Uncertainty in Semi-supervised Audio Classification – A Novel Extension for FixMatch

Sascha Grollmisch
Industrial Media Applications
Fraunhofer IDMT
Ilmenau, Germany
goh@idmt.fraunhofer.de

Estefanía Cano
Songquito UG
Erlangen, Germany
estefania.cano@songquito.com

Hanna Lukashevich
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany
lkh@idmt.fraunhofer.de

Jakob Abeßer
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany
abr@idmt.fraunhofer.de

***Abstract*—Semi-supervised learning (SSL) is a commonly used technique when annotated data is scarce but unlabeled data is easily available. In recent years, SSL has seen a large boost in the computer vision domain and methods such as FixMatch were successfully adapted to audio classification tasks. However, there still remains a gap between SSL methods and the fully supervised baselines, which were trained with all labels available. In this work, we first investigate the quality of the pseudo-labels, i. e., generated labels for unlabeled data, for musical instrument family classification and acoustic scene classification. Based on these insights, we propose and evaluate a novel extension of FixMatch that quantifies and considers the uncertainty of the pseudo-labels. Additionally, we highlight the problematic trade-off between pseudo-label quality and quantity. Our results show that Monte-Carlo Dropout combined with temperature scaling improved the pseudo-label accuracy from 78.4% to 86.7% for instrument family and from 87.9% to 89.9% for acoustic scene classification. Even though the accuracy on the test sets improved from 71.0% to 72.1% and from 69.2% to 70.8%, respectively, there is still a gap to the fully supervised baseline leaving room for future work.**

***Index Terms*—semi-supervised learning, deep learning, music information retrieval, acoustic scene classification**

I. Introduction

Annotated data for classification tasks is often difficult and expensive to obtain whereas unlabeled data is readily available or can be acquired at relatively low cost. Semi-supervised learning (SSL) reduces the amount of required annotated data for training a model by incorporating unlabeled data. In recent years, SSL has seen great advances in image classification tasks with algorithms such as ReMixMatch (RMM) [1], FixMatch (FM) [2], and FlexMatch [3], to name a few. These methods nearly closed the performance gap to the fully supervised baseline, that had all annotations available, using only a fraction of the labeled data. As an example, FM achieved 95.7% accuracy on *CIFAR-10* [4] using 400 examples per class and an only slightly reduced performance of 94.9% with 25 examples per class. In our previous work [5], we adapted FM for audio classification and evaluated it on classification tasks from different audio domains such as Acoustic Scene Classification (ASC), Music Information Retrieval (MIR), and Industrial Sound Analysis (ISA). With few labeled examples, FM outperformed models trained supervised from scratch, the SSL method Mean Teacher [6], and transfer learning. However, it did not reach the accuracy of the fully supervised baseline for ASC and MIR when the amount of labeled data was reduced to 5% or less. Similar findings were made by Cances et al. [7] for Sound Event Detection (SED) and Speech Command Recognition (SCR) where RMM and FM outperformed models trained supervised with 10% of the labeled data but SSL did not reach the accuracy of the fully supervised baseline on one of the datasets. In [8], FM and RMM achieved competitive results in the fields of computer vision, natural language processing, and audio classification compared to more recent SSL methods such as FlexMatch. In their work, the largest gap to the fully supervised results was assessed for audio classification even though pre-trained networks were used. This result was mainly attributed to inputting raw audio instead of spectral images. Following these results, we focus on FixMatch as a semi-supervised audio classification approach, and build upon our previous work [5] to improve SSL for audio classification.

The training procedure for audio classification using FM as proposed in [5] is shown in Figure 1. The audio files are transformed into (log) Mel spectrograms patches as input for a Convolutional Neural Network (CNN). These "spectral images" are augmented with common image augmentation techniques such as translation in x/y direction, additive Gaussian noise, and sharpening. For the labeled examples, the categorical cross-entropy (CCE) loss is calculated as one part of the loss used to train the CNN. For the unlabeled data, the model itself iteratively generates the target labels or *pseudo-labels* during training time: The unlabeled data in each batch is weakly augmented, passed through the model and only those examples with a highest class-wise output prediction above a pre-defined *confidence threshold* are binarized to pseudo-labels. Using this filtering step, we only keep pseudo-labels that the network is confident about. The same unlabeled examples are also strongly augmented. The pseudo-labels are then used as targets to calculate the CCE loss for the unlabeled data. The final training loss is computed as the unweighted sum between the labeled loss and the unlabeled loss.

