Adversarial Unsupervised Domain Adaptation for Harmonic-Percussive Source Separation

C. Lordelo, *Student Member, IEEE*, E. Benetos, *Senior Member, IEEE*, S. Dixon, S. Ahlbäck, and P. Ohlsson

***Abstract*—This letter addresses the problem of domain adaptation for the task of music source separation. Using datasets from two different domains, we compare the performance of a deep learning-based harmonic-percussive source separation model under different training scenarios, including supervised joint training using data from both domains and pre-training in one domain with fine-tuning in another. We propose an adversarial unsupervised domain adaptation approach suitable for the case where no labelled data (ground-truth source signals) from a target domain is available. By leveraging unlabelled data (only mixtures) from this domain, experiments show that our framework can improve separation performance on the new domain without losing any considerable performance on the original domain. The letter also introduces the Tap & Fiddle dataset, a dataset containing recordings of Scandinavian fiddle tunes along with isolated tracks for "foot-tapping" and "violin".**

***Index Terms*—Source separation, domain adaptation, semi-supervised learning, transfer learning.**

I. Introduction

BLIND source separation (BSS) is a fundamental problem in signal processing. It consists of separating a set of mixture signals into a set of source signals without using any extra information [1]. In this work, we will be considering the task of Music Source Separation (MSS), which is an ill-posed and underdetermined case of BSS, where multiple sources (instrumental signals) must be separated from a single mixture (music recording). Current MSS methods are based on Deep Neural Networks (DNNs) that need a lot of labelled data (mixtures and ground-truth isolated instrumental signals) to be trained under a supervised scenario [2], [3]. However, labelled audio data for MSS is difficult to obtain. In the literature, there are only a few large-scale public datasets for MSS, such as MUSDB18 [4] and Slakh [5].

Even though it is known that the use of data augmentation techniques such as random pitch-shifting and random mixing of source signals can improve model generalisation [6], [7], separation performance will always depend on the type of audio data used during training. When the data distribution of the training set is different from the data distribution of the test set, the performance of any predictor is degraded. This effect is known as dataset shift [8], and happens due to mismatched characteristics between data used for training and testing.

Under this scenario, domain adaptation techniques address this problem by adapting predictors from a *source domain*, where usually a large amount of labelled data is available, to a *target domain*, where only few or no labelled data is available. Domain adaptation is already consolidated as an important research topic in computer vision, where it is used in complex classification tasks [9]. Even in closer fields, such as acoustic scene analysis [10], [11], speech recognition [12] and speech enhancement [13], domain adaptation methods have already been proposed. However, to our knowledge, methods of this nature have not yet been investigated for MSS. Therefore, our work also attempts to fill this gap in the literature.

We propose an adversarial unsupervised domain adaptation approach for MSS. By using the mixtures and the available ground-truth signals from MUSDB18 and a set of unlabelled data (mixtures) from a different domain, we show that our framework is able to improve separation performance in the new domain while maintaining the original performance on MUSDB18, considerably reducing the degradation effect caused by dataset shift. Although our experiments are carried out for the particular task of Harmonic-Percussive Source Separation (HPSS), our framework can be easily adapted to other MSS tasks with different types of sources and domains.

In summary, our contributions include:

- The first work focused on unsupervised domain adaptation for MSS;
- An adversarial unsupervised domain adaptation framework for MSS that can be used with any neural network architecture, any type of audio representation and any number of sources;
- The public release of the "Tap & Fiddle Dataset," a dataset containing recordings of traditional Scandinavian fiddle tunes with accompanying foot-tapping along with isolated tracks for "foot-tapping" and "violin". This dataset has different timbral characteristics than MUSDB18 and is useful for domain adaptation experiments;

Manuscript received October 2, 2020; revised December 10, 2020; accepted December 11, 2020. Date of publication December 18, 2020; date of current version January 15, 2021. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Grant 765068 - MIP Frontiers Project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Odette Scharenborg. *(Corresponding author: Carlos Lordelo.)*

C. Lordelo is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K., and also with Doremir Music Research AB, 11140 Stockholm, Sweden (e-mail: c.p.viannalordelo@qmul.ac.uk).

E. Benetos and S. Dixon are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: emmanouil.benetos@qmul.ac.uk; s.e.dixon@qmul.ac.uk).

S. Ahlbäck and P. Ohlsson are with Doremir Music Research AB, 11140 Stockholm, Sweden (e-mail: sven.ahlback@kmh.se; patrik.ohlsson@doremir.com).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LSP.2020.3045915, provided by the authors

Digital Object Identifier 10.1109/LSP.2020.3045915

- A prototype experiment where we show an improvement over benchmark methods for the HPSS task.

