Invariance-based layer regularization for sound event detection

David Perera, Slim Essid, Richard Gaël

Invariance-based layer regularization for sound event detection

David Perera, Slim Essid, Gaël Richard
LTCI, *Télécom Paris*, *Institut Polytechnique de Paris*
Palaiseau, France
david.perera@telecom-paris.fr, gael.richard@telecom-paris.fr, slim.essid@telecom-paris.fr

***Abstract*—Experimental and theoretical evidences suggest that invariance constraints can improve the performance and generalization capabilities of a classification model. While invariance-based regularization has become part of the standard tool-belt of machine learning practitioners, this regularization is usually applied near the decision layers or at the end of the feature extracting layers of a deep classification network. However, the optimal placement of invariance constraints inside a deep classifier is yet an open question. In particular, it would be beneficial to link it to the structural properties of the network (*e.g.* its architecture), or its dynamical properties (*e.g.* the effectively used volume of its latent spaces). The purpose of this article is to initiate an investigation on these aspects. We use the experimental framework of the DCASE 2023 Task 4A challenge, which considers the training of a sound event classifier in a semi-supervised manner. We show that the optimal placement of invariance constraints improves the performance of the standard baseline for this task.**

***Index Terms*—DCASE task 4, invariance-based learning, semi-supervised learning.**

I. Introduction

Deep learning has proven to be effective for a wide array of tasks. However, it usually relies on the availability of large amounts of data, especially annotated data [1]. For supervised tasks, such as audio classification, collecting annotations at scale is costly and time consuming. To mitigate this difficulty, an extensive research effort has been devoted to algorithms that use less direct supervision or leverage unlabeled data [2].

In this context, invariance-based learning is a particularly interesting technique, because of its experimental efficiency [3], its theoretical properties [4] as well as its links with perception [5]. In its simplest version, this technique uses two data augmentation pipelines τ and τ' to artificially generate two views of an input data point x, and constrains the model outputs $f(\tau(x))$ and $f(\tau'(x))$ to be similar.

Building on this framework, many systems have been proposed to enforce the output's invariance [6], several strategies have been devised to sample the input points or the data augmentation pipeline [7], and different extraction points have been used when f is a deep neural classifier [8]. These variants have been studied through the lens of various criteria such as information bottleneck [9], latent space geometry [10], or confirmation bias [11]. However, the optimal placement of this regularization inside a deep neural network remains an open question.

In this article, we propose to study these aspects. We experimentally find the optimal extraction layer for invariance-based regularization, *i.e.* the network layer to which invariance regularisation should be applied, and show that it depends non trivially on the data augmentation and the target evaluation metric. We then correlate this extraction layer with statistical properties of the network, such as the distribution of class information and encoding complexity through its layers. For this study, we use the experimental framework proposed by the Task 4A of the DCASE 2023 challenge,[1] which involves training a sound event classifier using a combination of strongly labeled, weakly labeled and unlabeled data. We experimentally show that optimal regularization leads to a model that clearly outperforms the baseline system proposed for this task.

This article is organized as follows. In section II, we review related work. Section III provides a formal description of our training framework. Section IV describes the experimental setup used in this study. Finally, in Section V, we present our results and analyses.

II. Related work

There is a large literature on invariant representation learning using data augmentation [12]. The ladder network [8] is an early contribution on invariance-based semi-supervised learning, that trains a neural network to be invariant to small perturbations of its input and internal representations. Mean Teacher (MT) [13] builds upon this idea, and introduces exponential moving averaging techniques in order to mitigate the random aspects of training, such as mini-batch and data augmentation sampling. These techniques are key to ensure network convergence and generalization performances in this training setting. Data augmentation policies usually introduce many hyperparameters, and consequently require a time-consuming tuning phase. This observation has initiated work on the automatic search of augmentation strategies (AutoAugment [7]), or the gradual introduction of data augmentations during training (CREST [14]). In the opposite direction, the authors of RandAugment [15] have proposed a simplified sampling strategy, in order to reduce the hyperparameter search space: they apply one augmentation at a time, and factorize the distortion magnitude of all augmentations into a single scale.

[1] https://dcase.community/challenge2023/task-sound-event-detection-with-weak-labels-and-synthetic-soundscapes

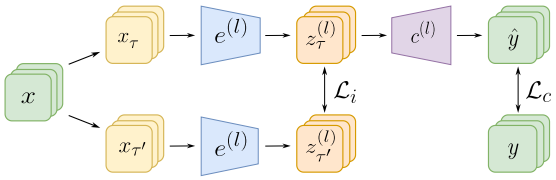

Fig. 1. Training framework.

