DEVICE GENERALIZATION WITH INVERSE CONTRASTIVE LOSS AND IMPULSE RESPONSE AUGMENTATION

Lorenz P. Schmidt, Nils Peters

Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
International Audio Laboratories, Erlangen, Germany
{lopa.schmidt, nils.peters}@fau.de

ABSTRACT

Acoustic Scene Classification poses a significant challenge in the DCASE Task 1 TAU22 dataset with a sample length of only a single second. The best performing model in the 2023 challenge achieves an accuracy of 62.7% with a gap to unseen devices of approximately 10%. In this study, we propose a novel approach using Inverse Contrastive Loss to ensure a device class invariant latent representation and a better generalization to unseen devices. We evaluate the interaction of this contrastive learning approach with impulse response augmentation and show the effectiveness for suppressing device related information in the encoder structure. Results indicates that both, contrastive learning and impulse response augmentation, improves generalization to unseen devices. Further the impulse response dataset should have a balanced frequency response to work effectively. Combining contrastive learning and impulse response augmentation yields embeddings with least device related information, but does not improve scene classification accuracy when compared to augmentation alone.

Index Terms— acoustic scene classification, contrastive learning, device impulse response, augmentation, passt, transformer

1. INTRODUCTION

Acoustic Scene Detection plays a vital role in various applications, such as hearing aids [1], smart homes [2], hands-free telephony and biological signal analysis [3]. The objective is to classify an acoustic scene into several, pre-defined, classes, enabling the application of algorithms under varying conditions. For example, the noise suppression and beamforming in hearing aids uses different approaches for a closed room and open-space [1]. With significant progress in recent years, especially with the introduction of data-based models, the DCASE Challenge [4] Task 1 for Acoustic Scene Classification (ASC) attracts a great number of contributions. Recently the focus shifted to resource-aware methods with complexity constraints [4].

One difficulty of low-complexity inference is generalization to unseen devices. The TAU Urban Acoustic Scenes 2022 Mobile dataset [5] poses a considerable challenge with an inference length of only a single second. Furthermore the data is heavily imbalanced towards one recording device. The discrepancy between recording devices arises due to variations in microphone characteristics, frequency responses and other device-specific (possible non-linear) factors that influences the signal captured. The TAU22 dataset contains audio recorded with three real devices (A: Soundman OKM II Classic/Studio A3, B: Samsung Galaxy S7, C: GoPro Hero5 Session) and three simulated devices (S1-S3). Further three unseen devices (S4-S6) are artificially generated for the testing dataset, emphasizing the importance of generalization to unseen devices. The dataset is heavily biased towards device A with 62.5% of all samples, while the remaining 8 devices contain only 37.5% of data.

We evaluate the interaction of contrastive learning and device impulse response augmentation. For our challenge submission [6] we used contrastive learning to improve device generalization. We show that combining both suppresses device related information in the model embedding better, than just using each method individually. We estimate the influence on classification performance for a state-of-the-art Transformer model with the TAU22 dataset.

By applying Inverse Contrastive Learning (ICL) [7] to the problem we encourage the model to learn device invariant representation. We use two device impulse response (DIR) datasets for augmentation. The first contains recordings of 66 vintage microphone impulse responses [8]. The second dataset is generated from 25 professional microphones recorded at different angles and distances [9], amounting in a total of 8138 DIRs.

We first discuss in Section 2 related work when dealing with heavily imbalanced datasets. This includes resampling methods, invariance learning and data augmentation. In our method part, Section 3 and Section 4, we discuss how contrastive learning helps to improve generalization ability and the differences between both device impulse response datasets. In the final Section 5 we explain our experiments, the outcomes and discuss implications for training.

2. RELATED WORK

2.1. Data Imbalance Resampling

The problem of device generalization is part of the broader issue of imbalanced training. A common countermeasure is oversampling of under-representated groups [10] to duplicate minority class samples. On the other hand, undersampling of over-represented removes considered samples during training epochs [11]. These techniques create therefore a more balanced training set and prevent the model to be biased for the majority class. The DCASE ASC dataset is characterised by a stark bias towards a single device (more than 60% of total data, where 16% would be uniformly distributed). Undersampling would prolong training time for the model until the whole dataset is seen at least once.

