


EMPIRICAL RESEARCH

Open Access



Comparison of semi-supervised deep learning algorithms for audio classification

Léo Cances¹, Etienne Labbé¹ and Thomas Pellegrini^{1,2*} 

Abstract

In this article, we adapted five recent SSL methods to the task of audio classification. The first two methods, namely Deep Co-Training (DCT) and Mean Teacher (MT), involve two collaborative neural networks. The three other algorithms, called MixMatch (MM), ReMixMatch (RMM), and FixMatch (FM), are single-model methods that rely primarily on data augmentation strategies. Using the Wide-ResNet-28-2 architecture in all our experiments, 10% of labeled data and the remaining 90% as unlabeled data for training, we first compare the error rates of the five methods on three standard benchmark audio datasets: Environmental Sound Classification (ESC-10), UrbanSound8K (UBS8K), and Google Speech Commands (GSC). In all but one cases, MM, RMM, and FM outperformed MT and DCT significantly, MM and RMM being the best methods in most experiments. On UBS8K and GSC, MM achieved 18.02% and 3.25% error rate (ER), respectively, outperforming models trained with 100% of the available labeled data, which reached 23.29% and 4.94%, respectively. RMM achieved the best results on ESC-10 (12.00% ER), followed by FM which reached 13.33%. Second, we explored adding the mixup augmentation, used in MM and RMM, to DCT, MT, and FM. In almost all cases, mixup brought consistent gains. For instance, on GSC, FM reached 4.44% and 3.31% ER without and with mixup. Our PyTorch code will be made available upon paper acceptance at <https://github.com/Labbeti/SSLH>.

Keywords: Audio classification, Semi-supervised deep learning, Wide-ResNet

1 Introduction

Semi-supervised learning (SSL) aims to reduce the dependency of deep learning systems on labeled data by integrating unlabeled data during the learning phase. It is essential since the conception of a large labeled dataset is expensive, dependent on the task to be learned, and time-consuming. On the contrary, the acquisition of unlabeled data is cheaper and quicker regardless of the task to perform. Using unlabeled data while maintaining high performance can be done in three different ways: (i) *consistency regularization* [1, 2], which encourages a model to produce consistent prediction whereas the input is perturbed; (ii) *entropy minimization* [3–5], which encourages the model to output high confidence predictions on unlabeled files; and (iii) *standard regularization*

by using weight decay [6, 7], mixup [8], or adversarial examples [9]. The most direct approach for SSL is pseudo-labeling [5], but since then, many new and better approaches came out such as Mean Teacher (MT) [10], Deep Co-Training (DCT) [11], MixMatch (MM) [12], ReMixMatch (RMM) [13], and FixMatch (FM) [14].

In previous work [15], we compared MT and DCT for the task of audio tagging (AT), a classification task that consists of automatically assigning an audio event label to an audio recording. Both approaches use two neural networks during training. In the present article, we extend our comparison by adapting to AT the three single-model SSL methods MM, RMM and FM. One difficulty lies in choosing which audio data augmentation techniques to use, that work for different types of sound events and spoken words [16]. The augmentations used on images for object recognition, such as flips and rotations, are most often not relevant for audio data. We compare the error rates on three audio datasets with different scopes

*Correspondence: thomas.pellegrini@irit.fr

¹ IRIT, University of Toulouse, CNRS, Toulouse, France
Full list of author information is available at the end of the article

and sizes: (i) Environmental Sound Classification 10 (ESC-10) [17], with audio event categories such as dog barking and helicopter; (ii) UrbanSound8k (UBS8K) [18], more specific to urban noises such as car horns, sirens and street music; and (iii) Google Speech Commands v2 (GSC) [19], containing spoken words exclusively.

In MM and RMM, a successful data augmentation technique called mixup [8] is used. It consists of mixing pairs of samples, both the data samples and the labels with a random coefficient. We propose to add mixup to the three other SSL approaches, namely MT, DCT, and FM, which do not already use it. The results reported in this article will highlight the positive impact of mixup in almost all our experiments.

The article contributions are mainly twofold: (i) the application and comparison of several recent SSL methods for audio tagging on three different datasets and (ii) the modification of these methods with the integration of mixup, which resulted in systematic error rate reductions. We shall see that in most cases, MM outperformed the other methods, closely followed by FixMatch+mixup.

The structure of the paper is as follows. Section 3 describes the augmentations we used and the mixup mechanism at the core of the present work. Section 4 describes the five SSL methods; Section 5 presents the experimental settings; and finally, Section 6 presents and discusses the results.

2 Related work

Semi-supervised learning (SSL) is a well-known machine learning setting, for which a lot of research has been conducted, before the rise in popularity of deep learning [20, 21]. In this work, we explore recent SSL approaches that were proposed in the framework of deep learning, since we use deep neural networks as state-of-the-art classifiers for audio tagging. These new approaches, as we shall see, were driven by the simplicity of incorporating unsupervised loss terms into the cost functions of neural networks [22].

2.1 Semi-supervised deep learning taxonomy

In their SSL survey [22], Van Engelen and colleagues proposed a detailed taxonomy for SSL methods in the framework in deep learning. The algorithms explored in the present article fit in the *intrinsically semi-supervised inductive methods* category, meaning methods that attempt to construct a classifier by directly optimizing an objective function for labeled and unlabeled samples. Most semi-supervised neural networks make use of perturbation-based learning methods, where the training data samples (labeled or unlabeled or both) are perturbed with data augmentation techniques. This is meant to incorporate the so-called *smoothness assumption* in SSL,

which states that a classifier should be robust to local perturbations in its input. This is the case of the five methods explored in our work: MT, DCT, MM, RMM, and FM. If we follow Van Engelen et al.'s taxonomy, MT is a *consistency regularization* method, in which predictions of a teacher and a student models are penalized when being different. DCT is described as a *pseudo-labeling* method, based on the *disagreement* between two models trained on two different views of the same data. As we shall see in the DCT description, the second view is automatically created by deriving adversarial examples of the original data samples. Finally, MM, RMM, and FM are considered as *hybrid* methods, in that they combine pseudo-labeling, consistency regularization and entropy minimization for performance improvement. Entropy minimization refers to methods that artificially lower the uncertainty of the predictions made on the unlabeled data. We will see, for instance, the use of a *sharpening* function in MM.

2.2 Semi-supervised deep learning in audio classification

In the seminal articles in which the five SSL methods were proposed, the experiments were carried out on image classification tasks only, not on audio related tasks. If we focus on SSL applied to sound event detection (SED), the most used technique in the literature is MT. In particular, the system ranked first in the Detection and Classification of Acoustic Scenes and Events (DCASE) task 4 2018 challenge (large-scale weakly labeled semi-supervised sound event detection in domestic environments) used MT with convolutional recurrent neural networks trained on a small labeled subset and a larger unlabeled one [23]. Since then, MT was used in the baseline system provided by the challenge organizers, and most of the systems proposed by the participants [24, 25]. Also in the framework of DCASE Task 4, Shi and colleagues adapted MM for the task [26]. Their MM method outperformed their solution based on MT¹. SED is a task consisting of segmenting an audio recording in possibly overlapping audio events. It is slightly different from audio classification, the target task of the present work, in which we more simply aim to tag audio recordings globally with a single audio event category per recording. Outside DCASE, MT has been favorably compared to supervised learning in [27] for audio classification. The authors show the importance of using diverse collections of noise as perturbations in MT. They also used MixUp successfully, as we will in the present article. Although they used two datasets in common with us (Google Speech Commands and UrbanSound8k), their

¹ <http://dcase.community/challenge2019/task-sound-event-detection-in-domestic-environments-results>

results cannot be compared to ours because of differences in the evaluation strategies: train/test splits different from the official ones and no cross-validation on UrbanSound8k, and a different number of target classes with Google Speech Commands. Finally, recently, FM and MT were compared on music, industrial sounds, and acoustic scenes classification data sets. FM outperformed MT and supervised learning in all cases [16].

2.3 An extension of our previous work

In previous work, we already compared two SSL methods for AT, namely MT and DCT, and we showed that DCT was consistently better than MT [15]. We build on this preliminary work to consider three simpler SSL methods, based on a single neural network instead of two models: MM, RMM and FM. Although some of these SSL methods were applied (in modified forms) to audio data in the context of audio classification before, as we just saw, the present work is among the first ones to compare a number of them in a systematic way.

As we shall see in their technical description, a key aspect in these three “holistic” methods is the extensive use of data augmentation techniques both on the labeled and unlabeled data subsets. In the results that we will report, we used the same augmentation techniques to train our fully-supervised baselines, which gave much stronger baselines than in our previous work [15]. Finally, another novelty of the present work is the addition of the mixup [8] augmentation to the SSL methods MT, DCT and FM.

3 Audio data augmentation

Augmentations are at the heart of most recent semi-supervised learning mechanisms. In this section, we begin by describing the mixup mechanism, which we extensively use in this work, and the other audio data augmentations used in some of the training settings.

