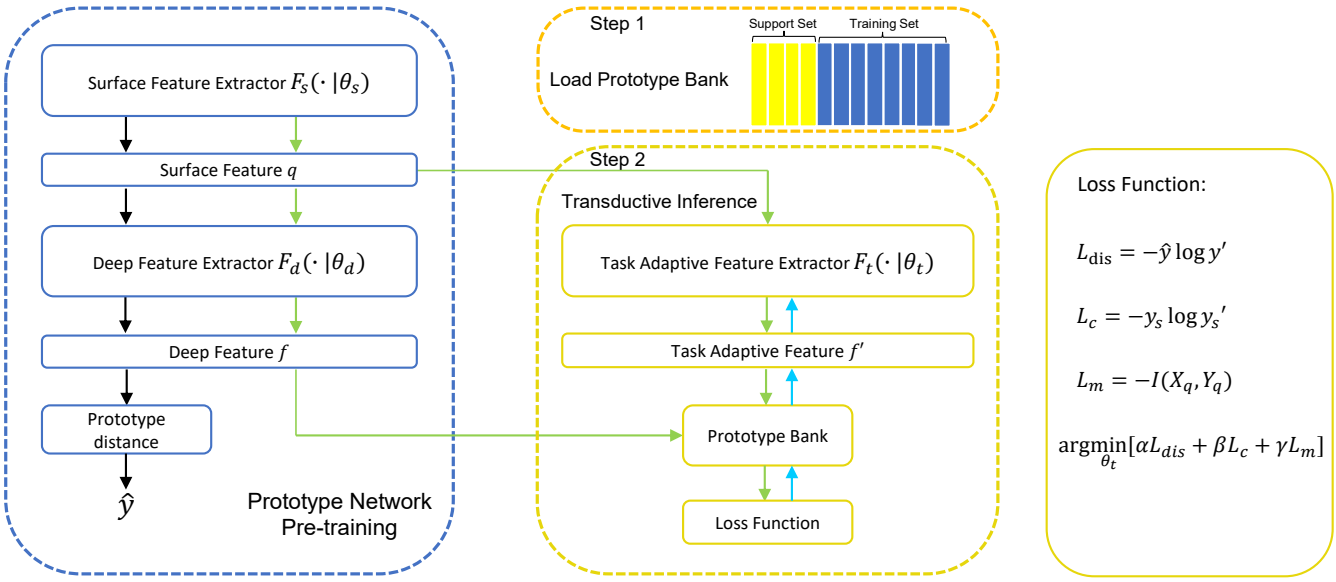

Figure 2: An overview of the proposed model.

3.2. Architecture

3.2.1. Feature Extractors Introduction

As the 2 shows, there are total three defined feature extractors in the architecture of our model, the surface feature extractor $F_s(\cdot|\theta_s)$, deep feature extractor $F_d(\cdot|\theta_d)$ and task adaptive feature extractor $F_t(\cdot|\theta_t)$, which all consist of several convolution layers. We artificially split a pre-trained model as the Surface and deep feature extractor and construct a new randomly initialized feature extractor as the Transductive Feature Extractor.

We suppose that the surface convolution layers in pre-trained model can only observe local information with limited view size so that the extracted surface features are relatively robust to any tasks. The deeper convolution layers in pre-trained model have a wider view of input mel-spectrum and learn more abstract feature corresponding to specific acoustic event, such as the the relationships between different time steps with various peaks of power. Based on these assumptions, the surface feature extractor $F_s(\cdot|\theta_s)$ will be fixed and directly used during the evaluation process. The deep feature extractor $F_d(\cdot|\theta_s)$ will be replaced by the Transductive Feature Extractor $F_t(\cdot|\theta_t)$ to extract the task adaptive feature during the evaluation. This design avoid updating the whole pre-trained model that accelerate the run-time of transductive inference. At the same time, through align the feature space of $F_t(\cdot|\theta_t)$ and $F_d(\cdot|\theta_s)$, less parameters updating design will reduce the probability of overfitting with less data requirement in query set.