Another possibility (for avoiding duplicating excessively when oversampling) is synthetic minority oversampling, where new samples are generated by interpolating between existing minority classes [12]. Freq-MixStyle [13] is an instance of this approach by

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS.

mixing frequency statistics of spectrograms, which shows a good generalization performance. Also Adaptive Synthetic Sampling [14] weights likelihood of samples by difficulty and generates minority classes that are harder to learn more often.

2.2. Invariance Learning

The concept of learning invariant representations is closely related to data imbalance in the sense that both address biases in datasets. While data imbalance focuses on unequal distributions in samples, learning invariant representations aims to extract features that are robust to variations introduced by factors in the data. In both cases the aim is reducing the impact of biases and promote generalization.

One example of invariance learning is the Generative Adversarial Network (GAN). An adversarial discriminator infers the device class during training and promotes learning invariant embeddings. The generator and discriminator are trained in tandem, where the generator creates realistic looking audio samples [15].

Other invariance learning methods extend contrastive learning to self-supervised settings. As reported in [16], an online model tries to predict the representation of a target model with an augmented view. This makes representation invariant in view differences and can be applied without any labels of data.

2.3. Data Augmentation

Data augmentation helps to manage imbalanced datasets by increasing the diversity and quantity of samples, therefore improving generalization ability of the model and reducing risk of overfitting. Common examples are SpecAugment [17] introducing random freq/temporal masking and warping, pitch shifting [18], time stretching and noise injection [19] into the data samples.

A simple, yet effective, method for generalizing to new devices is impulse response augmentation. In our case we convolve our training data with measured or simulated device impulse response to create a more diverse and realistic dataset. For ASC the generated data makes the training more robust and resembles a more realistic inference environment. To model non-linear effects, dynamic range compression [20] can simulate the microphone characteristics.

3. CONTRASTIVE LEARNING

The goal of contrastive learning is to find a latent representation where positive pairs are grouped together, while negative pairs are separated. Originally introduced for supervised learning [21], it recently finds extensions to unsupervised and self-supervised settings [22, 23] and application to audio [24]. In our dataset, we have device classes available making supervised methods possible.

In our approach, positive samples are selected from different device classes, and should exhibit a greater similarity on average compared to negative samples. Negative samples are from the same device class and the training process should maximize their dissimilarity. This method is used in ICL to find more mode-collapse robust latent representations, compared to approaches using Kullback-Leibler divergence or Maximum-Mean Discrepancy [7]. ICL utilizes a loss function defined as follows

$$\mathcal{L}_{\text{ICL}} = \mathop{\mathbb{E}}_{\substack{(z,c)\sim p(z,c)\\(\hat{z},\hat{c})\sim p(\hat{z},\hat{c})}} [\mathbb{1}(c=\hat{c}) f(z,\hat{z}) + \mathbb{1}(c\neq\hat{c}) g(z,\hat{z})] \quad (1)$$

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{ICL}} \mathcal{L}_{\text{ICL}} \quad (2)$$

with asymmetric penalties for positive and negative samples

$$g(z,\hat{z}) = d_Z^2(z,\hat{z}) \quad (3)$$

$$f(z,\hat{z}) = \exp((\alpha - d_Z(z,\hat{z}))/\beta), \quad (4)$$

where the threshold α and barrier strength β defines the extent to which latent similarity for the same device classes are penalized. The linear combination with the default cross-entropy loss term $\mathcal{L}_{\text{CE}}$ is controlled by λ_{ICL}. The distance $d_Z(\cdot,\cdot)$ is the ℓ_2 norm for all our experiments. We discuss the selection of barrier parameters α, β in our experiments described in Sec. 5.2.

The objective of ICL is to make training invariant to unwanted variables in the dataset. It can be used to address biases and confounding effects related to demographical variables [25, 26], for example age, gender, income etc. This helps mitigate biases and ensures correct model inference without unwanted side-effects.