3.1 Mixup

Mixup [8] is a successful data augmentation/regularization technique, that proposes to mix pairs of samples (images, audio clips, etc.). If x_1 and x_2 are two different input samples (spectrograms in our case) and y_1, y_2 their respective one-hot encoded labels, then the mixed sample and target are obtained by a simple convex combination:

$$\begin{aligned} x^{mix} &= \lambda x_1 + (1 - \lambda) x_2 \\ y^{mix} &= \lambda y_1 + (1 - \lambda) y_2 \end{aligned} \quad (1)$$

where λ is a scalar sampled from a symmetric Beta distribution at each mini-batch generation:

$$\lambda \sim \text{Beta}(\alpha, \alpha) \quad (2)$$

where α is a real-valued hyper-parameter to tune (always smaller than 1.0 in our case).

In the original MM algorithm, an “asymmetric” version of mixup is used, in which the maximum value between λ and $1 - \lambda$ is retrieved:

$$\lambda = \max(\lambda, 1 - \lambda) \quad (3)$$

This makes the λ values either close to one, allowing the resulting mixed batches to be closer to x_1 . This may be useful when the method mixes labeled and unlabeled samples, when only slight perturbations are wanted.

3.2 Audio signal augmentation methods

We tested several audio augmentation techniques and retained three of them: *Occlusion*, *CutOut* [28], and *Speed Perturbation* [29]. In addition to the three selected augmentations described below, we also tried to add uniform noise on the log-mel spectrograms, invert the mel frequency axis and the time axis, but no gains were observed with these techniques.

- *Occlusion*: applied to the raw audio signal, *Occlusion* consists of setting a segment of the waveform to zero. The size of the segment is randomly chosen up to a user-defined maximum size. The position of the segment is also chosen randomly.
- *CutOut*: applied to the log-mel spectrograms, *CutOut* sets the values within a random rectangle area with the -80 dB value, which corresponds to the silence energy level in our spectrograms. The length and width of the removed sections are randomly chosen from a predefined interval and depend on the spectrogram size.
- *Speed Perturbation*: we resample the raw audio signal up (nearest-neighbor upsampling) or down (decimation) according to a rate chosen randomly within a predefined interval. The resulting waveform is either shorter or longer. Padding or cropping is randomly applied at the start and the end of the stretched signal to keep the signal duration constant.

The difference between *Occlusion* and *CutOut* is that *CutOut* sets a time-frequency rectangle to the -80 dB value, whereas *Occlusion* sets to zero a whole portion of the waveform.

We used *Occlusion*, *CutOut*, and *Speed Perturbation* in augmented supervised learning settings, and in MM, RMM, and FM. During training, one of those is randomly applied to each audio sample.

RMM and FM make use of so-called “weak” and “strong” augmentations. The difference between the

Table 1 Augmentation hyperparameters

	Param.	Weak range	Strong range
Occlusion	Max size	[0.25, 0.25]	[0.75, 0.75]
CutOut	Scale	[0.10, 0.50]	[0.50, 1.00]
Speed perturb.	Rate	[0.50, 1.50]	[0.25, 1.75]

two lies in the strength and randomness with which an augmentation is applied. A “weak” augmentation has a 50% chance to be applied, and a “strong” one is always applied.

In order to tune these augmentations, we performed a grid-search on their hyperparameters, training Wide-Resnet28-2 models on the Google Speech Commands dataset (this architecture and dataset will be described here-after). The resulting hyperparameters are listed in Table 1.

No augmentation was used in DCT nor in MT, except Gaussian noise in MT.

4 Semi-supervised deep learning algorithms

This section provides a detailed description of the five SSL approaches we compare for audio classification. We chose them for their high performance reported for object recognition in images. Two of these approaches, Mean Teacher (MT) [10], and Deep Co-Training (DCT) [11] use the principle of consistency regularization between the outputs of two models. The other methods, MixMatch (MM) [12], ReMixMatch (RMM) [13], and FixMatch (FM) [14], use a single model and combine the three SSL mechanisms described in the introduction.

We provide a figure to illustrate each of the five methods. In Section 4.6, we explain how we add mixup to MT, DCT and FM, since MM and RMM already use it. We included a blue box in the method workflow figures, to show where mixup is optionally integrated. We will refer to the modified methods as “method+mixup,” for instance, FM+mixup.

4.1 Mean Teacher (MT)

MT uses two neural networks: a “student” f and a “teacher” g , which share the same architecture (Fig. 1). The weights ω of the student model are updated using the standard gradient descent algorithm, whereas the weights W of the teacher model are the Exponential Moving Average (EMA) of the student weights. The teacher weights are computed at every mini-batch iteration t , as the convex combination of its weights at $t-1$ and the student weights, with a smoothing constant α_{ema} :

$$W_t = \alpha_{\text{ema}} \cdot W_{t-1} + (1 - \alpha_{\text{ema}}) \cdot \omega_t \quad (4)$$

There are two loss functions applied either on the labeled or unlabeled data subsets. On the labeled data x_s , the usual cross-entropy (CE) is used between the student model’s predictions and the ground-truth y_s .

$$\mathcal{L}_s = \text{CE}(f(x_s), y_s) \quad (5)$$

The consistency cost is computed from the student predictions $f(x_s)$ and $f(x_u)$, and from the teacher prediction $g(x'_s)$ and $g(x'_u)$, where x'_s and x'_u correspond to the same samples but slightly perturbed with Gaussian noise with a 15 dB signal-to-noise ratio [24]. This cost is a Mean Square Error (MSE) loss:

$$\begin{aligned} \mathcal{L}_{\text{cc}} = & \text{MSE}(f(x_s), \perp g(x'_s)) \\ & + \text{MSE}(f(x_u), \perp g(x'_u)) \end{aligned} \quad (6)$$

The symbol \perp denotes the stop gradient operator, meaning that the teacher weights W_t are a constant with respect to optimization.

The final loss function is the sum of the supervised loss function and the consistency cost weighted by a factor λ_{cc} which controls its influence.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_s + \lambda_{\text{cc}} \cdot \mathcal{L}_{\text{cc}} \quad (7)$$

4.2 Deep Co-Training (DCT)

DCT has been recently proposed by Qiao et al. [11]. It is based on Co-Training (CT), the well-known generic framework for SSL proposed by Blum and colleagues in 1998 [30]. The main idea of Co-Training is based on the assumption that two independent views on a training dataset are available to train two models separately. Ideally, the two views are conditionally independent given the class. The two models are then used to make predictions on the unlabeled data subset. The most confident predictions are selected and added to the labeled subset. This process is iterative, like pseudo-labeling.

DCT is an adaptation of CT in the context of deep learning. Instead of relying on views of the data that are different, DCT makes use of adversarial examples to ensure the independence in the “view” presented to the models. Each batch is composed of a supervised and an unsupervised part. Thus, the unlabeled data are directly used, and the iterative aspect of the algorithm is removed.

Let \mathcal{S} and \mathcal{U} be the subsets of labeled and unlabeled data, respectively, and let f and g be the two neural networks that are expected to collaborate.

The DCT loss function is comprised of three terms, as shown in Eq. (8). These terms correspond to loss functions estimated either on \mathcal{S} , \mathcal{U} , or both. Note that during training, a mini-batch is comprised of labeled and unlabeled samples in a fixed proportion. Furthermore,

in a given mini-batch, the labeled examples given to each of the two models are sampled independently (Fig. 2).

$$\mathcal{L} = \mathcal{L}_s + \lambda_{\text{cot}} \mathcal{L}_{\text{cot}} + \lambda_{\text{diff}} \mathcal{L}_{\text{diff}} \quad (8)$$

The first term, \mathcal{L}_s , given in Eq. (9), corresponds to the standard supervised classification loss function for the two models f and g , estimated on examples x_1 and x_2 respectively, which are sampled from S .

In our case, we use categorical Cross-Entropy (CE), the standard loss function used in classification tasks with mutually exclusive classes.

$$\mathcal{L}_s = \text{CE}(f(x_1), y_1) + \text{CE}(g(x_2), y_2) \quad (9)$$

As in MT, a consistency cost on the unlabeled examples is used in DCT. It takes the form of the Jensen-Shannon (JS) divergence between the two sets of predictions on examples x_u sampled from the unlabeled subset \mathcal{U} , given by:

$$\begin{aligned} \mathcal{L}_{\text{cot}} = & H\left(\frac{1}{2}(f(x_u) + g(x_u))\right) \\ & - \frac{1}{2}(H(f(x_u)) + H(g(x_u))) \end{aligned} \quad (10)$$

where H denotes the entropy.