It is clear from the FM procedure that the correctness of the generated pseudo-labels is of critical importance for the

This work has been supported by the German Research Foundation (AB 675/2-2).

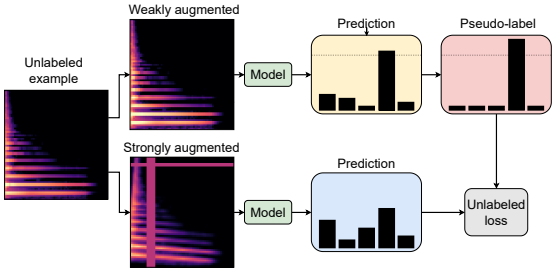


Fig. 1: Unlabeled part of the FM training process for time-frequency representations of audio data in [5]. The model generates a pseudo-label on a weakly augmented version of an unlabeled example and uses it as target for a strongly augmented version.

final performance of the system. Errors can amplify during the course of training and negatively impact performance [9]. This effect should be limited with the confidence threshold mechanism in FM. More recent FM variants such as Dash [10] or FlexMatch [3] tried to improve the confidence thresholding step in FM by applying class-dependent or dynamically changing threshold values. Still, neural networks can output wrong pseudo-labels with a high confidence [11]–[14]. We hypothesize that incorrect pseudo-labels with high confidence are one reason why FM and its variants fail to achieve the results of the fully supervised baseline on some of the investigated datasets.

The contributions of this work are as follows: We first investigate the quality of the pseudo-labels for one MIR dataset and one ASC dataset. We then propose and systematically evaluate a method that quantifies and considers the uncertainty of pseudo-labels in FM. Lastly, we investigate the impact of the proposed FM modification regarding the classification accuracy on the test data.

II. Related work

FM uses a confidence threshold to filter out pseudo-labels about which the model is uncertain. However, the output of the last softmax layer is often erroneously interpreted as the confidence of a model with respect to the class decision. This interpretation problem is called *deterministic overconfidence* [11]. As a result, not only the correct but also erroneous class decisions are getting high softmax outputs. This deterministic overconfidence is large when the data is far away from the model's decision boundary. It is also high when rectified linear units (ReLU) are used in the network [13].

There are multiple solutions to overcome the deterministic overconfidence. In the *temperature scaling* approach [12], the softmax outputs are calibrated post-hoc in order to soften the output predictions. This is achieved by dividing the output logits of the neural network by a fixed temperature T as follows

$$softmax_{Temp}(f_i) = \frac{exp(f_i/T)}{\sum_{j=1}^{J} exp(f_j/T)} \quad (1)$$

where J is the total number of classes and a_i are the logits for each class. Temperature scaling mitigates the deterministic overconfidence when data can be considered in-distribution.

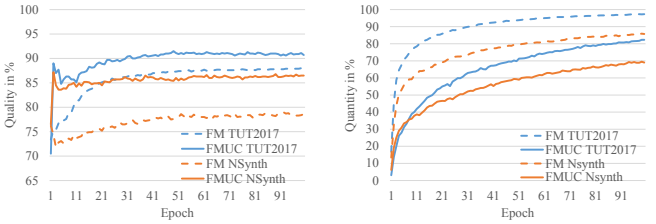
Other approaches approximate the Bayesian inference using the concepts of *aleatoric* and *epistemic* uncertainty. Aleatoric uncertainty is caused by inherent noise in the data and epistemic uncertainty is caused by missing knowledge in the model [15]. In the *Monte Carlo (MC)-Dropout* approach, dropout layers are added to the neural network to model the epistemic uncertainty in an efficient way [11]. During inference, the dropout layers are still active and the same input is passed several times through the neural network. Epistemic uncertainty can also be estimated using *deep ensembles* of multiple independently trained neural networks [16].