In our case, we employ a Transformer model [27] as the encoder structure to project a spectrogram into a lower-dimensional embedding. The encoder is expected to learn meaningful and robust representations that can be utilized for the downstream task of acoustic scene classification. Data augmentation plays a crucial role in training a good generalizing encoder. Furthermore, it can also improve the impact of contrastive learning in two ways, as shown in Sec. 5.4. First, augmentation leads to more device classes, which gives the contrastive learning more positive and negative samples for training. Second, the augmented device classes share some device characteristics with neighbouring classes. This makes the negative sampling more difficult, forcing the model to use a variety of device specific traits in the data. In our case the device of an acoustic scene sample is altered with an additional DIR.

4. DEVICE IMPULSE RESPONSES

In this section, we provide a description of two different datasets of microphone impulse responses that are used for augmenting the ASC TAU22 training set. Their characteristics are quite different.

The first dataset contains recordings of 66 vintage microphones produced by the MicIR project [8]. They are recorded in a booth with the swept-sine method. The source is placed in approximately 20-30cm distance from the microphone. Due to different room reflections, the recordings should not be considered as free-field. As seen in Fig. 1 the vintage DIRs have a frequency dependent variability and pronounced low-pass behaviour for frequencies above 10 kHz. Between the 1 kHz and 10 kHz region the data follows a narrow band in 0.1 to 0.9 quantiles with 20% of data in a wider 20dB variation.

The second dataset contains DIRs of 25 microphones for multiple angles and distances, and is henceforth called Multi DIRs. Incident angles are varied from $0°$ to $355°$ in steps of $5°$ and at source-to-microphone distances of 0.5m, 1.25m and 5m. The microphone is rotated with a computer-assisted turntable. The microphone characteristics include omnidirectional, cardioid, supercardioid and bi-directional polar patterns. The set is quite varied, due to different microphone transduction types (condenser, moving-coil, ribbon), single/dual and small/large diaphragms, and end/side address designs.

The distribution of frequency responses (see Fig. 1 for Multi DIRs) shows a more frequency independent variability of responses when compared to the Vintage DIRs. Further a smaller dip for frequencies above 10 kHz distorts the training data distribution less

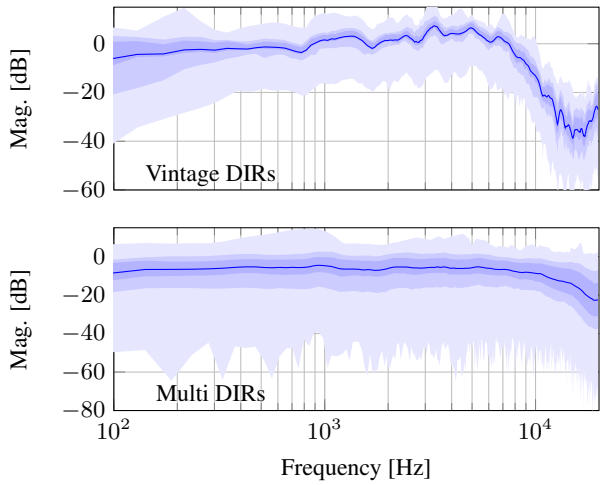

Figure 1: Microphone Frequency Responses for 100% ( ), 80% ( ), 50% ( ) of all data and mean (——) values.

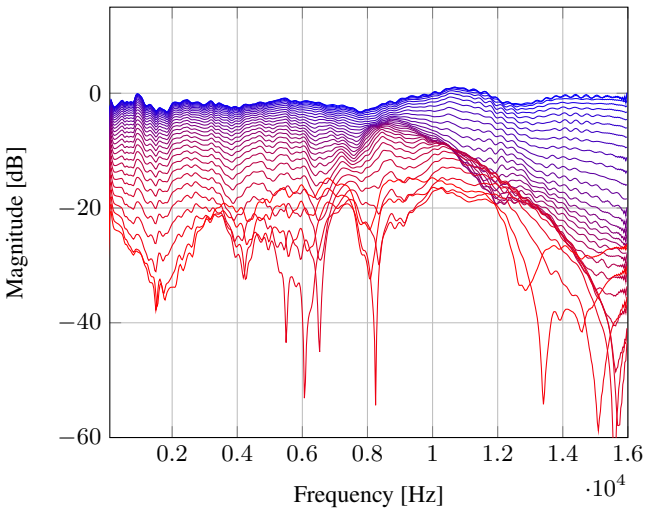

Figure 2: Frequency Responses of the Røde NT2-A cardioid microphone for angles 0° (——) to 180° (——) recorded at a distance of 50cm.

and avoids an inference mismatch. Finally the 0.1-0.9 quantile is wider and exposes more variations to the model during training.