3.2.2. Pre-trained Model

As our proposed transductive inference method will not involve in the training process, we use Prototype Network [4] as a specific pre-trained model on training set for a better comprehension. In brief, Prototype Network is a feature extractor that take the center of the embedding features of same classes samples in support set as the corresponding classes prototypes. The queried sample can be predicted through calculating the distance between its embedding feature and each class prototype, where the nearest class prototype is the result of prediction. In our work, we divide Prototype Network into two parts surface feature extractor $F_s(\cdot|\theta_s)$ and deep feature extractor $F_d(\cdot|\theta_d)$ as the figure 2 shows. This division will not influence the training process of Prototype Network:

$$f = F_d(F_s(x|\theta_s)|\theta_d) \tag{1}$$

$$\underset{\theta_s,\theta_d}{\arg\min} -y_c \log \frac{exp(-d_\phi(f, v_c))}{\sum_{c' \in \mathcal{C}} exp(-d_\phi(f, v_{c'}))} \tag{2}$$

where $\mathcal{C}$ is the classes set in training process, d_ϕ is the distance function (L2 distance here) and v_c is the prototype (center point) of class c in feature space. After training with meta learning technique, the Prototype Network will act as the pre-trained model in our following transductive inference process.

3.2.3. Feature Space Alignment

For each single few-shot task, the testing set will provide a batch of extra samples consist of few labelled samples (Support Set) and several unlabelled (Query Set). We construct a Transductive Feature Extractor $F_t(\cdot|\theta_t)$ to extract the task adaptive feature according to the information brought by testing set which will replace original deep feature extractor $F_d(\cdot|\theta_d)$ during testing process. The structure of Transductive Feature Extractor can be of any non-linear mapping function that can be back-propagated. In our work, we drop one convolution layer based on $F_d(\cdot|\theta_d)$ structure as the Transductive Feature Extractor with randomly parameter initialization.

The re-training of $F_t(\cdot|\theta_t)$ in transductive inference process will easily lead to a overfit problem with traditional cross-entropy classification loss due to the few labelled samples in support set. To solve this problem, we propose a feature space alignment method to transfer the original knowledge from $F_d(\cdot|\theta_d)$ to $F_t(\cdot|\theta_t)$. We concatenate the new prototypes retrieved from support set and all

the classes prototypes of training dataset as a prototype bank $W \in R^{(k+m)*z}$, where k is the number of classes in support set, m is the number of classes in training dataset and z is the dimension number of deep feature f. The prototype bank W can be considered as the classifier in pre-trained model, as the distance with each prototype represent the weight of each class. For the sample x in both support set and query set, we align task adaptive feature space with original deep feature space and as following:

$$q = F_s(x|\theta_s) \tag{3}$$

$$f = F_d(q|\theta_d), \quad f' = F_t(q|\theta_t) \tag{4}$$

$$\hat{y} = softmax(d_\phi(W, f)/t) \tag{5}$$

$$y' = softmax(d_\phi(W, f')) \tag{6}$$

$$L_{dis} = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{m+k}\hat{y}_i[j]\log y'_i[j] \tag{7}$$

where N is the batch size, t is a temperature coefficient, $y', \hat{y} \in R^{m+k}$, $y'[j]$ means the j^{th} dimension of y' and d_ϕ is the L2 distance here. L_{dis} encourage the task adaptive feature of each sample has the identical weight with deep feature to each prototypes. This alignment method can transfer the pre-trained knowledge in $F_d(\cdot|\theta_d)$ to $F_t(\cdot|\theta_t)$ through the task samples from both support set and query set.

3.2.4. Task Adaptation

We introduce the cross-entropy classification loss and regularier L_m of maximizing the mutual information [10] to adapt the specific few-shot task.

The cross-entropy classification L_c loss only involves the data in support set:

$$L_c = -\frac{1}{N_s}\sum_{i}^{N_s}\sum_{j}^{m+k} y_{s_i}[j]\log y'_{s_i}[j] \tag{8}$$

$$\tag{9}$$

where N_s is the batch size of support set data and y_{s_i} is the label of sample i in support set.

The regularier L_m of maximizing the mutual information only involve the data in query set:

$$y''_q = softmax(d_\phi(W[:k], f')) \tag{10}$$

$$\bar{y''_q} = \frac{1}{N - N_s}\sum_{i}^{N-N_s} y''_{qi} \tag{11}$$

$$L_m = \sum_{j}^{k}\bar{y''_q}[j]\log\bar{y''_q}[j] - \frac{1}{N - N_s}\sum_{i}^{N-N_s}\sum_{j}^{k} y''_{q_i}[j]\log y''_{q_i}[j] \tag{12}$$

where $W[:k] \in R^{k*z}$ is all the prototypes of support set and N is same with the batch size in equation (7). The L_m is aimed at maximizing the mutual information of X_q, Y_q, which can be considered as:

$$-I(X_q, Y_q) = -H(Y_q) + H(Y_q|X_q) \tag{13}$$

where terms corresponds terms in L_m.

3.2.5. Optimizing

The total loss of transductive inference optimizing is:

$$\arg\min_{\theta_t} \alpha L_{dis} + \beta L_c + \gamma L_m \tag{14}$$

where $\alpha = 0.85, \beta = 0.07, \gamma = 0.08$ in our experiments.

