INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Transductive Feature Space Regularization for Few-shot Bioacoustic Event Detection

Yizhou Tan[1], *Haojun Ai*[1,*], *Shengchen Li*[2], *Feng Zhang*[1]

[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China
[2] Department of Intelligent Science School of Advanced Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China

yizhou.tan@ieee.org, aihj@whu.edu.cn, Shengchen.li@xjtlu.edu.cn, fengzhang@whu.edu.cn

Abstract

In few-shot bioacoustic event detection, besides interested target events, background noises and various uninterested sound events lead to complex decision boundaries, which require regularized feature distributions in feature space. Due to the low label availability of uncertain noise events, existing few-shot learning methods with entropy-based regularizers suffer from overfitting during optimization. In this paper, we propose a transductive inference model with a prior knowledge based regularizer (PKR) to overcome the overfitting problem. We use a task-adaptive feature extractor to reconstruct a regularized feature space. A PKR is proposed to minimize the divergence between the original and reconstructed feature space. The development set of DCASE 2022 Task 5 is adopted as the experimental dataset. With the increasing iterations, the proposed model performs with long-lasting results around 55.43 F-measure, and well solves the overfitting problem in transductive inference.

Index Terms: Few-shot Learning, Transductive Inference, Bioacoustic Event Detection

1. Introduction

Due to the limited manual labor in data annotation, few-shot learning [1–3] has become a promising paradigm to solve tasks with few labeled data, such as few-shot bioacoustic event detection tasks [4]. In a few-shot bioacoustic event detection setting, a *backbone* model is first trained with sufficient labeled data of *base classes* in training set. For the model evaluation, each audio file is usually a separate few-shot task containing unlabeled data from new classes for prediction (Query Set), and a few labeled data from each new class are given (Support Set). Besides the interested target events, the background noises and various uninterested events are all considered as noise events. Due to the sparsity of the target events, the infrequent presence of target events causes a data imbalance problem with noise events in support set. The dense noise events with uncertain patterns make the few-shot bioacoustic event detection task more difficult.

Traditional few-shot learning methods [5–9] usually take the pre-trained backbone model as a general feature extractor and construct a simple classifier according to the support set. For example, Prototype Network [6] constructs a linear classifier by calculating the distance to the center point (prototype) of each class. However, plenty of noise events with uncertain patterns are embedded with an irregular feature distribution in the feature space. The decision boundaries between noise and target events are too complex to fit by these simple classifiers, as shown in Figure 1 (a). To address this issue, we start from the feature space reconstruction, as shown in Figure 1 (b), to enable the divisible decision boundaries with transductive inference.

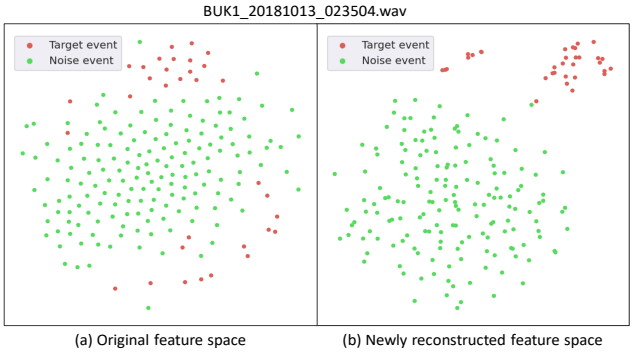

Figure 1: *This figure visualizes the embedded feature distribution of target and noise events in different feature spaces through the t-SNE algorithm [15]. The audio file BUK1_20181013_023504.wav is chosen with two classes of events (target and noise).*

It is natural to leverage the labeled data with the cross-entropy loss to optimize the feature space. Due to the limited labeled data in support set, the feature space suffers from an overfitting problem during the optimization. Transductive inference methods [8, 10–14] are recently proposed to leverage the unlabeled data in query set to regularize the supervised optimization process. A regularizer item is usually added to the loss function to prevent the overfitting problem. Most existing regularizers are based on Shannon Entropy [13, 16] and its deformations [8,17], which constrain the posterior predicted probability distribution to the lower Shannon entropy. Intuitively, these entropy-based regularizers can cluster similar samples together to reach higher classification confidence. However, the entropy-based regularizers can not handle the noise events well due to the high Shannon Entropy of noise events. Considering the data imbalance of dense noise events in both the support and query set, the loss function of cross-entropy loss and entropy-based regularizers will encourage the model to construct an abnormal feature space in a local optimization point, where all queried samples are classified as noise events for high confidence. To overcome the overfitting problem, a prior knowledge based regularizer (PKR) is proposed to regularize the optimization by transferring prior knowledge into transductive inference.