This strategy has been applied successfully on DCASE 2021 Task 4 dataset [16]. We build upon this method, denoted Random Consistency Training (RCT) in the following.

Several tools have been developed to analyze how deep neural networks encode information through their inner layers. Classical methods include statistics on the neural activations, saliency maps and clustering. Neural based approaches include Deconvnet [17] and Guided Back-Propagation [18], which estimates the parts of the input that are the most discriminative for a given neuron, Class Activation Map [19], which estimates the neurons that are most discriminative for a given class, or Mutual Information Neural Estimation [20], which estimates mutual information between two random variables (*e.g.*, between a latent representation of the network and the target annotation). However, these methods should be used with caution, as they can introduce artifacts and bias the analysis [21].

III. Method

A. Proposed training framework

The regularisation method that we study is depicted Fig. 1. We consider a dataset of annotated audio recordings (x, y). We use two random data augmentations $\tau, \tau' \sim \mathcal{T}$ drawn from a common distribution, in order to generate two different views x_τ and $x_{\tau'}$ of the audio input x.

We then process these two views using a deep neural classifier f, which is made of L layers. We denote by $e^{(l)}$ the first l layers of the classifier and by $c^{(l)}$ its last $L - l$ layers, using the convention $c^{(L)} = I_d$. Consequently, we have $f = c^{(l)} \circ e^{(l)}$ for each $l \in [\![1, L]\!]$. Note that $e^{(l)}$ can be seen as a feature extractor, and $c^{(l)}$ as a classification head. With this view in mind, we will regularize $e^{(l)}$ so that it becomes invariant to the data augmentation τ. We denote by $z_\tau^{(l)} = e^{(l)}(x_\tau)$ and $z_{\tau'}^{(l)} = e^{(l)}(x_{\tau'})$ the latent representations at layer l. Finally, we denote by $\hat{y} = f(x_\tau) = z_\tau^{(L)}$ the output of the classifier.

The data augmentation τ used by RCT [16] consists of low-level audio related augmentations. Consequently, the encoder $e^{(l)}$, which we train to be invariant to τ, discards acoustic variability that is irrelevant to the classification task. This invariance property reduces the dependency of the classifier f to this acoustic variability, thus reducing its error risk and improving its generalization capabilities.

B. Training objectives

In order to optimize the classifier f, we set an extraction layer l, and we use the combination of a classification objective $\mathcal{L}_c$ and an invariance objective $\mathcal{L}_i^{(l)}$:

$$\mathcal{L}_c = \mathcal{L}_c[\hat{y}, y], \tag{1}$$

$$\mathcal{L}_i^{(l)} = \mathcal{L}_i^{(l)}[z_\tau^{(l)}, z_{\tau'}^{(l)}]. \tag{2}$$

We also apply MT regularization [13] to the classifier f. More specifically, we denote by f^{ema} the exponential moving average of this network, which is updated during training after each gradient descent step. Let $\hat{y}_{\text{mt}} = f^{\text{ema}}(x_\tau)$ denote its output. We then optimize the following regularization objective:

$$\mathcal{L}_{\text{mt}} = \mathcal{L}_{\text{mt}}[\hat{y}, \hat{y}_{\text{mt}}]. \tag{3}$$

If we denote by w_i, w_c and w_{mt} the respective loss weights, the complete training objective can be written:

$$\mathcal{L}_{total} = w_i \cdot \mathcal{L}_i + w_c \cdot \mathcal{L}_c + w_{\text{mt}} \cdot \mathcal{L}_{\text{mt}}. \tag{4}$$

The terms w_c and w_{mt} are defined as in the baseline system [22]. The term w_i is taken using a grid search around the value proposed by the RCT method. Their exact values can be found in the code repository provided for reproducibility.[2]

IV. Experimental setup

A. Data

We use for our experiments the framework of DCASE Task 4, which introduces the DESED dataset [22]. The training dataset $\mathbb{D}$ is composed of three subsets: a strongly labeled dataset $\mathbb{D}_s$ of synthetic sounds, a weakly labeled dataset $\mathbb{D}_w$ of real recordings, and an unlabeled dataset $\mathbb{D}_u$ of real recordings. Each sample $x \in \mathbb{D}$ is a 10-s audio recording made in a domestic context. These samples are annotated using a set of 10 classes (*e.g.*, *speech*, *cat*, *blender*, *water*...). The annotation y is either a strong annotation y_s or a weak annotation y_w. A strong annotation is a list of triplets indicating the time onsets and offsets along the classes of the sound events in presence. A weak annotation consists of a list of classes, and does not provide any temporal information about the sound events. In addition to these training datasets, DESED also provides a test dataset $\mathbb{D}_{\text{test}}$ of strongly annotated synthetic recordings.