For DCT to work, the two models need to be complementary: on a subset different from $S \cup \mathcal{U}$, examples misclassified by one model should be correctly classified by the other model [31]. In DCT, this is achieved by generating adversarial examples with one model and training the other model to be robust to these adversarial samples. To generate adversarial examples, we used the Fast Gradient Signed Method (FGSM, [32]), as in Qiao's work. The $\mathcal{L}_{\text{diff}}$ loss term (Eq. (11)) is the sum of the Cross-Entropy losses between the predictions $f(x_3)$ and $g(x_3^f)$, where x_3 is sampled from $S \cup \mathcal{U}$ and x_3^f is the adversarial example generated with the model f from x_3 taken as input. The second term is the symmetric term for model g , with x_4 sampled from $S \cup \mathcal{U}$ and x_4^g the adversarial example generated with g from x_4 .

$$\begin{aligned} \mathcal{L}_{\text{diff}} = & \text{CE}(f(x_3), g(x_3^f)) \\ & + \text{CE}(g(x_4), f(x_4^g)) \end{aligned} \quad (11)$$

For more in-depth details on the technical aspects of DCT, the reader may refer to [11]. We implemented DCT as precisely as described in Qiao's article, using PyTorch, and made sure to accurately reproduce their results on CIFAR-10: about 90% accuracy when using only 10% of the training data as labeled data (5000 images).

4.3 MixMatch

MixMatch [12] (MM) uses entropy minimization and standard regularization, namely pseudo-labeling [5], mixup, and weak data augmentation, to leverage the unlabeled data and provide better generalization capabilities. Unlike MT and DCT, this approach uses only one model. The different steps are shown in Fig. 3 and detailed in the following paragraphs.

During the learning phase, each minibatch is composed of labeled x_s and unlabeled x_u samples in equivalent proportions. The first step consists of applying an augmentation to the labeled part of the mini-batch and K augmentations to the unlabeled part in parallel. These K augmentations are sampled from the three augmentations (weak) described in Section. In the second step, pseudo-labels y_u are generated for the unlabeled files using the model's prediction averaged on these K variants as shown in Eq. (12), where $x'_{u,i}$ denotes the i -th variant of an unlabeled augmented file.

$$\hat{y}_u = \frac{1}{K} \sum_{i=1}^K f(x'_{u,i}) \quad (12)$$

For encouraging the model to produce confident predictions, a post-processing step is necessary to decrease the output's entropy. To do so, the highest probability is increased and the other ones decreased. This process is called "sharpening" by the method authors, and it is defined as:

$$\text{sharpen}(p, T)_i := p_i^{1/T} / \sum_{j=1}^{|p|} p_j^{1/T} \quad (13)$$

The sharpen function is applied on to the pseudo-labels $p = \hat{y}_u$. The parameter T , called Temperature, controls the strength of the sharpen function. When T tends towards zero, the entropy of the distribution produced is lowered.

Finally, the labeled and unlabeled augmented samples are concatenated and shuffled into a W set then used as a pool of training samples used by the asymmetric mixup function. Asymmetric mixup is applied separately on the labeled and unlabeled parts of the mini-batch, as formulated here:

$$x_s^{\text{mix}} = \text{mixup}(x_s | W_{1..B_s}) \quad (14)$$

$$x_u^{\text{mix}} = \text{mixup}(x_u | W_{|x_s|+1..|W|}) \quad (15)$$

where B_s and $|W|$ are the number of labeled samples and of the whole W set. The W set and the corresponding labels are shuffled in the same order. Each labeled sample

is then perturbed by a second labeled or unlabeled sample. Mixing the two is done so that the original labeled sample remains the main component of the resulting sample. The operation has been detailed in Section 3.1. The same procedure is applied onto the unlabeled files using the remaining samples from W .

The original MixMatch loss function is composed of the standard CE cost for the supervised loss \mathcal{L}_s , and the MSE for the unsupervised loss \mathcal{L}_u . We replace MSE with CE in all our experiments, as proposed in the ReMixMatch paper. Indeed, it seems that CE performs better than MSE in our experiments.

$$\mathcal{L}_s = \frac{1}{B_s} \sum_{(x_s^{\text{mix}}, y_s^{\text{mix}})} \text{CE}(f(x_s^{\text{mix}}), y_s^{\text{mix}}) \quad (16)$$

$$\mathcal{L}_u = \frac{1}{K \cdot B_u} \sum_{(x_u^{\text{mix}}, \hat{y}_u^{\text{mix}})} \text{CE}(f(x_u^{\text{mix}}), \hat{y}_u^{\text{mix}}) \quad (17)$$

where B_s and B_u are the number of examples in the labeled and unlabeled mini-batches.

The final loss is the sum of the two components, with a hyper-parameter λ_u :

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \cdot \mathcal{L}_u \quad (18)$$

4.4 ReMixMatch (RMM)

ReMixMatch (RMM) [13] was presented as an improvement of MixMatch and introduced the concept of strong and weak augmentations and a so-called distribution alignment mechanism (Fig. 4).

At every iteration, the batch is composed of labeled x_s and unlabeled x_u samples. One weak augmentation and K strong augmentations are applied on x_u . The weakly-augmented sample is used to compute the pseudo-label vectors \hat{y}_u of the unlabeled examples.

$$\hat{y}_u = f(\text{weak}(x_u)) \quad (19)$$

A distribution alignment mechanism modifies the pseudo-labels to make them follow the class distribution of the labeled subset. Two “distributions” p_s and p_u are estimated in the form of vectors, which are respectively the averages of the true labels y_s and of the pseudo-labels \hat{y}_u , calculated over the samples of the N previous batches. Then, distribution alignment is applied to \hat{y}_u with this equation:

$$\hat{y}_u = \text{Normalize}\left(\hat{y}_u \cdot \frac{p_s}{p_u}\right) \quad (20)$$

Finally, we apply the sharpen function from Eq. (13) to the pseudo-labels \hat{y}_u , as done in MixMatch. The labels \hat{y}_u will be used as targets for the weakly and strongly augmented batches. Like in MixMatch, we concatenate the labeled and unlabeled batches to a set W for the mixup augmentation, and the labeled and unlabeled loss \mathcal{L}_s and \mathcal{L}_u remain the same.

ReMixMatch also introduced a strong-augmentation loss component for increasing stability and accuracy. This component will be computed with the first strongly-augmented version of x_u , called x'_{u_1} :

$$\mathcal{L}_{u_1} = \frac{1}{B_u} \sum_{(x'_{u_1}, \hat{y}_u)} \text{CE}(f(x'_{u_1}), \hat{y}_u) \quad (21)$$

In the original ReMixMatch, the authors added another loss term, a self-supervised learning component that predicts which transformation is applied to the x'_{u_1} batch. The transformation used was a rotation of 0, 90, 180, or 270°, and the model had to guess which angle the image had been rotated by (a four-class classification task). In some configurations, it was supposed to help the model to avoid collapsing during training. This component was removed because it did not show any positive impact on our experiments, and using rotations or flips on audio spectrograms is difficult to justify in terms of audio semantics.

In our experiments, the final loss is the sum of the three different components:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \cdot \mathcal{L}_u + \lambda_{u_1} \cdot \mathcal{L}_{u_1} \quad (22)$$

4.5 FixMatch

FixMatch [14] (FM) is another SSL method which proposes a simplification of MM and ReMixMatch. The method also uses one model, removes mixup and replaces the sharpen function by binary pseudo-labels. FM uses both weak augmentations (weak) and strong augmentations (strong). The strong augmentations can mislead the model predictions by disrupting too much the training data. Figure 5 shows the main pipeline of FixMatch. As in the other method illustrations, we added a mixup box in blue, to indicate where we add it to the algorithm in our modified FM algorithm, thus called FM+mixup.

The supervised loss component is the standard cross-entropy applied to the weakly-augmented data:

$$\mathcal{L}_s = \text{CE}(f(\text{weak}(x_s)), y_s) \quad (23)$$

Then, we guess the labels of the weakly augmented unlabeled data and apply a binarization (argmax) of these predictions to have a one-hot encoded label. This label is used as target for training the model with strongly

augmented unlabeled data. It allows the model to generalize with weak and strong augmentations and it also uses the guessed label to improve the model accuracy with unlabeled data:

$$\hat{y}_u = f(\text{weak}(x_u)) \quad (24)$$

To avoid training on incorrect guessed labels, FM uses a threshold τ that ensures that the unsupervised cost function can only be applied to predictions made with high confidence, i.e., above this threshold. This can be easily implemented in the form of a mask:

$$\begin{aligned} \text{mask} &= 1(\max(\hat{y}_u) > \tau) \\ \mathcal{L}_u &= \text{mask} \cdot \text{CE}(f(\text{strong}(x_u)), \arg\max(\hat{y}_u)) \end{aligned} \quad (25)$$

As in MixMatch, we sum the loss components to compute the final loss:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \cdot \mathcal{L}_u \quad (26)$$

4.6 Adding mixup to MT, DCT, and FM

As we described here-above, MM and RMM already uses mixup in its workflow. In order to measure the impact of mixup, we will report results when we remove mixup from MM and RMM. On the contrary, the three other SSL methods explored in our work (MT, DCT, FM) do not use mixup in their original version. We explored several ways to add mixup to them, and retained the best one for each of the three methods. Note that we illustrate where the mixup operation has been added in the figures describing the different methods in the previous section.

Since the labeled and unlabeled data flow is very similar in MM and FM, we added mixup to FM at the same place as in MM: both labeled and unlabeled samples are mixed up. Similarly, it is also the asymmetric mixup variant that we used in MM and FM since mixup is applied to labeled and unlabeled samples together, as in the original MM method. Using mixup on labeled and unlabeled examples separately seems to hurt performance with these two methods.