When looking at a specific example (the Røde NT2-A cardioid microphone) over variations in incident angle, a rich pattern can be seen in Figure 2. It follows the characteristics of a cardioid microphone polar pattern, with a deep notch at 180°. This shows a large variability for even a single microphone characteristic.

5. EXPERIMENTS

We evaluate the tandem setting of contrastive learning and DIR augmentation with the TAU22 [5] dataset split to 139,970 samples for training, 29,680 samples for validation and 29,680 samples for testing. They are recorded at a sampling rate of 44.1 kHz in 12 different European cities and 10 acoustic scenes. We first describe how we set-up our model for all our training sessions. Then the specifics of ICL and DIR augmentation are explained and their effects on device invariance and scene classification discussed.

5.1. PaSST Model

We extract Mel-scaled spectrograms with 128 bands from audio subsampled to 32 kHz sampling rate. Individual windows have a length of 800 samples and an overlap of 320 samples. We apply a logarithmic transformation to normalize the feature distribution.

We use the Patchout faSt Spectrogram Transformer (PaSST) [27, 28] as our encoder structure and a single feed-forward layer for classification. The transformer has a patch size of 16, depth of 12 and 12 heads. Furthermore, the embedding dimension is 768, where the classifier projects the final embedding to 10 scene classes. To speed up training, we apply patchout along frequency axis with a rate of 6 patches similar to the PaSST model [27].

To avoid overfitting and improve generalization, the dataset is augmented in the following ways. We merge recordings to 10s snippets and extract randomly sliding windows of 1s during training. We also apply independent frequency masking for 48 bins and do time masking for 24 windows and use randomized frequency cutoff of up to 500Hz. With this we follow the training approach described in PaSST [27], but we do not use mixstyle augmentation to compare augmentation and contrastive learning properly. We apply an Adam optimizer with the same settings as in our challenge submission [6]. An initial learning rate of 0.00042 is gradually reduced on plateau with a patience of 10 epochs and factor of 0.5. We operate the optimizer at $\beta = (0.957, 0.9514)$ and $\epsilon = 0.038$. Each configuration is trained three times for 250 epochs with a batch-size of 64 and we use the best performance in our results.

5.2. Inverse Contrastive Loss

We apply a inverse contrastive loss during training to make classification invariant to device characteristics. The augmented term penalizes latent distances of same device classes. This implies a tradeoff when choosing the hyperparameters for training. The exponential term (see Equation 4) acts as a barrier function for a shifted threshold α, where the strength is controlled by β (with indicator function in limit $\beta \to 0$). We choose $\beta = 1$ for all our experiments and perform grid search for suitable hyper-parameters which results are shown in Table 1. Even though the variables are not independent, we grid-searched them separately. We first fix $\lambda_{\text{ICL}} = 10$ to observe the effect of thresholding on the performance. It exhibits a slight decrease in performance when increasing to $\alpha = 0.2$, while a more drastical degradation when increasing further. Therefore we conclude with this value for the remaining of our experiments. The loss weight λ_{ICL} does not have such a drastic effect on the performance, but increasing too much decreases performance by 2% accuracy. We choose $\lambda_{\text{ICL}} = 0.5$ as a conservative measure. The accuracy improved on the validation dataset compared to $\lambda_{\text{ICL}} = 0.1$, indicating a positive effect of ICL.

5.3. Impulse Response Augmentation

For DIR augmentation, we use two dataset sources [8, 9]. We resample both to 32kHz sampling rate. Further, we window the Multi DIRs dataset [9] to 1024 samples with a Kaiser window ($\beta = 2$).