B. Augmentation pipeline

We apply to each sample x a data augmentation τ in the audio domain. Following RCT, we first draw an audio augmentation, then a distortion magnitude, and we apply their combination to the sample x. We apply modified versions of Mixup [23] and Time Shift. Our modifications ensure that the mixed audio is close to the original audio (mixing parameter α close to 1 in the case of Mixup), so that the invariance objective be relevant. In order to evaluate the impact of the augmentation diversity on the optimal extraction layer, we cluster the augmentations into three increasing sets : τ_0 is used

[2]https://github.com/daperera/irct

by the baseline system while τ_1 and τ_2 have been proposed by RCT. More specifically, τ_0 corresponds to Mixup [23] ; τ_1 additionally includes Time Masking, Time Shifting and Pitch Shifting ; τ_2 includes Frequency masking and FilterAugment [24]. The exact parameter ranges can be found in our code repository.

C. Model

We based our experiments on the baseline of the DCASE Task 4 challenge, which is an important reference largely adopted in the DCASE community.[3] This system won the challenge in 2017 and has been further improved in the subsequent years to match the SOTA. Therefore, we believe that results on this baseline are significant, and can be used to improve other SOTA models.

The architecture of the baseline model is a 7-layer Convolutional Neural Network (CNN) using 2d-convolution, followed by a 2-layer bidirectional Gated Recurrent Unit (GRU) and a Multi Layer Perceptron (MLP) classification head with Attention Pooling [25]. The classification head outputs two predictions for each audio sample x: a weak prediction $\hat{y}_w$ and a strong prediction $\hat{y}_s$. This model is trained using a classification loss $\mathcal{L}_c$, a MT regularisation loss $\mathcal{L}_{\text{mt}}$, and the Mixup data augmentation. It uses log Mel-spectrogram representation and min-max scaling in order to pre-process the audio input x. It uses median filtering in order to post-process the predictions $\hat{y}$.

We use the same pre/post-processing steps and the same architecture for the classifier f as this baseline system. In particular, the DCASE 2023 Task 4A baseline model can be obtained from our framework by using the augmentation set τ_0 and discarding the training objective $\mathcal{L}_i$. We use this re-implementation as a baseline for this study.

D. Training losses

Our approach follows the same choice of metrics as the baseline. We use binary cross-entropy as a classification metric $\mathcal{L}_c = \mathcal{L}_{BCE}$ for both strong annotations y_s and weak annotations y_w. We use the Euclidean distance for the MT and invariance objectives: $\mathcal{L}_{\text{mt}} = \mathcal{L}_i = \mathcal{L}_2$.

E. Evaluation metrics

We use three different metrics to compare the models. In order to evaluate the weak predictions $\hat{y}_w$, we use F1 macro scores. In order to evaluate the strong predictions $\hat{y}_s$, we use a collar-based score [26], which we denote by $event$, and the Polyphonic Sound Detection Score (PSDS) [27], which we denote by $intersection$. We additionally report the scenario 1 and 2 PSDS scores (denoted psds1 and psds2), as well as their threshold invariant versions (denoted by ti-psds1 and ti-psds2) [28]. These metrics are standard in the sound event detection community, and we use the same parameter values for theses metrics as in the evaluation of the DCASE 2023 Task 4A. In order to simplify visualizations, we will focus on the ti-psds2 score when plotting Figures. The other scores present similar behaviors, and support similar conclusions.

Following the recommended protocol for the DCASE 2023 Task 4A challenge, we trained 3 versions of each presented variant, and we report the mean score and standard deviation in Table 1 and Figure 2.