The Transductive Feature Extractor $F_t(\cdot|\theta_t)$ will be updated following Adam optimizer with 1e-4 learning rate for 10 epoches.

3.2.6. Evaluation

During Evaluation, we drop the prototypes of training set and use task adaptive feature to calculate the distance of prototypes of support set. The class of nearest prototypes is the prediction result.

4. EXPERIMENT

4.1. Dataset

All used data belong to the dataset of DCASE 2022 task5. The training dataset of DCASE 2022 task5 consists of five sets of audio files deriving from a different source each. For evaluation, each audio file will be considered as an independent few-shot task while various audio files make up the evaluation set. Each evaluation audio file will be annotated the earliest 5 target event segments as the positive samples and the rest segments before the end of the fifth positive segment are labelled as negative samples.

4.2. Settings

For the training process, we adopt the same setting of baseline during our pre-trained model training. The audio clip segment follows 0.2s segments length and 0.1s hopping length for training dataset. We use the Short Time Fourier Transform with 22.05kHz down sampling rate, 1024 window size and 256 hop size to extract the 128 dimensions mel-spectrum. The training dataset will be divided as 0.9 and 0.1 for training and validation. We train the the prototype network for 50 epochs and choose the best model in validation as the pre-trained model.

For transductive inference, we construct two convolution layers as the task adaptive feature extractor. The Adam optimizer is used for re-training the task adaptive feature extractor with learning rate 0.0001 for 10 epochs. Due to the segment size of each file in evaluation set are not constant, we split each samples' mel-spectrum as a group of 17x128 mel-spectrums, which is same with the mel-spectrum size in training process, to do the data augmentation.

4.3. Experimental Results

The metrisc for evaluation are the event based F-measure, precision and recall. Table 1 shows the result of our model and the competitors. Baseline (official) is based on prototype network [4] and the result is provided by DCASE community. Baseline is also the prototype network but trained by ourselves. TIM [**?**] is one of the state-of-the-arts transductive inference method, which utilizes maximizing the mutual information to update the classifier. Our model is the model using the given parameters ($\alpha = 0.85, \beta = 0.07, \gamma = 0.08$). Our model (ablation) is the ablation study result with parameters ($\alpha = 0.85, \beta = 0.07, \gamma = 0$). Our model (best) is the best result of experiments of our model after the fine tuning process of parameter ($\alpha = 0.854, \beta = 0.067, \gamma = 0.79$).

Table 1: The experiments results.

Model	Precision	Recall	F-measure
Baseline(official)	36.34	24.96	29.59
Baseline	33.09	43.65	37.64
TIM	52.21	40.46	45.59
Our model	64.46	47.90	54.96
Our model (ablation)	66.95	43.71	52.89
Our model (best)	**67.92**	**49.69**	**57.40**

The results of our model shows that effectiveness of the proposed framework and the ablation study results indicate that the mutual information regularier is benefit to our model but not the major factor of the huge improvement of our model.

5. CONCLUSION

In this report, we analyze the problems of existing transductive methods when facing the extraordinary feature distribution of newly appeared classes. Then we propose a new transductive inference framework that introduce the knowledge transfer techniques to restrict the feature distribution of newly appeared classes in a dedicate sub-space. This design effectively alleviate the overfit problem of large-scale parameter updating in transductive inference and well limited the feature distribution of newly appeared classes in few-shot learning task, which shows a great performance in DCASE 2022 task5.

6. ACKNOWLEDGEMENT

This work was supported by Acoustic Scenes Identification based on Domain Adaptation, a research project funded by the National Natural Science Foundation of China (62001038).

7. REFERENCES

[1] E. G. Miller, N. E. Matsakis, and P. A. Viola, "Learning from one example through shared densities on transforms," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, vol. 1. IEEE, 2000, pp. 464–471.

[2] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, no. 33, 2011.

[3] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[4] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[5] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[6] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1842–1850.

[7] T. Munkhdalai and H. Yu, "Meta networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2554–2563.

[8] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[9] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[10] M. Boudiaf, I. Ziko, J. Rony, J. Dolz, P. Piantanida, and I. Ben Ayed, "Information maximization for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2445–2457, 2020.

[11] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," *arXiv preprint arXiv:1909.02729*, 2019.

[12] I. Ziko, J. Dolz, E. Granger, and I. B. Ayed, "Laplacian regularized few-shot learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 660–11 670.

[13] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," *Advances in Neural Information Processing Systems*, vol. 32, 2019.