This paper proposes a transductive inference model to re-

The research project is supported partly by the National Natural Science Foundation of China (No: 62001038 and No: 61971316), and Gusu Innovation and Entrepreneurship Leading Talents Programme - Youth Innovation Leading Talent (ZXL2022472).

*Corresponding Author

10.21437/Interspeech.2023-579

construct the feature space with a PKR. The Prototype Network [6] is used as the backbone model. Rather than optimizing the whole backbone model, we build a task-adaptive feature extractor to construct a new feature space with transductive inference. A few labeled data in the support set will be used to optimize the newly constructed feature space via the cross-entropy loss. A PKR is proposed to regularize the optimization by minimizing the divergence between the original and newly constructed feature spaces. The prototypes of base classes in the training set are introduced into the transductive inference process as prior knowledge. The PKR uses prior knowledge to measure the feature distribution divergence with two metrics, distance-based and angle-based. Intuitively, PKR aligns the original and newly constructed feature spaces at the global level, while cross-entropy can optimize decision boundaries at the local level without overfitting. The traditional entropy-based regularizer is also added in the final loss function for higher classification confidence and acts as the control group in the ablation study to verify the effectiveness of the PKR. All experiments are conducted on the development dataset of DCASE 2022 Task 5. The proposed model outperforms prior transductive inference methods with 55.43 F-measure. Along with the increasing iterations of optimization, PKR shows stable performance instead of the overfitting problem in entropy-based regularizers.

2. Method

2.1. Few-shot Scenario

In few-shot bioacoustic event detection, the training dataset $X_{train} = \{(x_i, y_i)|y_i \in Y_{train}\}$ is a large scale of labeled dataset, where x_i is an audio clip, y_i is corresponding event class label and Y_{train} is the set of all base classes in the training dataset. The testing set for each few-shot task consists of a support set $X_s = \{(x_i, y_i)|y_i \in Y_s\}$ and a query set $X_q = \{x_i\}$, where $Y_s = \{target, noise\}, Y_{train} \cap Y_s = \emptyset$. The few-shot event detection task can be converted to a classification task by splitting the whole audio file into several segments. The support set is a continuous piece of labeled audio that contains 5-shot target events, and the left pieces of duration between two target events are taken as noise events. As the length of noise events duration is far longer than target events in common sense, there are many more labeled segments of noise events in support set with a data imbalance problem.

2.2. Transductive Inference

The transductive inference is a few-shot strategy that deems that the unlabeled data in query set can be utilized to further improve the model performance instead of dropping them during the prediction process. The whole query set is accessible in transductive inference to further optimize the model before the final prediction.

2.3. Pre-trained Backbone Model

We train Prototype Network [6] in the training set as the pre-trained backbone model for later transductive inference. In brief, Prototype Network is a model that takes the center of the embedding features of the same classes samples in support set as the corresponding class prototypes. The queried sample can be predicted by calculating the distance between its embedding feature and each class prototype, where the nearest class prototype is the prediction result. In our work, the Prototype Network

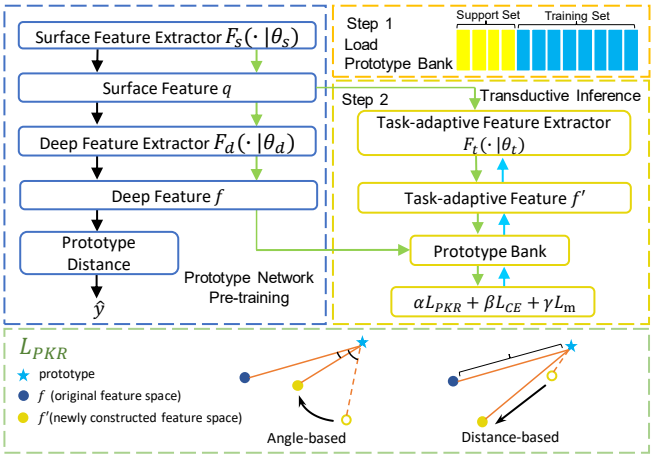

Figure 2: *An overview of the proposed model. The prototype bank contains the center points (prototypes) of all base classes in the training set. The green arrows indicate the forward propagation process, while the blue arrows are the backpropagation process.*