V. Results

A. Impact of augmentation diversity and regularization

The impact of augmentation diversity and regularization's position on the test score is presented in Table I. First, we can see that invariance-based regularization improves for most metrics the performance of the baseline (first line of Table I). Moreover, using a wider augmentation set τ_2 gives better results in most configurations. However, we notice that regularizing the output of the network is never optimal (lines corresponding to layer 9 in Table I), even though this strategy is commonly used in invariance-based learning. In fact, we find that the optimal extraction layer l is non trivially dependent on the audio augmentation set and the target metric.

Fig. 2 shows that the profile of the regularization impact on performance along the network layers has a characteristic double-peak shape for both augmentation sets τ_1 and τ_2. This shape, which favours either early or late regularization, is the same for all metrics. It is exacerbated by using the larger augmentation set τ_2. In the following, we relate this behavior with the statistical properties of the network.

B. Latent space geometry

Fig. 3 provides insights about the distribution of latent vectors in each layer. We use two statistical properties of the unconstrained network, that we compute on the test set $\mathbb{D}_{\text{test}}$. First, we use the maximum absolute activation of a layer l, defined as $\max_{x \in \mathbb{D}_{\text{test}}} \|z^{(l)}\|_\infty$. Note that it is independent from the dimension of the latent space. Second, we use the average distance to the centroid of a layer l, defined as $z^{(l)}_{\text{centroid}} = \sum_{x \in \mathbb{D}_{\text{test}}} e^{(l)}(x)$. These quantities achieve a maximum around the center of the network, suggesting that the network expands the volume taken by its internal representations. Their evolution through the network seems to be complementary to the regularization efficiency profile shown in Fig. 2. In the following, we investigate how this additional volume is used by the network to encode relevant information.

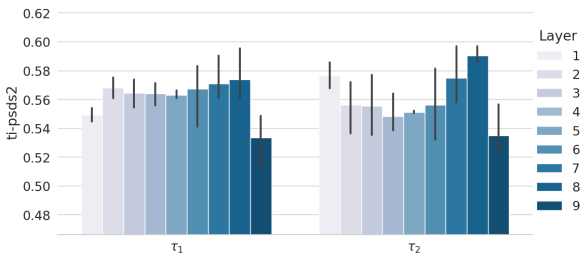

Fig. 2. Impact of layer regularization on the ti-psds2 score, for augmentation sets τ_1 and τ_2.

[3]https://dcase.community/challenge2023/task-sound-event-detection-with-weak-labels-and-synthetic-soundscapes#baseline-system

TABLE I
COMPARISON OF MODEL PERFORMANCE, DEPENDING ON THE DIVERSITY OF THE AUGMENTATION τ AND THE REGULARIZATION LAYER.