II. Related Work

A. Harmonic-Percussive Source Separation

The task of HPSS consists of separating a music signal into two source signals, one with the harmonic components and other with the percussive sounds [14]. Signal processing methods for HPSS perform separation by exploiting the fact that percussive signals form vertical lines in the mixture time-frequency representation, while the harmonic components tend to form horizontal structures, e.g. [15]–[17]. However, due to their strict assumptions and hand-crafted features, methods of this nature have intrinsic performance limitations.

Over the years, data-driven approaches have shown significant improvements over traditional methods for HPSS and current state-of-the-art methods are based on DNNs [18]–[21]. In previous work carried out by the authors [21], the 3W-MDenseNet, an encoder-decoder DNN that uses convolutions with several kernel shapes to perform HPSS, was proposed. In this work, the same architecture is used, but here we add a domain discriminator into the framework and modify the loss function to support adversarial domain adaptation.

Moreover, since our approach is also grounded in Generative Adversarial Networks (GANs) [22], it is important to point out some key aspects in which our proposal is different from other GAN-based source separation methods [23]–[25].

1) Discriminator: Works on GAN-based MSS use a *source discriminator*, which is trained to differentiate *real* source signals from *fake* source signals. This is different from our work, where we use a *domain discriminator* trained to differentiate mixtures across two different domains.

2) Unlabelled Data: In order to train a source discriminator, a large number of single-source signals are required, even though those signals do not necessarily have to be paired with a music mixture. Here, we only need mixtures from each of the two domains to successfully train our domain discriminator.

3) Input to Discriminator: The input to a source discriminator of GAN-based MSS works is the *output* of the separator network. Our approach applies the domain discriminator on the *encoded feature-maps*, in the middle of the separator network and not directly on its output.

B. Domain Adaptation

Domain adaptation methods can be either supervised or unsupervised depending on the type of data from the target domain that is used. While Supervised Domain Adaptation (SDA) methods use labelled data, Unsupervised Domain Adaptation (UDA) exploits only unlabelled data (mixture signals) from the target domain.

A typical SDA approach is to first train a model using a large number of labelled samples from the source domain and then retrain some (or all) of its layers using a smaller labelled dataset of interest (target domain). This technique is known as *fine-tuning* [26], [27]. Another SDA approach is *joint training*, where the two datasets are merged into a new dataset and only a single training stage is done, using labelled data from both domains in every batch [5], [28].

UDA methods usually consider that the system is under the *covariate shift paradigm*, assuming that, even though the marginal distribution of source domain data is different from the marginal distribution of target domain data, the conditional probability of the output remains the same. Therefore, if the marginal distributions can be matched, the same predictor can be applied successfully over samples from either of the two domains [29]. In order to do this, some UDA methods propose to re-weight [30] or select samples from the source domain [31], while others project the data through an embedding function such that not only the marginals become similar on the embedded space, but also the embedded features keep their discrimination potential [32], [33]. The latter case is also the type of UDA method in our proposal. We look for a transformation that creates an embedded space in which the confusion between the two domains is maximised. Similar to [34], we propose to find a *domain-invariant* and *separation-discriminative* embedded space that is learned from data via adversarial training. However, differently from [34], we deal with the task of source separation (regression) instead of image recognition (classification). In addition, we use CNNs for the encoder-decoder and the domain discriminator, while in [34] simple feed-forward networks are used, and while [34] performs adversarial training using the gradient reversal layer method, we conduct conditional GAN iterative optimisation as in [22].

III. Proposed Framework

We assume that both the input data and the outputs are $F \times T$ magnitude spectrograms, where F is the number of frequency bins and T the number of frames. To simplify the notation, we treat them as vectors in $\mathbb{R}^K$, where $K = FT$. Hence, the input (mixture signal) is notated as $\mathbf{x}$ and its labels (ground-truth isolated source signals) as the $K \times 2$ matrix $\mathbf{Y} = [\mathbf{h}\,\mathbf{p}]$, where the first column is the original harmonic vector $\mathbf{h} \in \mathbb{R}^K$ and the second column is the original percussive vector $\mathbf{p} \in \mathbb{R}^K$. Furthermore, we consider that the mixture-label pairs follow the joint distribution $p_{\mathcal{A}}(\mathbf{x}, \mathbf{Y})$, or, in other words, we say that the data "come from domain $\mathcal{A}$". For the general supervised HPSS case, the goal is to train a model based on this data that can be a good predictor of $p(\mathbf{Y}|\mathbf{x} \sim p_{\mathcal{A}}(\mathbf{x}))$.