In MT, mixup is applied on labeled and unlabeled samples separately and only for the teacher model. The perturbation with Gaussian noise applied to the unlabeled samples is removed, since no gain was observed when mixup is used instead.

For DCT, mixup is applied on the unlabeled samples only, common to both models in each minibatch during training. Applying mixup on the labeled samples, which are sampled differently for the two models at each training step, lead to worse results. It is then, not necessary to use the asymmetrical variant for MT and DCT.

Finally, in all cases, we apply mixup on the log-mel spectrograms, which are the input features given to our

deep neural networks (feature extraction is detailed in the Section 5).

5 Experiments

In this section, we describe our experimental setup. We give a brief description of the datasets and metrics, describe the Wide ResNet architecture we used, together with the training strategy details.

5.1 Datasets and evaluation metrics

Environmental Sound Classification 10 (ESC-10) [17] is a selection of 400 5-s-long recordings of audio events separated into ten balanced categories. The dataset is provided with five uniformly sized cross-validation folds that will be used to perform the evaluation. The files are sampled at 44 kHz and are converted into 431×64 log-mel spectrograms.

UrbanSound8k (UBS8K) [18] is a dataset composed of 8742 files between 1 and 4 s long, separated into ten balanced categories. The dataset is provided with ten cross-validation folds of uniform size that will be used to perform the evaluation. The files are zero-padded to 4 s, resampled to 22 kHz, and converted to 173×64 log-mel spectrograms.

Google Speech Commands Dataset v2 (GSC) [19] is an audio dataset of spoken words designed to evaluate keyword spotting systems. The dataset is split into 85,511 training files, 10,102 validation files, and 4890 testing files. The latter is used for the evaluation of our systems. We ran the task of classifying the 35 word categories of this dataset. The files are zero-padded to 1 s if needed and sampled at 16 kHz before being converted into 32×64 log-mel spectrogram.

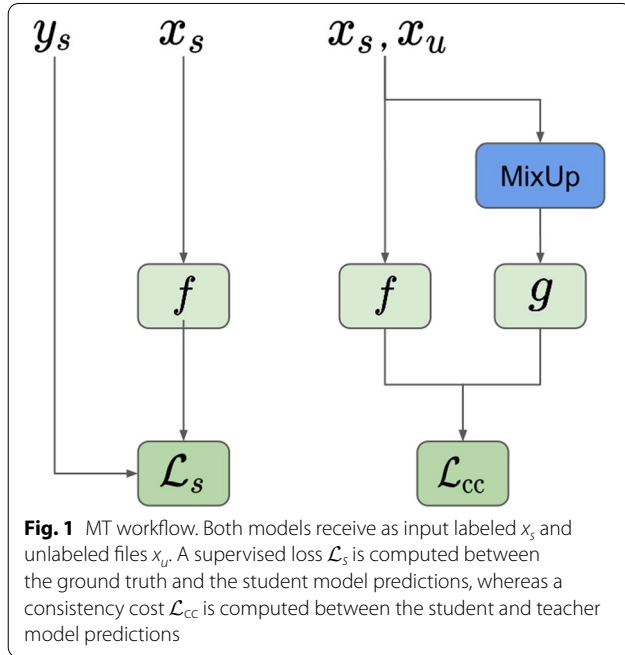
In all cases, the 64 mel-coefficients were extracted using a window size of 2048 samples and a hop length of 512 samples. For ESC-10 and UBS8K, we used the official cross-validation folds. We report the average classification Error Rate (ER) along with standard deviations. ER is defined as the number of errors divided by the total number of samples.

5.2 Models

We used the Wide-ResNet-28-2 [33] architecture in all our experiments (Table 2). This model is very efficient, achieving SOTA performance on the three datasets when trained in a 100% supervised setting. Moreover, its small size, comprised of about 1.4 million parameters, allows to experiment quickly. Its structure consists of an initial convolutional layer (conv1) followed by three groups of residual blocks (block1, block2, and block3). Finally, an average pooling and a linear layer act as a classifier. The residual blocks, composed of two BasicBlock, are repeated three times and their structure is defined in Eq.

Table 2 Architecture of Wide ResNet28-2. Downsampling is performed by the first layers in block2 and block3

Layer	Architecture
input	Log mel spectrogram
conv1	BasicBlock(32)
	Max pool
block1	$\begin{bmatrix} \text{BasicBlock}(32) \\ \text{BasicBlock}(32) \end{bmatrix} \times 4$
block2	$\begin{bmatrix} \text{BasicBlock}(64) \\ \text{BasicBlock}(64) \end{bmatrix} \times 4$
block3	$\begin{bmatrix} \text{BasicBlock}(128) \\ \text{BasicBlock}(128) \end{bmatrix} \times 4$
	Avg pool
	ReLU
	Linear



(27). The number of channels of the convolution layers is referred as l , BN stands for Batch Normalization and ReLU [34] for the Rectified Linear Unit activation function. We used the official implementation available in PyTorch [35].

$$\text{BasicBlock}(l) = (\text{conv } 3 \times 3 @ l, \text{BN}, \text{ReLU}) \quad (27)$$

5.3 Training configurations

Each model was trained using the ADAM [36] optimizer. Table 3 shows the hyper-parameter values used for each method, such as the learning rate lr , the

mini-batches' size bs , the warmup length wl if used, and the number of epochs e . These parameters are identical regardless of the dataset used, unless otherwise specified. They were obtained by performing a reasonable short grid-search using UBS8K dataset first validation fold.

For supervised training, MM and FM, the learning rate remains constant throughout training. For MT and DCT, the learning rate is weighted by a descending cosine rule, function of the learning epoch t :

$$lr = 0.5 \left(1.0 + \cos \left((t - 1) \frac{\pi}{N_e} \right) \right) \quad (28)$$

where N_e denote the number of epochs.

All the SSL approaches, but FixMatch, introduce one or more subsidiary terms to the loss. To alleviate their impact at the beginning of the training, these terms are weighted by a lambda λ ratio, which ramps up to its maximum value within a warmup length wl . The ramp-up strategy is defined in Eq. (29) for MT and DCT, and is linear in MM during the first 16k learning iterations.

$$\lambda = \lambda_{\max} \times e^{-5 \times (1 - (t/wl))^2} \quad (29)$$

In MT, the maximum value of λ_{cc} is 1 and α_{ema} is set to 0.999. In DCT, the maximum values of λ_{cot} and λ_{diff} are 1 and 0.5, respectively. In MM the maximum value of λ_u is 1. FM and RMM do not use a ramp up strategy. In FM, the value of λ_u is set to 1 and in RMM the values of λ_u , λ_{u1} and λ_r are set to 1.5, 0.5, and 0.5, respectively.

In MM and RMM, we use two augmentations ($k=2$), the sharpening temperature T is set to 0.5. In FM, we use a threshold $\tau=0.8$ on ESC-10 and GSC datasets, and $\tau=0.95$ for UBS8K. In RMM, the number of labels N kept for distribution alignment is set to 128.

For MM, FM and RMM, on ESC-10, the batch size is 60 because ESC-10 is a small dataset of 400 files only. During training, only four folders are used, that is, 320 files. In a 10% configuration and due to the whole division's restrictions, this represents only 30 supervised files in total. Each mini-batch must contain as many labeled as unlabeled files, hence the batch size of 60. Moreover, because of this small number of files, the training phase only lasts for 2700 iterations, and therefore, warm-up ends prematurely.

For our proposed variants, which include mixup, we kept the same configurations and parameter values.

6 Results

We first report the results obtained in a supervised setting, with and without the same data augmentation methods used in the SSL algorithms, including mixup.

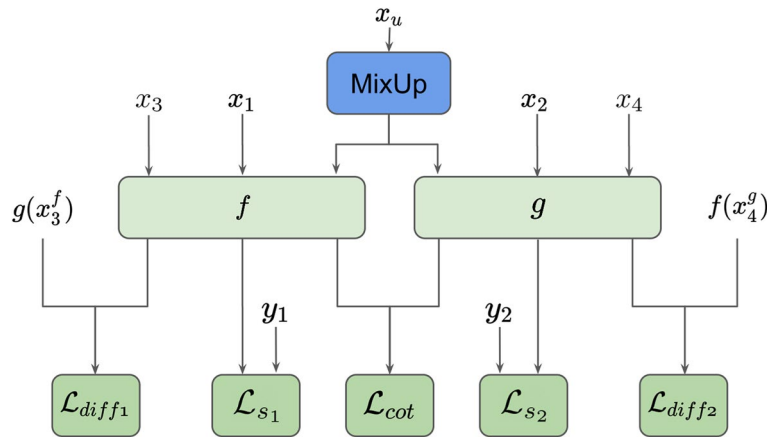


Fig. 2 DCT workflow. Each model is trained on its own labeled samples x_i , unlabeled samples x_u , and the adversarial examples generated by the other model. Model f makes predictions on x_1 and x_4^g , and model g on x_2 and x_3^f . In our DCT+mixup variant, mixup is used on the unlabeled samples only