Model		Scores*						
τ	Layer	weak	event	intersection	psds1	ti-psds1	psds2	ti-psds2
τ_0	*none*	**0.762 ± 0.009**	**0.645 ± 0.006**	**0.415 ± 0.005**	**0.350 ± 0.008**	**0.535 ± 0.011**	**0.360 ± 0.008**	**0.553 ± 0.011**
τ_1	1	0.757 ± 0.005	0.638 ± 0.008	0.400 ± 0.009	0.336 ± 0.005	0.525 ± 0.005	0.349 ± 0.002	0.549 ± 0.004
	2	0.763 ± 0.008	0.643 ± 0.006	0.395 ± 0.004	0.339 ± 0.004	0.537 ± 0.006	0.355 ± 0.004	0.568 ± 0.006
	3	0.764 ± 0.002	0.638 ± 0.003	0.407 ± 0.008	0.336 ± 0.007	0.532 ± 0.015	0.352 ± 0.002	0.564 ± 0.008
	4	0.758 ± 0.005	0.636 ± 0.013	0.403 ± 0.005	**0.343 ± 0.002**	0.538 ± 0.008	**0.356 ± 0.003**	0.564 ± 0.006
	5	0.763 ± 0.007	0.637 ± 0.007	**0.408 ± 0.010**	0.329 ± 0.006	0.529 ± 0.003	0.344 ± 0.007	0.563 ± 0.002
	6	0.759 ± 0.008	0.644 ± 0.004	0.392 ± 0.007	0.339 ± 0.004	0.540 ± 0.023	0.353 ± 0.002	0.567 ± 0.019
	7	0.759 ± 0.014	0.640 ± 0.010	0.406 ± 0.006	0.341 ± 0.006	0.547 ± 0.018	0.353 ± 0.005	0.571 ± 0.014
	8	**0.769 ± 0.008**	**0.655 ± 0.009**	0.397 ± 0.014	0.337 ± 0.008	**0.552 ± 0.016**	0.346 ± 0.010	**0.574 ± 0.016**
	9	0.754 ± 0.010	0.624 ± 0.015	0.394 ± 0.013	0.326 ± 0.008	0.519 ± 0.016	0.333 ± 0.007	0.533 ± 0.015
τ_2	2	0.776 ± 0.006	0.651 ± 0.004	0.409 ± 0.006	**0.356 ± 0.004**	0.560 ± 0.008	**0.366 ± 0.003**	0.577 ± 0.007
	2	0.762 ± 0.007	0.633 ± 0.018	0.399 ± 0.019	0.334 ± 0.016	0.531 ± 0.014	0.347 ± 0.014	0.556 ± 0.015
	3	0.765 ± 0.009	0.646 ± 0.018	0.399 ± 0.014	0.334 ± 0.007	0.530 ± 0.018	0.347 ± 0.006	0.555 ± 0.017
	4	0.765 ± 0.005	0.637 ± 0.006	0.394 ± 0.014	0.331 ± 0.009	0.518 ± 0.019	0.347 ± 0.006	0.548 ± 0.011
	5	0.753 ± 0.006	0.616 ± 0.008	0.378 ± 0.016	0.335 ± 0.007	0.519 ± 0.002	0.351 ± 0.007	0.551 ± 0.001
	6	0.770 ± 0.011	0.634 ± 0.006	0.388 ± 0.009	0.335 ± 0.001	0.533 ± 0.018	0.349 ± 0.002	0.556 ± 0.020
	7	0.772 ± 0.016	**0.656 ± 0.024**	**0.422 ± 0.018**	0.347 ± 0.009	0.551 ± 0.014	0.358 ± 0.010	0.575 ± 0.016
	8	**0.777 ± 0.003**	0.655 ± 0.004	0.414 ± 0.007	0.336 ± 0.012	**0.570 ± 0.003**	0.345 ± 0.012	**0.590 ± 0.005**
	9	0.757 ± 0.012	0.624 ± 0.019	0.390 ± 0.011	0.331 ± 0.012	0.520 ± 0.015	0.339 ± 0.012	0.535 ± 0.015

* For each evaluation metric and augmentation set, the best scores are indicated in bold. The first line corresponds to the baseline, which does not apply invariance-based regularization.

C. Implicit pretext tasks

The authors of [29] argue that a deep classifier solves implicit pretext tasks, such as context prediction or noise classification, when it is trained on large realistic datasets. Indeed, they show experimentally that its latent representations are highly informative about these pretext tasks, and hypothesize that the network uses this additional information to perform conditional classification. Likewise, we suggest that the deep classifier f uses the observed additional volume in its central layers in order to encode acoustic, semantic and contextual information, that it then uses to solve its target task.

The augmentations τ_1 and τ_2 have been carefully selected for the target classification task, via grid search. However, it is not obvious that these augmentations are still aligned with the network's implicit pretext tasks. Consequently, it is critical to constrain the network f as loosely as possible on its central layers (low data augmentation diversity, high latent space dimension), so that the regularization objective does not compete with the implicit pretext tasks. This observation explains the *lateralization* of the regularization efficiency profile (Fig. 2) when we increase the diversity of the data augmentation set from τ_1 to τ_2. On the contrary, when the latent space distribution is concentrated in a small volume, it is useful to help the network to make the best use of available space using regularization. In order to confirm these hypotheses, we need to take a closer look on how class information is distributed across the network.

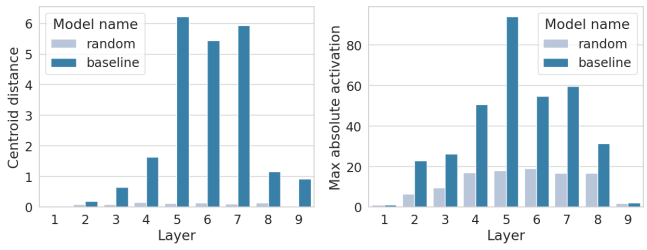

Fig. 3. Average distance to the centroid (on the left), and maximum absolute activation value (on the right), computed for each layer of the network. We compare statistics computed for a trained and a randomly initialized network, in order to prevent architecture bias in the analysis.