With this, we train an IR generator, similar to the FAST-RIR [29] diffuse room impulse generator. The model is conditioned on

α	LogLoss	Acc. [%]	λ_{ICL}	LogLoss	Acc. [%]
0.0	1.265	51.72	0.1	1.139	56.87
0.2	**1.270**	51.28	0.5	**1.100**	58.32
0.4	1.313	50.34	1.0	1.113	58.52
0.8	1.331	49.02	3.0	1.139	56.87
1.5	2.166	21.60	6.0	1.183	56.30

Table 1: Results for different α and λ_{ICL} values. We fixed $\lambda_{\text{ICL}} = 10.0$ for the α search and $\alpha = 0.2$ for the λ_{ICL} search. Based on the results we choose $\alpha = 0.2$, $\lambda_{\text{ICL}} = 0.5$ for further experiments.

Method	Accuracy [%]
PaSST	82.04
+ ICL	65.65
+ Multi DIRs	41.08
+ ICL + Multi DIRs	**17.53**

Table 2: Device classification accuracy results for the embedding of a PaSST model with different generalization methods (see Sec 5.4). Lower accuracy indicates better invariance to device class.

the microphone characteristics (1) directivity (2) transducer (3) diaphragm properties and angle/distance in cartesian coordinates in total of 12 variables. We train the generator in the same GAN framework as the original method with a final MSE of 0.00527. Unfortunately the approximately 8000 samples of Multi DIR are not sufficient for training a microphone impulse response model. We see good generalization for varying incident angles, but not for source distance and new synthetic device classes. Applying the generator to our ASC model gives only a best log-loss of 1.56 and we drop it therefore for our next comparison.

5.4. Device Related Latent Information

As an additional study we measure the device related information during training. We create a separate device classifier with the same capacity as the acoustic scene classifier and train it with the default Adam optimizer until convergence. Since the device class is imbalanced, we use a balanced cross-entropy term as our loss measure.

The results in Table 2 are evaluated on the validation set for the 6 devices of the training set. Because the device occurrence is balanced, random guess is set at $1/6$.

The use of contrastive learning does not lower device accuracy as much as impulse response augmentation. A possible explanation for this is that we can use augmentation aggressively, while use of contrastive learning has a negative effect on training (see Table 1). Further augmentation adds variability to the dataset and does not necessarily inhibit the primary task.

Interestingly, combining augmentation and contrastive learning reduces device accuracy further to the points of random guess. This indicates, that the latent space for acoustic scene classification does not have device-related information. When looking at the final results in Table 3, on the other hand, the results are still biased towards the more common devices. To illustrate, see that the acoustic scene classifier benefits from a robust latent representation. Even though we have minimized device related information the encoder still generalizes the spectrogram for the majority class better. To mitigate this effect we would have to resample to even class distribution, for example with synthetic augmentation (see related work in Sec. 2). Another possibility is that the device classify is too shallow to model the benefiting factors for the scene classification, even though they have the same capacity.

5.5. Acoustic Scene Classification Results

As the final experiment we train the PaSST model with the illustrated four different settings for device generalization. We see a large gap of 0.3 log-loss between real devices and simulated/unseen devices in Table 3 for the vanilla PaSST model - with device A best performing of 1.012 log-loss.

The vintage DIR augmentation improves the performance for unseen devices, but degrades that of real devices. This gives a worse overall performance. The multi angle DIR dataset on the other hand improves performance for all three device families, with the largest improvement in unseen devices of approximately 0.1 log-loss. When applying contrastive learning we see a similar effect, but not as pronounced as the impulse response augmentation. Further the performance for real devices suffer slightly.

Finally combining impulse response augmentation with contrastive learning improves performance slightly, compared to contrastive learning alone. On the other hand, it does not improve performance when comparing to Multi DIRs augmentation alone.

6. CONCLUSION

To summarize, contrastive learning makes latent space invariant to device classes and improves generalization. To that effect, impulse response augmentation works better, but best device invariance is achieved by combining both methods. The Multi DIRs shows a greater variability and less bias for frequency responses and works better for data augmentation when compared to the Vintage DIR. In the final ASC experiment, contrastive learning improves log-loss, but is outperformed by applying proper data augmentation alone. Nevertheless, contrastive learning can be advantageous compared to domain specific augmentation, especially when the training is only affected by data imbalance and not by unseen classes or no effective augmentation technique is available.