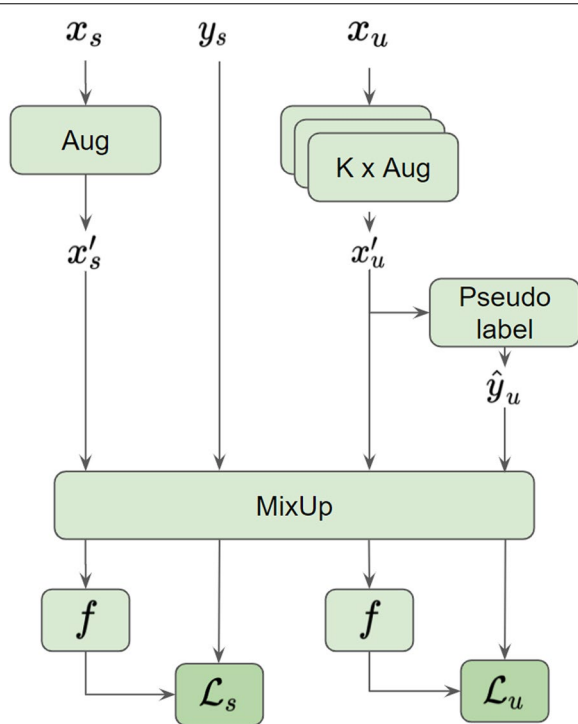


Fig. 3 MM workflow. K augmentations are applied to the unlabeled data x_u , and the averaged model predictions are used as pseudo labels \hat{y}_u . The labeled and augmented unlabeled data are mixed up and used to compute the supervised and unsupervised loss values

6.1 Supervised learning

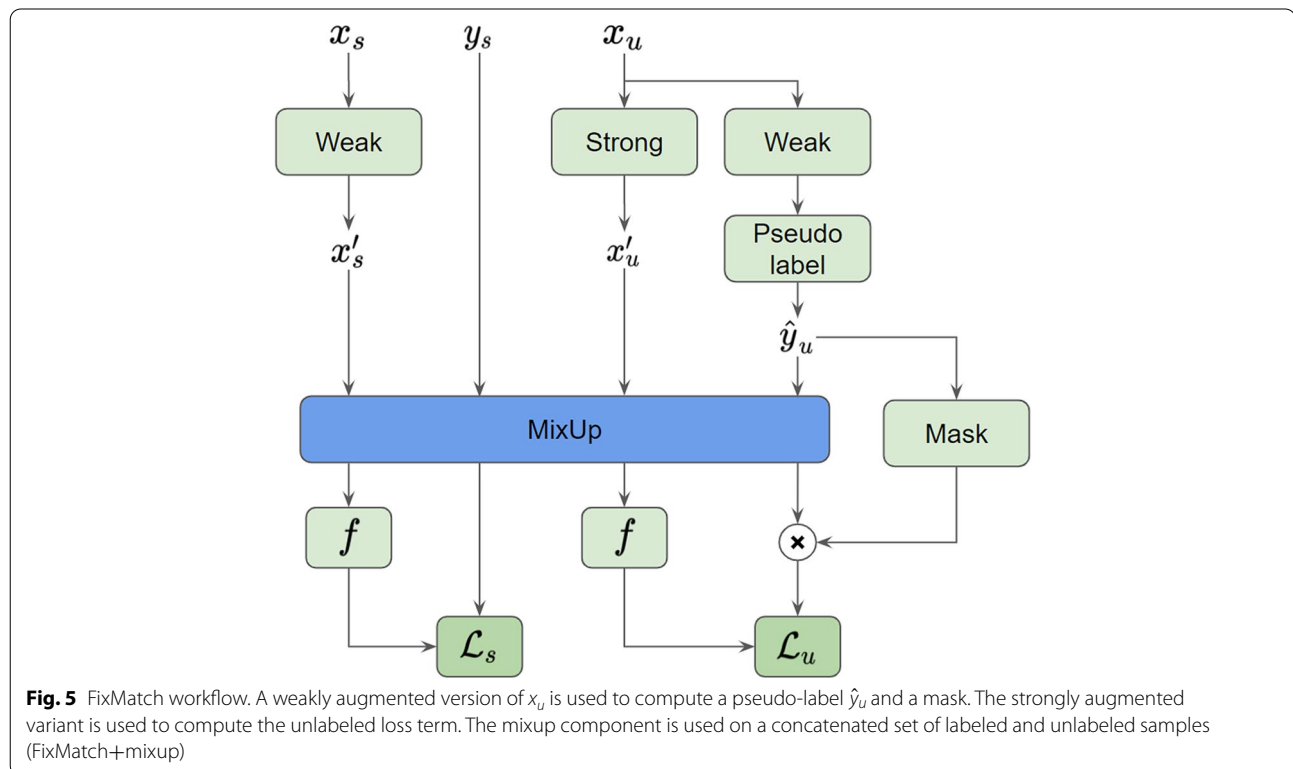
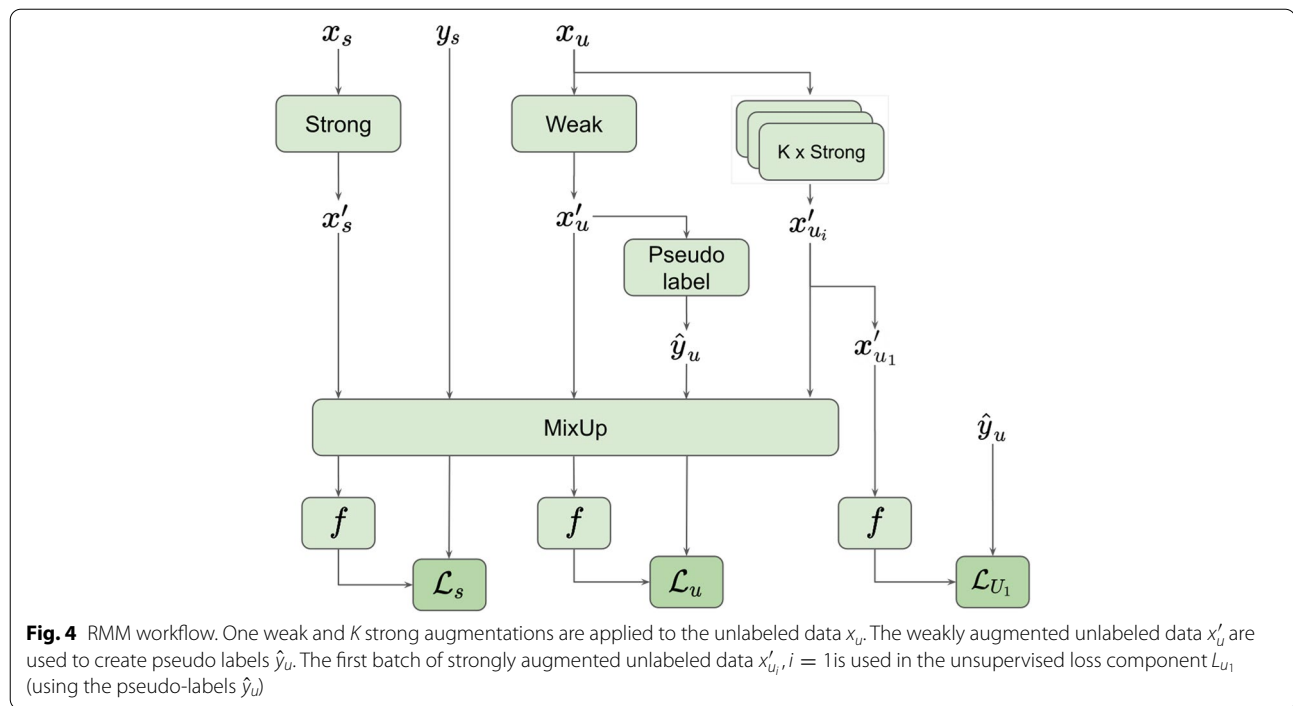
This section presents the results obtained with supervised learning in different settings while using either 10% or 100% of the labeled data available. MM, RMM and FM use augmentations as their core mechanism. RMM and FM use weak and strong augmentations, while MM uses a combination of weak augmentations and mixup. Therefore, it seems essential for fair comparisons to use the same augmentations in the supervised settings too.

We trained models without any augmentation (Supervised), using mixup alone (mixup), weak augmentations alone (Weak), a combination of weak augmentations and mixup (Weak+mixup), strong augmentations alone (Strong), and to finish, a combination of strong augmentations with mixup (Strong+mixup). Table 4 presents the results on ESC-10, UBS8K, and GSC. In order to give an idea of how our results compare to the literature, we reported three results from the literature, in the “CNN models (literature)” row in the table. We chose to report results from works in which the models are primarily based on a CNN architecture, to be fair with the Wide-ResNet we used in our case. There are better results from the recent literature, but that involved large transformer models, sometimes pretrained on AudioSet. For instance, the state-of-the-art result on UBS8K is 10.0% ER, obtained with a 25-M parameter transformer, pretrained on AudioSet [37].