In [21] we proposed the 3W-MDenseNet, a convolutional encoder-decoder for HPSS, where the network output is an estimate $\hat{\mathbf{Y}} = [\hat{\mathbf{h}}\,\hat{\mathbf{p}}]$ of $\mathbf{Y}$. Here, we model the encoder-decoder-based separation process as a sequence of two mappings. First, the encoder $\mathcal{E}$ with parameters $\theta_{\mathcal{E}}$ maps the input to an embedded feature space $\mathbf{z} = \mathcal{E}(\mathbf{x}; \theta_{\mathcal{E}})$ and then the decoder $\mathcal{D}$, with parameters $\theta_{\mathcal{D}}$, maps $\mathbf{z}$ to the output $\hat{\mathbf{Y}}$ such that:

$$\hat{\mathbf{Y}} = \mathcal{D}(\mathbf{z}; \theta_{\mathcal{D}}) = \mathcal{D}(\mathcal{E}(\mathbf{x}; \theta_{\mathcal{E}}); \theta_{\mathcal{D}}). \tag{1}$$

This separator can be optimised for the general supervised HPSS case using the mean square error as the loss $\mathcal{L}_{\mathrm{S}}$ [21]:

$$\begin{aligned}\mathcal{L}_{\mathrm{S}}(\theta_{\mathcal{E}}, \theta_{\mathcal{D}}) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\mathcal{A}}(\mathbf{x})}\left[\lambda_h ||\hat{\mathbf{h}} - \mathbf{h}||^2 + \lambda_p ||\hat{\mathbf{p}} - \mathbf{p}||^2\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\mathcal{A}}(\mathbf{x})}\left[||(\hat{\mathbf{Y}} - \mathbf{Y})\mathbf{\Lambda}||_F^2\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\mathcal{A}}(\mathbf{x})}\left[||(\mathcal{D}(\mathcal{E}(\mathbf{x}; \theta_{\mathcal{E}}); \theta_{\mathcal{D}}) - \mathbf{Y})\mathbf{\Lambda}||_F^2\right], \end{aligned} \tag{2}$$

where λ_h and λ_p are weights for the harmonic and percussive outputs respectively — we use 0.5 for each since we want to assign equal importance to each source —, $||\dots||$ represents

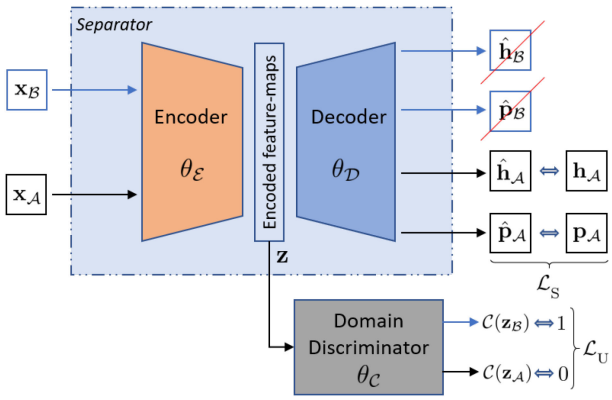

Fig. 1. Schematic of proposed adversarial UDA for HPSS.

the Euclidean norm, $||\ldots||_F$ the Frobenius norm and $\mathbf{\Lambda}$ is the diagonal matrix $\left[\begin{smallmatrix}\sqrt{\lambda_h} & 0\\ 0 & \sqrt{\lambda_p}\end{smallmatrix}\right]$.

However, in this work we assume there also exists a new domain $\mathcal{B}$, where mixtures follow the marginal distribution $p_\mathcal{B}(\mathbf{x})$, which is considered different from $p_\mathcal{A}(\mathbf{x})$. Our main goal is now to be able to robustly predict labels $\hat{\mathbf{Y}}$ given that the input can be from either domain $\mathcal{A}$ or $\mathcal{B}$. Apart from the labelled samples from domain $\mathcal{A}$, we have access to set of mixtures from $\mathcal{B}$ that can be used for performing UDA.

Our approach adopts a similar methodology to [34] and [35]. We propose to learn encoded features $\mathbf{z}$ that can not only guarantee a good separation performance, but that are also invariant to domain changes. This means that $\mathbf{z}$ must not contain any discriminative information about the origin of the input ($\mathcal{A}$ or $\mathcal{B}$). By doing so, we can make the distributions $p(\mathbf{z}|\ \mathbf{x} \sim p_\mathcal{A}(\mathbf{x})) = \{\mathcal{E}(\mathbf{x};\theta_\mathcal{E})|\ \mathbf{x} \sim p_\mathcal{A}(\mathbf{x})\}$ and $p(\mathbf{z}|\ \mathbf{x} \sim p_\mathcal{B}(\mathbf{x})) = \{\mathcal{E}(\mathbf{x};\theta_\mathcal{E})|\ \mathbf{x} \sim p_\mathcal{B}(\mathbf{x})\}$ to become as similar as possible. In order to measure their similarity, we use a domain discriminator $\mathcal{C}(\mathbf{z}, \theta_\mathcal{C})$ to discriminate the encoded feature-maps between the two domains. Such domain discriminator is a binary classifier that can be trained using only mixture signals by minimising the binary cross-entropy $\mathcal{L}_\mathrm{U}$:

$$\mathcal{L}_\mathrm{U}(\theta_\mathcal{C}, \theta_\mathcal{E}) = -\mathop{\mathbb{E}}_{\mathbf{z}\sim p_\mathcal{B}(\mathbf{z})}\big[\log \mathcal{C}(\mathbf{z}, \theta_\mathcal{C})\big] - \mathop{\mathbb{E}}_{\mathbf{z}\sim p_\mathcal{A}(\mathbf{z})}\big[\log(1-\mathcal{C}(\mathbf{z}, \theta_\mathcal{C}))\big]. \tag{3}$$