To measure the uncertainty of the pseudo-labels in SSL, Rizve et al. [14] proposed the Uncertainty-Aware Pseudo-Label Selection (UPS) framework. They used only predictions with low uncertainty to reduce the effect of poor neural network calibration. Furthermore, they applied temperature scaling with a temperature of $T = 2$ combined with MC-Dropout to compute the uncertainty of predictions from the output of the temperature-scaled softmax layer over all passes through the network. One filtering mask only includes pseudo-labels with an average prediction above 0.7 and a second mask only the ones with a standard deviation below 0.05. In each epoch, the pseudo labels to be included are obtained from the intersection of both masks. In contrast to FM, UPS uses two neural networks, the pseudo-labels are generated once per epoch anew, and the neural networks are re-initialized after each pseudo-labeling step. The reported results for UPS were below the ones from RMM.

III. Datasets

Building up upon our previous work [5], we focus on two datasets from two different audio domains. The first dataset is the TUT Urban Acoustic Scenes 2017 (*TUT2017*) for acoustic scene classification (ASC) with its corresponding training [17] and test splits [18]. Each of the ten seconds long audio clips was recorded in one of 15 acoustic scenes including grocery store, park, library, etc. The recording locations differ between training and test set. The fully supervised baseline achieved 77.2% file-wise accuracy while FM obtained 69.2% with 5% of the labeled data[1]. The second dataset is the *NSynth* dataset [19] for instrument family recognition with instruments belonging to 11 instrument families such as brass, strings, and vocals. The audio snippets are four seconds long and include synthesized instrument notes. The training and test sets cover different instruments from the same families. With 1% of the labeled data, the classification accuracy of FM dropped from 77.1% (fully supervised baseline) to 71.0%. Both datasets resulted in comparable performance gaps between FM and the fully supervised baselines. Therefore, we use these datasets

[1]In [5], a file-wise accuracy of 61.2% was reported. This was due to an error which excluded three quarters of the unlabeled data (only the first part of each recording was used). We report the corrected results here.



(a) Pseudo-label quality (b) Pseudo-label quantity

Fig. 2: Pseudo-label quality and quantity for each training epoch for FM and FMUC for both investigated datasets.

throughout this work. For details on the state of the art, transfer learning, and few-shot techniques with these datasets see [5].

To enable comparability with our previous work [5], we use the same processing pipeline: From the audio recordings, (log) Mel spectrograms are extracted and several time frames form the spectral images. The extraction parameters for each dataset are detailed in [5]. The spectral images are normalized to zero mean and unit variance for each Mel band before being input to the CNN, namely the `CNN420`. It has a ResNet-based [20] architecture with 420k trainable parameters. Each batch consists of 32 labeled and 224 unlabeled examples. We use the Adam optimizer with a learning rate of 10^{-3} and a fixed number of 500 iterations per epoch. The `CNN420` is trained for 100 epochs on *TUT2017* and for 200 epochs on *NSynth* since the dataset is larger.

IV. Investigating Pseudo-Label Quality

We first measure the quality of the pseudo-labels to prove the hypothesis that they are one of the reasons for the gap to the fully supervised results. We define *quality* as the accuracy of the filtered and binarized pseudo-labels above the confidence threshold. We can compute the accuracy of the pseudo-labels since we know the true labels while artificially reducing the amount of labeled examples in our experiments.

The quality of the pseudo-labels in each training epoch is shown in Figure 2a. For both datasets, it starts at around 75%, meaning that one in every four pseudo-labels is wrong. Over the course of the training, the quality converges at around 88% for *TUT2017* and 78% for *NSynth*. Interestingly, FM can recover from mistakes that were made in early epochs and steadily improve the quality to an upper ceiling.

To get a better understanding of the type of errors that are made when generating the pseudo-labels, Figure 3 illustrates the confusion matrix for *TUT2017* after 20 epochs of training. Interestingly, most of the pseudo-label errors make sense: forest path is confused 20.9% of the times with park, grocery store is confused 26.1% of the times with cafe/restaurant, and train is confused 25.2% of the times with bus. The model did not make arbitrary but rather understandable mistakes which indicate that the basic concept of the task is learned.Still, there is potential for improving the pseudo-label quality, and therefore the FM performance.

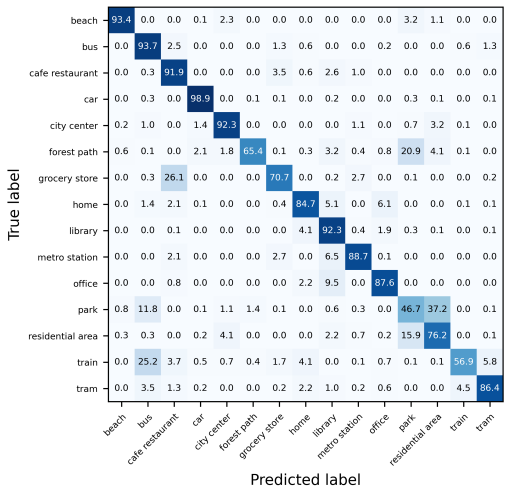


Fig. 3: Confusion matrix of the pseudo-labels after 20 epochs on *TUT2017*.