6.1.1 ESC-10

In the 10% setting, the supervised model reached an ER of 32.00%. The use of Weak yielded the best performance with 22.67% ER, outperforming the supervised model

We compare the error rates obtained by the five SSL methods and then show that adding mixup is almost in all cases beneficial.



by 9.3 points (29.16% relative). In the 100% setting, the supervised model reached an ER of 8.00%, and the best ER of 4.67% was achieved when using Weak+mixup. The gain is 3.33 points (41.62% relative).

6.1.2 UBS8K

In a 10% setting, the supervised model reached 33.80% ER, and the best supervised result was obtained with Weak+mixup, with a 23.75% ER. It represents an

Table 3 Training parameters used on the datasets. Bs: batch size, lr: learning rate, wl: warm-up length in epochs, N_e : number of epochs, α : mixup Beta param

	bs	lr	wl	N_e	α
Supervised	256	0.001	-	300	-
Mixup	256	0.001	-	300	0.40
MT	64	0.001	50	200	-
MT+mixup	64	0.001	50	200	0.40
DCT	64	0.0005	160	300	-
DCT+mixup	64	0.0005	160	300	0.40
MM-mixup	256	0.001	-	300	-
MM	256	0.001	-	300	0.75
RMM-mixup	256	0.001	-	300	-
RMM	256	0.001	-	300	0.75
FM	256	0.001	-	300	-
FM+mixup	256	0.001	-	300	0.75

improvement of 10.05 points, 29.73% relative improvement. In the 100% setting, the same augmentation combination reached an ER of 17.96%, outperforming the 23.29% ER from the supervised model by 5.33 points, 22.88% relative improvement.

6.1.3 GSC

In a 10% setting, the supervised model reached 10.01% ER, and Weak+mixup yielded the best ER of 6.58%. It represents an augmentation of 3.43 points, 34.26% relative improvement. In the 100% setting, the Strong+mixup reached an ER of 2.98%, outperforming the 4.94% ER from the supervised model by 1.96 point, 39.68% relative improvement.

Overall, we observe that in a supervised setting, the combination of mixup with a weak or a strong augmentation is systematically better than using a single augmentation, except in the ESC-10 dataset.

Table 4 Supervised learning Error Rates (%) on ESC-10, UBS8K, and GSC

Dataset	ESC-10		UBS8K		GSC	
Labeled fraction	10%	100%	10%	100%	10%	100%
CNN models (literature)	-	3.00 [38]	-	14.50 [39]	-	3.00 [40]
Supervised	32.00 ±6.17	8.00 ±5.06	33.80 ±4.82	23.29 ±5.80	10.01	4.94
+mixup	36.00 ±5.22	8.33 ±4.56	31.41 ±5.56	22.04 ±5.99	8.83	3.86
+weak	22.67 ±3.46	4.67 ±3.43	27.08 ±4.58	20.09 ±5.50	7.62	3.90
+weak+mixup	24.67 ±4.92	4.67 ±1.39	23.75 ±4.73	17.96 ±3.64	6.58	3.00
+strong	23.00 ±5.19	5.00 ±2.64	25.58 ±4.15	20.69 ±4.92	7.60	3.27
+strong+mixup	24.00 ±8.71	5.00 ±4.25	24.73 ±4.42	18.52 ±4.38	6.86	2.98

Table 5 Semi-supervised learning error rates (%) on ESC-10, UBS8K, and GSC

Dataset	ESC-10		UBS8K		GSC	
Labeled fraction	10%	100%	10%	100%	10%	100%
Supervised	32.00 ± 6.17	8.00 ±5.06	33.80 ±4.82	23.29 ±5.80	10.01	4.94
Best supervised	22.67 ± 3.46	4.67 ±1.39	23.75 ±4.73	17.96 ±3.64	6.58	2.98
MT	28.28 ± 5.28	-	32.80 ±4.21	-	8.51	-
MT+mixup	27.81 ± 2.25	-	32.00 ±5.80	-	8.50	-
DCT	25.16 ± 4.42	-	27.85 ±4.29	-	6.22	-
DCT+mixup	23.75 ± 2.36	-	25.77 ±4.73	-	5.63	-
MM-mixup	17.33 ± 3.84	-	20.42 ±4.88	-	4.49	-
MM	15.33 ± 5.58	-	18.02 ±4.00	-	3.25	-
RMM-mixup	32.50 ±11.71	-	38.23 ±6.15	-	5.15	-
RMM	12.00 ±5.55	-	28.41 ±6.54	-	3.54	-
FM	13.33 ± 2.89	-	21.44 ±4.16	-	4.44	-
FM+mixup	14.67 ± 7.21	-	18.27 ±3.80	-	3.31	-

6.2 Semi-supervised learning

We report in Table 5 the results of the SSL methods. For MM and RMM, mixup is already used in the original methods, thus, we compare MM to MM without mixup (MM-mixup) and RMM to RMM without mixup (RMM-mixup). For the three other methods, we denote for instance FM+mixup the FM algorithm augmented with mixup.

In all the three datasets, the five SSL methods brought ER decreases compared to the 10% supervised learning setup, when no augmentation is performed. Only MM, RMM, and FM performed better than the best supervised training result, that used the weak augmentations. Furthermore, they also significantly outperformed MT and DCT in all but one cases (DCT better than RMM on UBS8K), showing that using single-model SSL methods is more efficient than two-model-based methods, at least on these three datasets and among the five methods that were compared.

For ESC-10, in the 10% setting, the lowest ER was achieved by RMM with a 12.00% value, compared to a 22.67% for a weakly augmented supervised training. It represents a 10.67 points improvement, 47.1% relative. The difference with a fully supervised training using weak augmentations reaching a 4.67% ER is still notable with a 7.33 points difference.

On UBS8K, the best ER was achieved using MM with an 18.02% ER, very closely followed by FM+mixup with 18.27%. The difference with the best supervised training Weak+mixup, reaching 23.75%, represents a difference of 5.73 points (24.13% relative). The performance of MM is also very close to the best fully supervised training Weak+mixup, which reached a 17.96% ER. The difference is only 0.06 points. Similarly to ESC-10, if MT and DCT outperformed the supervised training methods, they performed worse than supervised learning with augmentation. UBS8K is the only dataset for which RMM performed worse than DCT.

The GSC dataset results confirm the previous observations. The MM method is the best method with an ER of 3.25%, representing a relative gain of 6.76 (67.53%) or 3.33 points (50.61%) compared to supervised training without and with Weak+mixup augmentations, respectively. RMM and FM+mixup obtained results very similar to MM: 3.54% and 3.31% ER, respectively.

6.3 Impact of mixup

Given that the best SSL methods so far were MM and RMM, and that mixup is used in these approaches, we decided to try to add mixup to MT, DCT, and FM, in different ways for each method as explained in Section 4.6. In [14], Appendix D.2, mixup on the entries (not on the labels) was added to FM, removing all the other image

augmentations. In this setting, FM was shown to reach an accuracy very close to that of MM on CIFAR-10.

In Table 5, we reported the results when adding mixup to MT, DCT and FM, (MT+mixup, DCT+mixup, FM+mixup). We also give the ER when removing mixup from MM and RMM, in the row named MM-mixup and RMM-mixup.

As a first comment, MM-mixup and RMM-mixup are always worse than with mixup. For instance, with MM on UBS8K, ER increased from 18.02 to 20.42%. This is particularly visible with RMM on ESC-10 and UBS8K. Moreover, adding mixup to the other SSL methods brought performance improvements on all the datasets tested. The only counter-example observed is FM on ESC-10, which went from 13.33 to 14.67% ER. The standard deviation value also increased significantly from 2.89 to 7.21%.

Similarly, FM on UBS8K went from 21.44% ER without mixup to 18.24% with mixup. On GSC, RMM presented the largest gap between 5.15 and 3.54% ER without and with mixup, respectively.

It is also important to note that using mixup allowed to get ER values very close to the ones obtained with fully (100% setting) supervised training using augmentations, on UBS8K and GSC. This is observable with MM, RMM, and FM+mixup. For instance, compared to Weak+mixup 100% supervised, MM has only 0.06 point difference on UBS8K, and 0.27 point difference on GSC.

When we look at our supervised training performance, we can observe that an improvement does not systematically follow the use of weak or strong augmentations. However, when combined with mixup, ER is frequently improved. This can be partly explained by the fact that audio augmentations are often difficult to choose and that their impact is often dependent on the dataset and the task at hand [27]. With this in mind, mixup seems to be beneficial regardless of the dataset used.