is artificially divided into two components, surface feature extractor $F_s(\cdot|\theta_s)$ and deep feature extractor $F_d(\cdot|\theta_d)$ as Figure 2 shows. This division will not influence the training process of Prototype Network:

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

$$\arg\min_{\theta_s,\theta_d} -y \log \frac{exp(-d_\phi(f, v_y))}{\sum_{c\in\mathcal{C}} exp(-d_\phi(f, v_c))} \tag{2}$$

where $\mathcal{C}$ is the classes set in the 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.

2.4. Architecture

As Figure 2 shows, there are three 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 $F_t(\cdot|\theta_t)$. $F_s(\cdot|\theta_s)$ and $F_d(\cdot|\theta_d)$ constitutes the backbone model, while $F_t(\cdot|\theta_t)$ is newly constructed with random parameter initialization for optimization in transductive inference.

We suppose that the beginning shallow convolution layers in the backbone model can only observe local information with a limited receptive field so that the extracted surface features are relatively general to different bioacoustic events. The deeper convolution layers in the back part of the backbone model have a larger receptive field of input spectrum to learn more abstract semantic features of base classes [18], which may be unhelpful to the novel events in query set. Based on these assumptions, the surface feature extractor $F_s(\cdot|\theta_s)$ will be fixed and consistent in transductive inference. The task-adaptive feature extractor $F_t(\cdot|\theta_t)$ will be optimized to construct a new feature space in transductive inference to replace the original feature space constructed by $F_d(\cdot|\theta_s)$. This design enables the adjustment of updated parameter amount to accelerate the transductive inference run-time by the discretionary $F_t(\cdot|\theta_t)$. To optimize the $F_t(\cdot|\theta_t)$, we propose a prior knowledge based regularizer to prevent the overfitting problem caused by cross-entropy loss in support set.

2.4.1. Prior Knowledge based Regularizer (PKR)

In few-shot bioacoustic event detection, the uncertain noise events do not have stable patterns. The plenty of noise events are embedded with an irregular feature distribution and make the class decision boundaries too complex to fit by a linear classifier. The limited capability of classifiers requires the reconstruction of feature space for simple divisible decision boundaries of noise events. In transductive inference, $F_t(\cdot|\theta_t)$ is optimized to construct a new task-adaptive feature space with cross-entropy loss in support set. Due to the limited labeled data in support set, the cross-entropy loss will rapidly drop $F_t(\cdot|\theta_t)$ into overfitting during the optimization.

To solve this overfitting problem, we propose a prior knowledge based regularizer aimed at transferring the original knowledge from $F_d(\cdot|\theta_d)$ to $F_t(\cdot|\theta_t)$ through feature space alignment. We concatenate the new prototypes [6] (mean feature point of a class) of support set and all base 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 base classes in training dataset and z is the dimension number of deep feature f. The prototype bank W can be considered as the anchors of both the original deep feature space and the newly constructed task-adaptive feature space. The principle of feature space alignment is to keep the relative position of each sample to anchors as constant as possible in two feature spaces. This principle enables the general consistent feature distribution of the entire query set and can optimize local feature distribution through cross-entropy loss. Based on the above principle, we propose the two metrics, angle-based and distance-based, to measure the divergence of two feature distributions in original and newly constructed feature space. Although the distance-based metric performs better in experiments, the angle-based metric is still introduced to show the potential idea of regularizing the feature space reconstruction.