V. Including Uncertainty to Pseudo-Label Selection

A. Method

We propose to integrate uncertainty measurements based on MC-Dropout and temperature scaling into FM in order to improve the selection and quality of the pseudo-labels. Since we focus on a lightweight approach, we did not further consider deep ensembles as they require to train several models. The main metric for this experiment is the quality of the selected pseudo-labels. As a second metric, we report the *quantity* in % where 100% means that all unlabeled data in each batch is used. A perfect pseudo-label quality is not the best choice when only very few unlabeled examples are considered and the benefit of SSL can be actually neglected.

We illustrate the proposed FM approach with uncertainty measures as *FixMatch-Uncertainty (FMUC)* in Figure 4. Every weakly augmented training example is repeated N times per batch. The weakly augmented examples are passed through the neural network with MC-Dropout. The mean and standard deviations of the temperature scaled softmax output are calculated over N repetitions of each weakly augmented example. The standard deviation for each example is subtracted from its mean prediction and only examples above the pre-defined threshold are used. We propose this combination of both values to avoid an additional threshold which needs to be optimized. The filtered and binarized pseudo-labels are used as targets for the strongly augmented version as in FM.

In order to add MC-Dropout to a CNN, dropout layers are required. The `CNN420` already contains dropout layers for regularization. The architecture is ResNet-based [20] with modifications proposed by Chen et al. in [21]. Each ResNet-block starts with ReLU activation, followed by Batch Normalization and dropout of 0.1 before being passed to the convolutional layer. These dropout layers allow us to include MC-Dropout without changing the model and affecting its

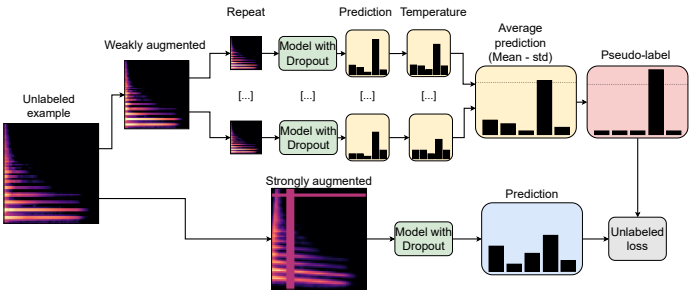


Fig. 4: Extended training process of the unlabeled data part for FMUC. The weakly augmented unlabeled examples are passed several times through the model with dropout enabled. The mean and standard deviation of the temperature scaled predictions are binarized to pseudo-labels for the strongly augmented version when the threshold is exceeded.

overall performance. In this experiment, the repetitions for MC-Dropout are set to 5. We were not able to investigate larger numbers of repetitions due to memory restrictions.

The temperature is set to the suggested value of 2 as in UPS. The influence of the temperature scaling is evaluated separately by changing the temperature value to 1 which turns off the scaling and 4 which softens the predictions further. Another way to increase the accuracy of the pseudo-labels in FM could be raising the confidence threshold, e. g., from 0.95 to 0.99. Even though this performed worse in the original publication [2], we add it for comparison. Based on preliminary experiments, the default confidence threshold was lowered for FMUC from 0.95 to 0.9 due to the stronger filtering introduced with the uncertainty measures. However, a threshold of 0.95 is additionally investigated. Furthermore, we change FMUC to using two different thresholds for mean and standard deviation of the confidence outputs similar to UPS to see if the proposed simplified version using only one combined threshold leads to comparable results. Here, we apply the suggested thresholds from [14] of 0.7 for mean and 0.05 for standard deviation.

B. Results

The results of FM and FMUC with different parameters are shown in Table I. For both datasets, over 90% of unlabeled data is included in FM with a pseudo-label accuracy of 78.4% for *NSynth* and 87.9% for *TUT2017*. Raising the confidence threshold of FM to 0.99 slightly increases the pseudo-label quality while reducing the quantity by a larger margin. This demonstrates the quality-quantity trade-off in SSL.