6.4 Training time

The normalized training duration means for all the five methods are shown in Fig. 6. The values were computed on the three datasets using the following equation:

$$\text{mean} = \frac{d}{N_f \cdot N_e \cdot \text{bs}} \quad (30)$$

Where d is the total duration, N_f the number of folds in the dataset, N_e the number of epochs, and bs the batch size used in each method. We compute the three means for each dataset then we report the average of the three values. Finally, we use the *supervised 100%* execution time as the reference (training duration of one). We also

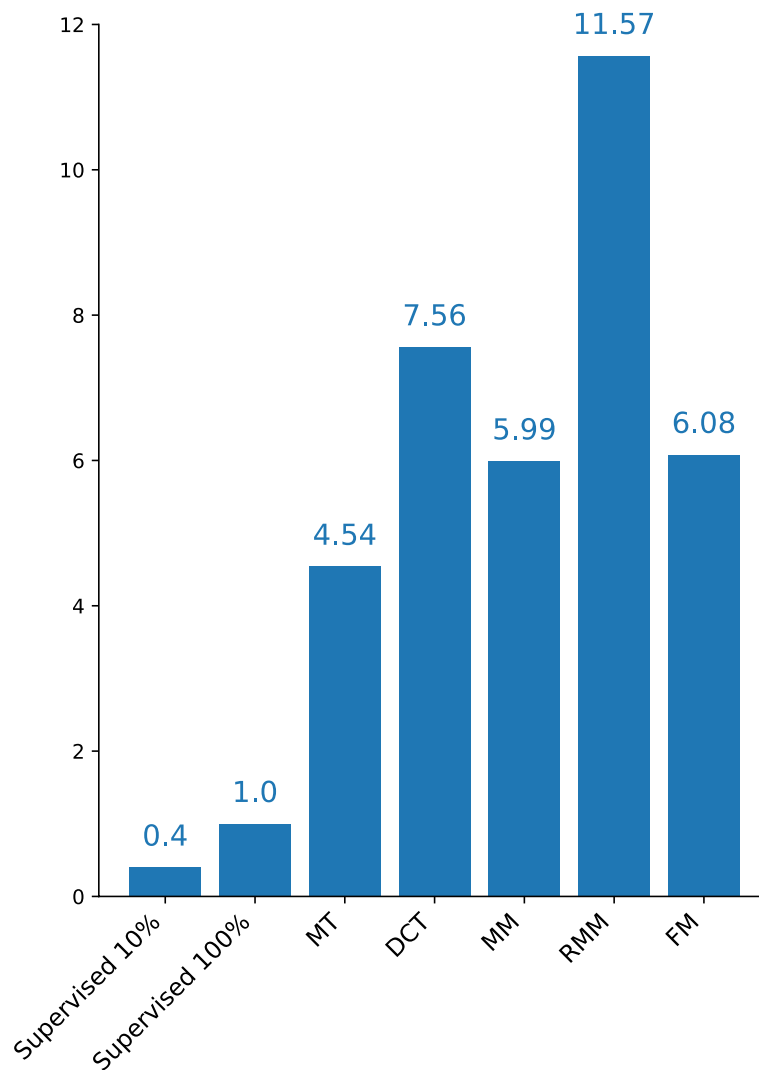


Fig. 6 Normalized mean training duration for all methods without mixup

assessed the impact of adding mixup, but it had a negligible impact of about 0.5%.

Among the SSL approaches, the fastest one is MT, which has a training time 4.5 times longer than the fully supervised training. Then, FM and MM follow with 6 times longer. DCT, with its high complexity and use of adversarial data, took up to 7.6 times longer, and finally the longest of all is RMM, 11.6 times longer, due to the large number of augmentations involved.

7 Discussion

7.1 Why are MM, RMM, and FM better than MT and DCT?

This question remains open. Several key components may explain this gap in performance. First, data augmentation is extensively used in these methods (weak

and strong ones), both on the labeled data and on the unlabeled subset to satisfy the consistency criterion of SSL. No data augmentation is used in the DCT nor MT basic methods, except the addition of noise in MT, on the unlabeled subset at the input of the teacher model. Nevertheless, when mixup was added to MT, no significant gain was observed. Thus, other augmentations should be explored for MT. Second, MM, RMM, and FM use pseudo-labeling, with either explicit entropy minimization (sharpening in MM and RMM) or threshold-based selection (confidence masking in FM). In DCT and MT, no entropy minimization is used, the predictions on the unlabeled part of the data are used as is for a consistency criterion between the two collaborating networks.

7.2 Which augmentations?

We used three augmentations (besides mixup): Occlusion, CutOut, and Speed perturbation. An advantage of those is that they are task-agnostic. We tuned their hyperparameters once on GSC, and then, we used them on ESC-10 and UBS8K as is, bringing performance improvements. Exploring more audio-specific augmentations is an avenue still to be explored. For instance, we did not try pitch shifting nor dynamic range compression [41]. Those would need careful parameter tuning depending on the audio event types and on the dataset involved in the experiments.

Finally, Occlusion and CutOut could be replaced by SpecAugment [42], originally proposed in automatic speech recognition and very often used nowadays in audio processing tasks, such as audio tagging. There is two small differences, though, in using SpecAugment, since it drops out one or several vertical and horizontal stripes from the spectrograms, while CutOut drops out a single rectangle of random shape. Another difference is that we applied randomly either Occlusion or CutOut, but not a combination of the two. To evaluate the effect of SpecAugment, we ran supervised learning experiments on GSC, using Speed Perturbation and mixup, and SpecAugment instead of Occlusion and CutOut, in the 100% of the labeled training data setting. We tested several configurations for SpecAugment. Our best setting was zero, one or two frequency stripes of width between 0 and 7 bins, and zero or one stripe of width also between 0 and 7 bins in time. This setting led to a 2.51% ER, which is better than the 2.98% value of our best supervised baseline method. This confirms experimentally that SpecAugment could replace Occlusion and CutOut, as a combination of the two. We did not rerun all the SSL experiments with SpecAugment, but one might expect slightly better results than those obtained with Occlusion and CutOut.

8 Conclusions

In this article, we reported audio classification experiments in a semi-supervised setting on three standard datasets of different sizes and content, the very small-sized ESC-10 with generic audio events, urban noises with UrbanSound8K, and speech with Google Speech Commands. We used only 10% of the labeled training data samples and the remaining 90% as unlabeled samples. We adapted and compared five SSL algorithms for this task, two methods that use two neural networks in parallel: Mean Teacher and Deep Co-Training, and the three single-model methods MixMatch, ReMixMatch and FixMatch, that strongly rely on data augmentation.

All the five methods brought significant gains compared to a supervised training setting using 10% of

labeled data. They performed better than supervised learning without augmentation. On UBS8K, MixMatch and FixMatch were very close to fully supervised learning with augmentation (100% of labeled training data). On ESC-10, ReMixMatch reached the best Error Rate of 12.00%. The relative gains were 62% and 47%, when compared to a supervised training using 10% of labeled data, without and with augmentation, respectively. On UrbanSound8K, MixMatch obtained the best results, reaching 18.02% Error Rate. Compared to a 10% supervised training without and with augmentation, the respective relative improvements were 47% and 24%. On Google Speech Commands, MixMatch again reached the best Error Rate of 3.25%. The relative improvement was 68% and 51%, compared to a 10% supervised training without and with augmentation, respectively. Mixup is an efficient regularization technique that is at the heart of the MixMatch and ReMixMatch algorithms. Its consistent impact in MM and RMM encouraged us to add it to the other SSL approaches. In almost all the experiments, adding mixup brought consistent improvements, which allowed us to get closer to the best supervised learning settings using 100% of the labeled data available. For instance, adding mixup to FixMatch reduced the error rates on UrbanSound8K from 21.4% to 18.3%, and from 4.4% to 3.3% on Google Speech Commands, to be compared with 17.9% and 3.0% respectively, obtained in the best supervised learning settings.

In conclusion, if we were to recommend a method out of the ones tested in our work, we would recommend MixMatch, and FixMatch+mixup also, with very similar performances. Their good results are consistent across the three datasets. The gains brought by these methods is worth their training time, about six times the 100% supervised setting training time. ReMixMatch obtained the best results on ESC-10, but this method is more demanding in training time.

Many questions remain open, though. The fact that MM and RMM were slightly better than FM needs to be further investigated, in particular the use of audio augmentations different in nature for the weak and the strong ones may be a direction to explore. MT and DCT do not use augmentations in their original version. It would be interesting, though, to try the weak augmentations used in the holistic methods with them. We also plan to adapt the SSL methods to multi-label audio tagging, for instance on Audioset [43] or FSD50K [44]. In particular, we would have to adapt the sharpen method in MixMatch, and the thresholding operations in FixMatch. Finally, new SSL methods have been very recently proposed and could be added to our list, such as Unsupervised Data Augmentation (UDA) [45], and the recent Meta Pseudo Labels method [46].

Acknowledgements

Not applicable

Authors' contributions

LC implemented the DCT and MT methods, wrote their descriptions. EL implemented the holistic methods (MM, RMM and FM) and wrote their descriptions. They both conducted all the experiments. They were both major contributors in writing the manuscript. TP did the conception and design of the work, wrote the related work section, and substantively revised the manuscript. All authors read and approved the manuscript.

Funding

This work was partially supported by the French ANR agency within the LUDAU project (ANR-18-CE23-0005-01) and the French "Investing for the Future — PIA3" AI Interdisciplinary Institute ANITI (Grant agreement ANR-19-PI3A-0004). We used HPC resources from CALMIP (Grant 2020-p20022) and from the Osirim platform.