Angle-based Divergence (PKR-A): The angle-based divergence measures the consistent angles between the task-adaptive feature, original deep feature, and the prototype bank:

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

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

$$L_{\text{PKR}} = -\frac{1}{N}\sum_{i}^{m+k} -cos\_similarity(f - W[i], f' - W[i]) \tag{5}$$

where x comes from both support and query set, N is the batch size, and cos_similarity is the function to calculate the cosine value of the angle between two vectors.

Distance-based Divergence (PKR-D): It is worth noting that the prototype bank W can be considered as the classifier in the backbone model, as the distance with each prototype represents the weight of each class. As a result, the distance-based divergence can be represented as the prediction probability distribution as follows:

$$\hat{y} = \textit{softmax}(d_\phi(W, f)/t) \tag{6}$$

$$y' = \textit{softmax}(d_\phi(W, f')) \tag{7}$$

$$L_{\text{PKR}} = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{m+k} \hat{y}_i[j] \log y'_i[j] \tag{8}$$

where 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. Noteworthy, the number of dimensions is not a significant parameter, as the L_{PKR} reflects relative distance relationships. Coincidentally, L_{PKR} has a consistent form with the knowledge distillation technique [19], while we speculate that the better performance of PKR-D may be ascribed to the temperature coefficient.

2.4.2. Feature Space Reconstruction

We further introduce the cross-entropy classification loss and entropy-based regularizer L_m of maximizing the mutual information [8] for higher classification confidence. The cross-entropy loss L_{CE} only involves the data in support set:

$$L_{\text{CE}} = -\frac{1}{N_s}\sum_{i}^{N_s}\sum_{j}^{m+k} y_{s_i}[j] \log y'_{s_i}[j] \tag{9}$$

where N_s is the batch size of support set data, and y_{s_i} is the label of a sample i in support set. The entropy-based regularizer L_m of maximizing the mutual information [8] only involves the data in query set:

$$y''_q = \textit{softmax}(d_\phi(W[:k], f')), \quad \bar{y''_q} = \frac{1}{N - N_s}\sum_{i}^{N-N_s} y''_{q_i} \tag{10}$$

$$L_m = \sum_{j}^{k} \bar{y''_q}[j] \log \bar{y''_q}[j] - \frac{1}{N - N_s}\sum_{i}^{N-N_s}\sum_{j}^{k} y''_{q_i}[j] \log y''_{q_i}[j] \tag{11}$$

where $W[:k] \in R^{k*z}$ is all the prototypes of support set, and N is the same with the batch size in Equation (8). The total loss of transductive inference optimizing is:

$$\arg\min_{\theta_t} \alpha L_{\text{PKR}} + \beta L_{\text{CE}} + \gamma L_m \tag{12}$$

3. Experiment

3.1. Experimental setups

Dataset All used data belong to the development dataset of DCASE 2022 task 5 [20]. The training dataset consists of five sets of audio files derived from different sources, containing 174 audio recordings, 21 hours duration, 47 classes, and 14229 events. The testing set consists of 18 audio recordings with a total of about 6 hours duration and 1077 total positive events.

Feature and Segment The Short Time Fourier Transform is used with 22.05kHz down-sampling rate, 1024 window size and 256 hop size to extract the 128 dimensions per-channel energy normalization (PCEN) features. For training, the audio clip segment follows 0.2s segments length and 0.1s hopping length for sampling. For testing, we follow the setting of the official baseline [20].

Training Setting The training dataset will be divided into 0.9 and 0.1 for training and validation. The prototype network is composed of 3 residual block layers and will be trained for 50 epochs, where the best model in validation will act as the pre-trained backbone model. In transductive inference, we construct two convolution layers as the task-adaptive feature extractor. The Adam optimizer is used to optimize the task-adaptive feature extractor with a 0.0001 learning rate for 15 epochs. The parameters of loss function is $\alpha = 1, \beta = 0.1, \gamma = 0.1$. This choice is due to the fact that L_{PKR} should have a larger weight than L_{CE} to overcome overfitting, while L_m follows the same weight with L_{CE} according to Yang et al. [17].

Model Evaluation The model performance is evaluated by an event-based *F*-measure metric [21].