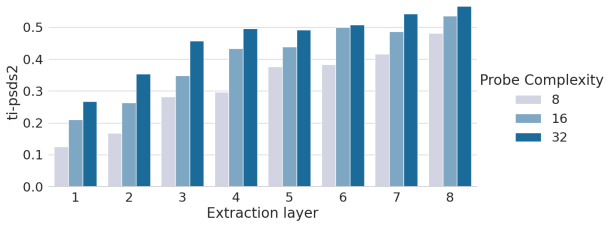

Fig. 4. Comparison of the performance of several probes, depending on the layer they are applied to (grouped by color), and their complexity. Performance is measured using ti-psds2 score. Note that we did not connect probes to the outputs of the network (layer 9).

D. Class information and encoding complexity

The usual clustering and linear correlation algorithms that we tested could not extract class information from the latent representations $z^{(l)}$ of the network. This motivated the study of non-linear probes, which have higher expressivity.

We proceed as follows. We connect a deep probe $p^{(l)}$ to an arbitrary layer l of the frozen network f, and train it using the Task 4A framework. In order to study the complexity of class information encoding at each layer, we gradually increase the complexity of the probe $p^{(l)}$. The probe is a MLP followed by a GRU and a classification head with Attention Pooling. In order to adjust the probe's complexity, we modify the latent

dimension of its classification head from 8 to 32. The result of this experiment is shown Fig. 4.

We observe that the score of the probe increases when it is attached to later layers. This shows that class information increases throughout the network. We also observe that increasing the probe complexity improves its score, for all considered layers. Moreover, as we increase the complexity of the probe, the considered metric quickly reaches a plateau, due to over-parametrization. This suggests that the network simplifies the encoding of class information on its last layers. Consequently, it is safe to regularize the network at these points, as is experimentally verified in Section V-A.

VI. Conclusion

In this paper, we have studied the impact of the audio augmentation-based regularization of the internal layers of a sound event deep classifier, using the framework of DCASE 2023 Task 4A. We have shown experimentally that output regularization is not optimal in this setting, and that proper internal regularization improves the baseline system for this task. Moreover, our results suggest that the optimal placement of this regularization is non trivially related to the diversity of the set of audio augmentations and to the target evaluation metric. Finally we have studied this behavior through the lens of the classifier's implicit pretext tasks, and its latent representation encoding complexity. We believe that the study of these two properties can lead to insight on how a deep classifier solves its target task, how to select the best augmentation strategies, and how to best regularize it.

Acknowledgment

The material contained in this document is based upon work funded by the Agence National de la Recherche en Intelligence Artificielle (PhD program in AI) and Hi! PARIS through its PhD funding program. This work was performed using HPC resources from GENCI–IDRIS (Grant 2021-AD011013406R1).

References

[1] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[2] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[3] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos *et al.*, "A cookbook of self-supervised learning," *arXiv preprint arXiv:2304.12210*, 2023.

[4] C. Lyle, M. van der Wilk, M. Kwiatkowska, Y. Gal, and B. Bloem-Reddy, "On the benefits of invariance in neural networks," *arXiv preprint arXiv:2005.00178*, 2020.

[5] M. Leyton, *Symmetry, causality, mind*. MIT press, 1992.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[7] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 113–123.

[8] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates, Inc., 2015.

[9] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1947–1980, 2018.

[10] D. Bang and H. Shim, "Mggan: Solving mode collapse using manifold-guided training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2347–2356.

[11] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[12] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[13] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[14] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, "Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 857–10 866.

[15] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[16] S. Nian, L. Erfan, and L. Xiaofei, "Rct: Random consistency training for semi-supervised sound event detection," in *Interspeech*, 2021.

[17] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[18] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[20] M. I. Belghazi, A. Baratin, S. Rajeshwar, Y. Ozair, S. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International conference on machine learning*. PMLR, 2018, pp. 531–540.

[21] W. Nie, Y. Zhang, and A. Patel, "A theoretical explanation for perplexing behaviors of backpropagation-based visualizations," in *International conference on machine learning*. PMLR, 2018, pp. 3809–3818.

[22] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[24] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4308–4312.

[25] N. Turpault, R. Serizel, and E. Vincent, "Analysis of weak labels for sound event tagging," Apr. 2021, working paper or preprint. [Online]. Available: https://hal.inria.fr/hal-03203692

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

[27] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 61–65.

[28] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Post-processing independent evaluation of sound event detection systems," *arXiv preprint arXiv:2306.15440*, 2023.

[29] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers," *arXiv preprint arXiv:2307.03183*, 2023.