Availability of data and materials

The three datasets used in this article are publicly archived datasets: <https://github.com/karolpiczak/ESC-50>; <https://urbansounddataset.weebly.com/urbansound8k.html>; http://download.tensorflow.org/data/speech_comma_nids_v0.02.tar.gz.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

¹IRIT, University of Toulouse, CNRS, Toulouse, France. ²Artificial and Natural Intelligence Toulouse Institute (ANITI), Toulouse, France.

Received: 13 December 2021 Accepted: 3 August 2022

Published online: 19 September 2022

References

1. M. Sajjadi, M. Javanmardi, T. Tasdizen, in *Advances in Neural Information Processing Systems*, vol. 29, ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Regularization with stochastic transformations and perturbations for deep semi-supervised learning (Curran Associates, Inc., 2016), Barcelona, pp. 1163–1171.
2. S. Laine, T. Aila, in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*. Temporal ensembling for semi-supervised learning (OpenReview.net, 2017). <https://openreview.net/forum?id=BJ6oOfqge>. Accessed 11 Aug 2022
3. T. Miyato, S. -i. Maeda, M. Koyama, S. Ishii, Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning (2018). <http://arxiv.org/abs/1704.03976>. Accessed 11 Aug 2022
4. Y. Grandvalet, Y. Bengio, in *Advances in Neural Information Processing Systems*, vol. 17, ed. by L. Saul, Y. Weiss, and L. Bottou. Semi-supervised learning by entropy minimization (MIT Press, 2005), Vancouver, pp. 529–536.
5. D. -H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proc. ICML 2013 Workshop: Challenges in Representation Learning (WREPL)*, Atlanta.
6. I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization (2019). <http://arxiv.org/abs/1711.05101>. Accessed 11 Aug 2022
7. G. Zhang, C. Wang, B. Xu, R. Grosse, Three Mechanisms of Weight Decay Regularization (2018). <http://arxiv.org/abs/1810.12281>. Accessed 11 Aug 2022
8. H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond Empirical Risk Minimization (2018). <http://arxiv.org/abs/1710.09412>. Accessed 11 Aug 2022
9. R. R. Wiyatno, A. Xu, O. Dia, A. de Berker, Adversarial Examples in Modern Machine Learning: A Review (2019). <http://arxiv.org/abs/1911.05268>. Accessed 11 Aug 2022
10. A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results (2018). <http://arxiv.org/abs/1703.01780>. Accessed 11 Aug 2022
11. S. Qiao, W. Shen, Z. Zhang, B. Wang, A. Yuille, in *Proc. ECCV*. Deep co-training for semi-supervised image recognition (Munich, 2018), Springer, pp. 135–152.
12. D. Berthelot, N. Carlini, I. Goodfellow, A. Oliver, N. Papernot, C. Raffel, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. MixMatch: a holistic approach to semi-supervised learning, Article 454 (Curran Associates Inc., Red Hook, 2019), pp. 5049–5059.
13. D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, C. Raffel, ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring (2020). <http://arxiv.org/abs/1911.09785>. Accessed 11 Aug 2022
14. K. Sohn, D. Berthelot, C. -L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, C. Raffel, FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence (2020). <http://arxiv.org/abs/2001.07685>. Accessed 11 Aug 2022
15. L. Cances, T. Pellegrini, in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Comparison of deep co-training and mean-teacher approaches for semi-supervised audio tagging (IEEE, 2021), pp. 361–365.
16. S. Grollmisch, E. Cano, Improving semi-supervised learning for audio classification with fixmatch. *Electronics*. **10**(15), 1807 (2021).
17. K. J. Piczak, in *Proc. ACM Multimedia*. Esc: Dataset for environmental sound classification (Brisbane, 2015), pp. 1015–1018. <https://doi.org/10.1145/2733373.2806390>.
18. J. Salamon, C. Jacoby, J. P. Bello, in *Proc. ACM Multimedia*. A dataset and taxonomy for urban sound research, (2014), pp. 1041–1044. <https://doi.org/10.1145/2647868.2655045>.
19. P. Warden, Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *arxiv:1804.03209* (2018). <http://arxiv.org/abs/1804.03209>. Accessed 11 Aug 2022
20. X. J. Zhu, Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences (2005).
21. O. Chapelle, M. Chi, A. Zien, in *Proceedings of the 23rd International Conference on Machine Learning*. A continuation method for semi-supervised svms, (2006), pp. 185–192.
22. J. E. Van Engelen, H. H. Hoos, A survey on semi-supervised learning. *Mach. Learn.* **109**(2), 373–440 (2020).
23. L. Jia Kai, Mean teacher convolution system for dcase 2018 task 4. Technical report, DCASE Challenge, Surrey (2018).
24. N. Turpault, R. Serizel, A. Parag Shah, J. Salamon, in *Workshop on Detection and Classification of Acoustic Scenes and Events*. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis (New York City, 2019). <https://hal.inria.fr/hal-02160855>. Accessed 11 Aug 2022
25. K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, K. Takeda, Convolution-augmented transformer for semi-supervised sound event detection. Technical report, DCASE2020 Challenge (2020).
26. Z. Shi, L. Liu, H. Lin, R. Liu, A. Shi, in *Proc. DCASE Workshop*. Hodgepodge: Sound event detection based on ensemble of semi-supervised learning methods, (2019), Ed. University of New York, New York, pp. 224–228.
27. K. Lu, C. -S. Foo, K. K. Teh, H. D. Tran, V. R. Chandrasekhar, in *Proc. INTERSPEECH*. Semi-supervised audio classification with consistency-based regularization (Graz, 2019), Ed. ISCA, pp. 3654–3658.
28. T. DeVries, G. W. Taylor, Improved Regularization of Convolutional Neural Networks with Cutout (2017). <http://arxiv.org/abs/1708.04552>. Accessed 11 Aug 2022
29. T. Ko, V. Peddinti, D. Povey, S. Khudanpur, in *Proc. Interspeech*. Audio augmentation for speech recognition (Dresden, 2015), Ed. ISCA, pp. 3586–3589.
30. A. Blum, T. Mitchell, in *Proc. COLT*. Combining labeled and unlabeled data with co-training (Madison, 1998), ACM, pp. 92–100.
31. M. -A. Krogel, T. Scheffer, in *Machine Learning*. Multi-relational learning, text mining, and semi-supervised learning for functional genomics, Vol 57, (2004), Kluwer Academic Publishers, pp. 61–81.
32. I. Goodfellow, J. Shlens, C. Szegedy, in *Proc. ICLR*. Explaining and harnessing adversarial examples (San Diego, 2015), Ed. Yoshua Bengio, Yann LeCun.

33. S. Zagoruyko, N. Komodakis, Wide Residual Networks (2017). <http://arxiv.org/abs/1605.07146>. Accessed 11 Aug 2022
34. A. F. Agarap, Deep Learning using Rectified Linear Units (ReLU) (2019). <http://arxiv.org/abs/1803.08375>. Accessed 11 Aug 2022
35. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, in *Proc. NeurIPS*. Pytorch: An imperative style, high-performance deep learning library, (2019), pp. 8026–8037. <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>. Accessed 11 Aug 2022
36. D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization (2017). <http://arxiv.org/abs/1412.6980>. Accessed 11 Aug 2022
37. A. Gazneli, G. Zimmerman, T. Ridnik, G. Sharir, A. Noy. End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network. arXiv preprint arXiv: 2204.11479 (2022).
38. A. Guzhov, F. Raue, J. Hees, A. Dengel, ESResNet: Environmental Sound Classification Based on Visual Domain Models. arXiv (2020). <https://doi.org/10.48550/ARXIV.2004.07301>.
39. A. Gazneli, G. Zimmerman, T. Ridnik, G. Sharir, A. Noy, End-to-End Audio Strikes Back: Boosting Augmentations Towards An Efficient Audio Classification Network. arXiv (2022). <https://doi.org/10.48550/ARXIV.2204.11479>.
40. R. Vygion, N. Mikhaylovskiy, in *Speech and Computer*. Learning efficient representations for keyword spotting with triplet loss (Springer, 2021), pp. 773–785. https://doi.org/10.1007/978-3-030-87802-3_69.
41. J. Salamon, J. P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.* **24**(3), 279–283 (2017). <https://doi.org/10.1109/lsp.2017.2657381>.
42. D. S. Park, W. Chan, Y. Zhang, C. -C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779 (2019).
43. J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter, in *Proc. IEEE ICASSP 2017, Ed. IEEE*. Audio set: An ontology and human-labeled dataset for audio events (New Orleans, 2017).
44. E. Fonseca, X. Favory, J. Pons, F. Font, X. Serra, FSD50K: an Open Dataset of Human-Labeled Sound Events (2020). <http://arxiv.org/abs/2010.00475>. Accessed 11 Aug 2022
45. Q. Xie, Z. Dai, E. Hovy, M. -T. Luong, Q. V. Le, Unsupervised Data Augmentation for Consistency Training (2020). <http://arxiv.org/abs/1904.12848>. Accessed 11 Aug 2022
46. H. Pham, Z. Dai, Q. Xie, M. -T. Luong, Q. V. Le, Meta Pseudo Labels (2021). <http://arxiv.org/abs/2003.10580>. Accessed 11 Aug 2022

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