Table 1: *The comparison of different methods.*

Model	Precision	Recall	*F*-measure
Baseline (TM) [20]	2.42	18.32	4.28
Baseline (PN) [20]	36.34	24.96	29.59
Prototype Network [6]	28.76	38.85	33.05
TIM [8]	52.31	30.10	38.21
Fine-tuning [13]	47.50	34.40	39.90
TI-ML [17, 22]	59.20	46.34	51.99
Dcase-Top-1 [23]	-	-	**74.40**
Dcase-Top-2 [24]	-	-	50.00
Dcase-Top-3 [25]	-	-	60.00
Ours (PKR-A)	49.64	43.91	46.60
Ours (PKR-D)	**60.69**	**51.02**	55.43

Table 2: *The ablation study of our proposed models.*

Our Model	Regularizer		*F*-measure
	L_m (entropy-based)	L_{PKR}	
Backbone			33.05
Entropy-based	✓		38.66
PKR-A-only		✓	45.54
PKR-A	✓	✓	46.60
PKR-D-only		✓	55.37
PKR-D	✓	✓	**55.43**

3.2. Experimental Results and Analysis

Competitors: Baseline(TM/PN) are the official baselines provided by DCASE community [20]. Prototype Network [6] is the famous framework in few-shot learning. TIM [8], Fine-tuning [13] are state-of-the-art transductive inference methods that optimize the classifier and whole model respectively with entropy-based regularizers. TI-ML [17, 22] is the state-of-art transductive inference mutual learning framework with extra data augmentation in few-shot bioacoustic event detection. As the model performance is related to several factors such as model structure and external datasets, we choose the top 3 models in the official rank of DCASE to show the potential model capabilities instead of a strict performance comparison.

Results: Table 1 shows that our models outperform all the competitors in all metrics. This distance-based proposed regularizer (PKR-D) outperforms the pre-trained backbone model (Prototype Network) and prior transductive inference methods. Dcase-Top-3 outperforming Dcase-Top-2 in the public validation set indicates the few-shot bioacoustic sound detection still facing a generalization problem in different datasets. Considering the external datasets in Dcase-Top-2, better performance of the proposed PKR-D shows the effectiveness of the feature space reconstruction, while PKR-A performance indicates the potential of other prior-based regularizers.

Ablation Study: As the third term L_m in the loss function Equation (12) is the entropy-based regularizer, the ablation study is conducted to verify that improvement of our model performance is ascribed to the proposed PKR instead of the existing entropy-based regularizer, as Table 2 shows. Compared with Backbone, PKR-D-only and PKR-A-only, the results indicate that our proposed PKR L_{PKR} greatly improves the performance based on the pre-trained backbone independently. When there is no PKR in feature space reconstruction, the Entropy-based result is inferior to both PKR-A-only and PKR-D-only, which verifies the irreplaceable state of the PKR in the proposed architecture. The entropy-based regularizer L_m can only benefit the PKR models with subtle improvement.

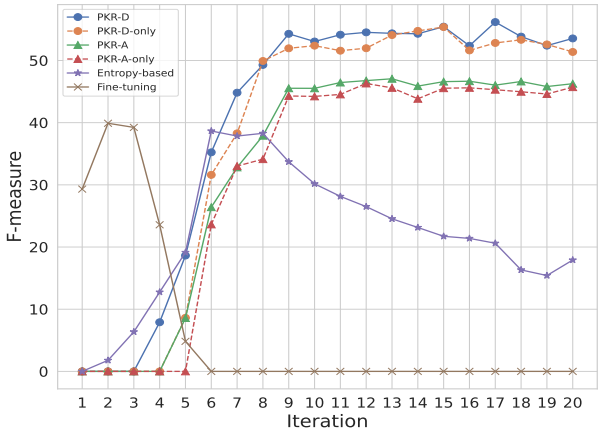

Figure 3: *The figure shows the robustness of models with different regularizers during the optimization. The regularizers are from prior works and the ablation study.*