Method	Real Devices				Simulated Devices				Unseen Devices				Overall
	A	B	C	Avg.	S1	S2	S3	Avg.	S4	S5	S6	Avg.	
PaSST	1.012	1.266	1.070	1.116	1.371	1.492	1.326	1.396	1.401	1.36	1.509	1.423	1.181
+ Vintage DIRs	1.082	1.360	1.212	1.218	1.462	1.449	1.361	1.424	1.343	1.289	1.557	1.396	1.212
+ Multi DIRs	**0.979**	1.221	1.090	1.097	**1.347**	1.427	**1.318**	**1.364**	**1.277**	**1.302**	**1.425**	**1.334**	**1.139**
+ ICL	1.021	1.297	**1.074**	1.131	1.375	1.465	1.325	1.388	1.284	1.364	1.467	1.372	1.167
+ ICL + Multi DIRs	1.030	**1.190**	1.139	**1.096**	1.372	**1.412**	**1.318**	1.367	1.302	1.327	1.474	1.368	1.156

Table 3: Log-loss validation performance of the proposed methods on the TAU Urban Acoustic Scenes 2022 Mobile dataset [5] with provided split. The PaSST model is trained for three different seeds and best performance is picked. The validation results are grouped into real devices (A, B, C), simulated devices (S1, S2, S3) and unseen devices (S4, S5, S6) and averaged values given to compare device families.

7. REFERENCES

[1] V. Vivek, S. Vidhya, and P. Madhanmohan, "Acoustic scene classification in hearing aid using deep learning," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, 2020, pp. 0695–0699.

[2] S. Krstulović, "Audio event recognition in the smart home," *Computational Analysis of Sound Scenes and Events*, 2018.

[3] A. I. Humayun, S. Ghaffarzadegan, M. I. Ansari, Z. Feng, and T. Hasan, "Towards domain invariant heart sound abnormality detection using learnable filterbanks," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2189–2198, aug 2020. [Online]. Available: https://doi.org/10.1109%2Fjbhi.2020.2970252

[4] I. Martín-Morató *et al.*, "Low-complexity acoustic scene classification in dcase 2022 challenge," 2022. [Online]. Available: https://arxiv.org/abs/2206.03835

[5] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 56–60. [Online]. Available: https://arxiv.org/abs/2005.14623

[6] L. Schmidt, B. Kiliç, and N. Peters, "Submission to DCASE 2023 task 1: Device invariant training with structured filter pruning for low complexity acoustic scene classification," DCASE2023 Challenge, Tech. Rep., May 2023.

[7] A. K. Akash, V. S. Lokhande, S. N. Ravi, and V. Singh, "Learning invariant representations using inverse contrastive loss," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6582–6591.

[8] Xaudia, "Microphone impulse response project," accessed 2023-06-14. [Online]. Available: http://micirp.blogspot.com/

[9] J. C. Franco Hernández, B. Bacila, T. Brookes, and E. De Sena, "A multi-angle, multi-distance dataset of microphone impulse responses," *J. Audio Eng. Soc*, vol. 70, no. 10, pp. 882–893, 2022. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=22014

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002. [Online]. Available: https://doi.org/10.1613%2Fjair.953

[11] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, p. 20–29, jun 2004. [Online]. Available: https://doi.org/10.1145/1007730.1007735

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[13] B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," *arXiv preprint arXiv:2206.12513*, 2022.

[14] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328.

[15] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," 2019.

[16] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," 2020.

[17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*. ISCA, sep 2019. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2019-2680

[18] H. Su, H. Zhang, X. Zhang, and G. Gao, "Convolutional neural network for robust pitch determination," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 579–583.

[19] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[20] H. Hu *et al.*, "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," DCASE2020 Challenge, Tech. Rep., June 2020.

[21] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 539–546 vol. 1.

[22] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," 2022.

[23] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," 2021.

[24] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3875–3879.

[25] V. S. Lokhande, R. Chakraborty, S. N. Ravi, and V. Singh, "Equivariance allows handling multiple nuisance variables when analyzing pooled neuroimaging datasets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 432–10 441.

[26] Z. Cao, H. Yu, H. Yang, and A. Sano, "Pirl: Participant-invariant representation learning for healthcare," 2022.

[27] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Interspeech 2022*. ISCA, sep 2022. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2022-227

[28] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "Fast-rir: Fast neural diffuse room impulse response generator," 2022.