FMUC improves the quality by 8.3 percentage points for *NSynth* and by 2 percentage points *TUT2017* showing the benefit of uncertainty measurements in pseudo-label selection. Increasing the confidence threshold for FMUC to 0.95 only improves the pseudo-label quality on *TUT2017* while reducing the quantity for both datasets. Therefore, the threshold of 0.9 seems a better choice and is used in further experiments. Using two separate thresholds as in UPS leads to similar results compared to FMUC on *TUT2017* but lowers the quality

TABLE I: Pseudo-label quality and quantity in % for FM and different FMUC versions including different confidence thresholds (thresh.), two separate thresholds as in UPS (FMUC two thresh.), temperature scaling of 4 (temp. 4) or disabled (w/o temp.), and MC-Dropout disabled (w/o MC-Dr.).

Method	NSynth quality	NSynth quantity	TUT2017 quality	TUT2017 quantity
FM	78.4	**91.1**	87.9	**97.3**
FM thresh. 0.99	80.7	75.8	89.9	90.8
FMUC	86.7	78.1	89.9	88.7
FMUC thresh. 0.95	85.1	69.2	91.9	72.8
FMUC two thresh.	84.1	77.5	90.4	87.9
FMUC w/o temp.	77.2	90.1	89.3	94.1
FMUC w. temp. 4	**92.3**	34.1	**94.0**	38.6
FMUC w/o MC-Dr.	82.6	78.8	89.8	89.2

and quantity on *NSynth*. The proposed combined threshold of FMUC removes one hyperparameter without sacrificing any performance and is therefore a useful simplification. Temperature scaling is very effective for increasing the pseudo-label quality as can be seen at "FMUC w/o temp." and "FMUC w. temp. 4". This indirectly confirms the overconfidence of the neural network since many incorrect pseudo-labels with a high confidence are excluded thanks to temperature scaling. The highest pseudo-label quality can be obtained by setting the temperature to 4. At the same time, the pseudo-label quantity dropped below 40% for both datasets. Therefore, the number of epochs needs to be increased too much for being feasible in this study. Furthermore, we kept a temperature value of 2 as it is not guaranteed that the quantity is not converging to lower values. Disabling MC-Dropout ("FMUC w/o MC-Dr") lowers only the pseudo-label quantity on *NSynth*. For *TUT2017* the influence of MC-Dropout is negligible which might be caused by using only 5 repetitions.

For deeper insights into the effect of the added uncertainty, Figure 2 shows how the pseudo-label quality and quantity for FM and FMUC progresses over the training duration. The improvement of the quality using the proposed uncertainty approach is clearly visible, see Figure 2a. In early epochs, the improved filtering leads to a constantly higher pseudo-label accuracy. Therefore, FMUC is a valid strategy for countering badly calibrated networks in the beginning of the training. On the downside, FMUC is still bound to the quality-quantity trade-off requiring longer training times, see Figure 2b.

VI. Full Evaluation

Even though the quality of the pseudo-labels improved, FMUC needs to be compared to FM and the fully supervised baseline regarding the file-wise accuracy on the test set. All experiments are repeated three times to account for randomness in the CNN initialization and data selection when reducing the amount of labeled data. The number of training epochs is doubled for both datasets since the amount of included unlabeled data did not fully converge. The default FM results are recalculated with the current number of iterations and epochs. Additionally, the file-wise accuracy is calculated on

TABLE II: Mean file-wise accuracy and standard deviation of FM and FMUC in % on the separated test sets and left-out unlabeled data plus the results of the fully supervised baseline.

Dataset	FMUC	FM	Fully supervised
TUT2017 test	70.8±0.4	69.2±0.3	77.2±1.2
TUT2017 unlabeled	88.4±0.6	84.9±0.7	-
NSynth test	72.1±0.7	71.0±1.9	77.1±0.2
NSynth unlabeled	76.5±1.1	74.0±1.9	-

the unlabeled data to assess the performance on data with the exact distribution as the labeled training data.

As shown in Table II, the accuracy on the test data improved by 1.6 percentage points for *TUT2017* and 1.1 percentage points for *NSynth* with FMUC. It reduces the gap to the fully supervised baseline that has all labels available. However, this improvement is less than the gain in pseudo-label quality. The difference can be explained by overfitting to the training domain (exact instruments and recordings locations) for the investigated datasets. This is evident by looking at the accuracy on the unlabeled data which increased by 2.5 to 3.5 percentage points.