Influence of optimizing Iterations: In transductive inference, the termination condition is important but difficult to determine, especially involving large-scale parameters optimization. The usual practice (Fine-tuning [13]) sets a fixed optimizing iteration with a small learning rate, which highly relies on empiricism and the robustness of the model. Particularly in few-shot bioacoustic event detection, we assume that entropy-based regularizers are heavily influenced by optimizing iterations with an overfitting problem. To verify the overfitting problem in entropy-based regularizer and the robustness of our PKR, we optimize the Fine-tuning model [13] and ablation components of our model with different iteration, as shown in Figure 3. The Entropy-based and Fine-tuning models are all suffering a performance drop along with the increasing iterations of optimization after the best performance point, as both of them only use entropy-based regularizers disabled by the uncertain noise events. These results verify the overfitting problem of entropy-based regularizers with uncertain noise events and illustrate the difficulty of ending point choice with entropy-based regularizers. In contrast, our proposed PKR shows a great stable performance along with increasing optimizing iterations, which well solves the overfitting problem and verifies the effectiveness of our reconstructed feature space. As a result, we recommend considering the prior knowledge in a few-shot transductive inference design, which can effectively reduce the difficulty of ending point selection in optimization.

4. Conclusion

This paper proposes a transductive inference model to reconstruct a regular feature space for few-shot bioacoustic event detection. A novel prior knowledge based regularizer is further proposed to address the overfitting problem during the feature space reconstruction. The proposed model outperforms existing transductive inference methods with more robust performance in DCASE 2022 Task 5. This recommends introducing prior knowledge into few-shot transductive inference for more stable performance. In the future, we will further explore the adaptive ending point of optimization to accelerate transductive inference. The source code has been released[1].

[1]https://github.com/Voltmeter00/DCASE2022Task5

5. 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] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015, p. 0.

[4] V. Morfi, I. Nolasco, V. Lostanlen, S. Singh, A. Strandburg-Peshkin, L. F. Gill, H. Pamula, D. Benvent, and D. Stowell, "Few-shot bioacoustic event detection: A new task at the DCASE 2021 challenge." in *DCASE*, 2021, pp. 145–149.

[5] 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.

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

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

[8] 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.

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

[10] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. Hwang, and Y. Yang, "Learning to propagate labels: Transductive propagation network for few-shot learning," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[11] 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.

[12] J. Kim, T. Kim, S. Kim, and C. D. Yoo, "Edge-labeling graph neural network for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11–20.

[13] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," in *International Conference on Learning Representations, ICLR*, 2020. [Online]. Available: https://openreview.net/forum?id=rylXBkrYDS

[14] M. Boudiaf, H. Kervadec, Z. I. Masud, P. Piantanida, I. Ben Ayed, and J. Dolz, "Few-shot segmentation without meta-learning: A good transductive inference is all you need?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 979–13 988.

[15] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[16] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2004.

[17] D. Yang, H. Wang, Y. Zou, Z. Ye, and W. Wang, "A mutual learning framework for few-shot sound event detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 811–815.

[18] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[20] I. Nolasco, S. Singh, A. Strandburg-Peshkin, L. Gill, H. Pamula, J. Morford, M. Emmerson, F. Jensens, H. Whitehead, I. Kiskin, E. Vidana-Villa, V. Lostanlen, V. Morfi, and D. Stowell, "DCASE 2022 Task 5: Few-shot Bioacoustic Event Detection Development Set," Mar. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6482837

[21] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162–178, 2016.

[22] D. Yang, Y. Zou, F. Cui, and Y. Wang, "Improved prototypical network with data augmentation technical report," DCASE2022 Challenge, Tech. Rep., June 2022.

[23] J. Tang, Z. Xueyang, T. Gao, D. Liu, X. Fang, J. Pan, Q. Wang, J. Du, K. Xu, and Q. Pan, "Few-shot embedding learning and event filtering for bioacoustic event detection technical report," DCASE2022 Challenge, Tech. Rep., June 2022.

[24] H. Liu, X. Liu, X. Mei, Q. Kong, W. Wang, and M. D. Plumbley, "Surrey system for dcase 2022 task 5 : Few-shot bioacoustic event detection with segment-level metric learning," Tech. Rep., 2022.

[25] J. Martinsson, M. Willbo, A. Pirinen, O. Mogren, and M. Sandsten, "Few-shot bioacoustic event detection using a prototypical network ensemble with adaptive embedding functions technical report," DCASE2022 Challenge, Tech. Rep., June 2022.