VII. Conclusion

In this work, we first investigated the pseudo-label quality of FixMatch (FM) for audio classification. Instrument family recognition and acoustic scene classification were selected as tasks to cover a wide variety of audio content. We demonstrated that the pseudo-label quality is not perfect, leading to wrong targets for Semi-supervised Learning (SSL). These errors are one possible reason for the gap between SSL and the fully supervised baseline with all labels available. To improve the quality of the pseudo-labels, we propose and evaluate a novel extension of FM which combines temperature scaling and MC-Dropout to measure the uncertainty of the pseudo-labels, called FixMatch-Uncertainty (FMUC). It increases the pseudo-label quality while the quantity is not lowered by the same amount. Temperature scaling was the most important component to filter out incorrect pseudo-labels while the positive impact of MC-Dropout was dataset-dependent.

FMUC reduces the gap to the fully supervised baseline for both datasets and the proposed changes can be integrated into other FM variants with little effort. On the downside, FMUC cannot completely overcome the quality-quantity trade-off, meaning that longer training times are required for a performance gain. While the accuracy of the pseudo-labels already increased in the early stages of the training, the final performance on the test data was improved less. One reason is the overfitting to the training domain which needs to be approached differently. Another reason is the imperfect quality of the pseudo-labels with confusions between semantically close classes. Pre-training the neural networks or active learning strategies with human supervision could be promising directions for future work.

Nevertheless, the proposed FM changes may have a stronger impact for out-of-domain problems. The current experimental setup of artificially reducing the number of labeled data represents an ideal version of SSL. In real-world use cases it is likely that not all unlabeled data will contain the desired classes or be from the same distribution as the labeled examples. Here, measuring the uncertainty already proved its suitability [22] and might therefore be more valuable for SSL in real-world applications.

References

[1] D. Berthelot *et al.*, "ReMixMatch: Semi-supervised Learning with distribution matching and augmentation anchoring," in *ICLR*, Online, 2020.

[2] K. Sohn *et al.*, "FixMatch: Simplifying Semi-supervised Learning with consistency and confidence," in *NeurIPS*, Online, 2020, pp. 596–608.

[3] B. Zhang *et al.*, "Flexmatch: Boosting Semi-supervised Learning with curriculum pseudo labeling," in *NeurIPS*, Online, 2021.

[4] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[5] S. Grollmisch and E. Cano, "Improving Semi-supervised Learning for audio classification with FixMatch," *Electronics*, vol. 10, no. 15, 2021.

[6] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NIPS*, Long Beach, California, USA, 2017, pp. 1195–1204.

[7] L. Cances, E. Labbé, and T. Pellegrini, "Comparison of semi-supervised deep learning algorithms for audio classification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 23, 2022.

[8] Y. Wang *et al.*, "USB: A unified semi-supervised learning benchmark for classification," in *NeurIPS Datasets and Benchmarks Track*, Online, 2022.

[9] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep Semi-supervised Learning," in *IJCNN*, Glasgow, UK, 2020.

[10] Y. Xu *et al.*, "Dash: Semi-supervised learning with dynamic thresholding," in *ICML*, Online, 2021.

[11] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016, p. 1050–1059.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*, 2017, p. 1321–1330.

[13] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in *CVPR*, 2019.

[14] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for Semi-supervised Learning," in *ICLR*, Online, 2021.

[15] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[17] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Acoustic scenes 2017, development dataset," Website https://zenodo.org/record/400515, last accessed 23/08/2022, 2017.

[18] ——, "TUT Acoustic scenes 2017, evaluation dataset," Website https://zenodo.org/record/1040168, last accessed 23/08/2022, 2017.

[19] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *ICML*, Sydney, Australia, 2017, pp. 1068–1077.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for image recognition," in *CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.

[21] G. Chen, P. Chen, Y. Shi, C.-Y. Hsieh, B. Liao, and S. Zhang, "Rethinking the usage of Batch Normalization and Dropout in the training of deep neural networks," *arXiv preprint arXiv:1905.05928*, 2019.

[22] Y. Ovadia *et al.*, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems*, vol. 32, 2019.