Fig. 1 summarises the domain adaptation scenario.

In addition, we ensure that $\mathbf{z}$ will become domain-invariant by forcing the encoder sub-network to generate feature-maps that can fool the domain discriminator. This is achieved by maximising $\mathcal{L}_\mathrm{U}$ when training the encoder weights. Such a min-max game is played by the encoder sub-network and the domain discriminator during training just like in GAN training [22]. At the same time, $\mathbf{z}$ can keep its separation-discriminative properties if we include the minimisation of $\mathcal{L}_\mathrm{S}$ in the loss function. The final encoder loss is, therefore, a combination of the (unsupervised) *adversarial* loss $\mathcal{L}_\mathrm{U}$, which can be optimised using only mixture signals from each of the two domains, and the (supervised) loss $\mathcal{L}_\mathrm{S}$, which can be optimised based only on samples from $\mathcal{A}$ since it requires labelled data. In summary, the loss functions of each sub-network are:

$$\hat{\theta}_\mathcal{C} = \arg\min_{\theta_\mathcal{C}} \mathcal{L}_\mathrm{U}(\theta_\mathcal{E}, \theta_\mathcal{C}) \tag{4}$$

$$\hat{\theta}_\mathcal{E} = \arg\min_{\theta_\mathcal{E}} \Big[-\gamma_\mathrm{U}\mathcal{L}_\mathrm{U}(\theta_\mathcal{E}, \hat{\theta}_\mathcal{C}) + \gamma_\mathrm{S}\mathcal{L}_\mathrm{S}(\theta_\mathcal{E}, \hat{\theta}_\mathcal{D})\Big] \tag{5}$$

$$\hat{\theta}_\mathcal{D} = \arg\min_{\theta_\mathcal{D}} \mathcal{L}_\mathrm{S}(\theta_\mathcal{E}, \theta_\mathcal{D}) \tag{6}$$

where γ_U and γ_S are weights given to the unsupervised part and to the supervised part of the loss.

It should be noted that $\mathcal{C}$, $\mathcal{E}$ and $\mathcal{D}$ must be trained together in an iterative way as in GAN training [22]. If $\mathcal{C}$ is optimised to completion, the encoder sub-network will not be able to increase the domain-discriminator confusion, causing the separator performance to overfit over domain $\mathcal{A}$ [22]. In our experiments, at every training iteration, we perform 5 updates on $\theta_\mathcal{C}$ before updating $\theta_\mathcal{E}$ and $\theta_\mathcal{D}$. The full training algorithm can be found in the supplementary material of this letter.

IV. Datasets

MUSDB18 [4] is the largest public dataset for MSS containing real-world audio recordings. It contains full-track songs and includes both the mixtures and the original sources, divided between a training subset of 100 music recordings and a test subset of 50. The available isolated tracks are vocals, bass, drums and "other". We use the drum track as the ground-truth for the percussive source, while the sum of the other tracks is used as ground-truth for the harmonic source.

As a different domain, we collected and publicly release the Tap & Fiddle (T&F) dataset [36]. The T&F dataset contains stereo recordings of traditional Scandinavian fiddle tunes with accompanying foot-tapping, which is standard performance practice within these musical styles. It consists of 28 recordings with completely separate fiddle and foot-tapping sounds as well as mixed signals. The dataset is divided into a training set with 23 files and a test set with 5. All recordings are solo and have an average duration of 65 seconds. Detailed information regarding the T&F Dataset can be found in [36].

V. Experimental Setup

In our experiments, the music signals are converted to mono and resampled to 16 KHz. The inputs are normalised magnitude spectrograms of size 256×256 generated by the application of an STFT of size 512 with 75% overlap. A validation split of 20% of all labelled data available for training is set.

We use the 3W-MDenseNet [21] as the separator architecture. As a post-processing step, we apply Wiener filtering [37] to the source estimates and use the mixture phase to return to the time domain. We concatenate the encoded feature-maps of each of the three branches of the 3W-MDenseNet to form $\mathbf{z}$. Details about hyper-parameter choices can be found in the letter's supplementary material. The architecture of the domain-discriminator network is depicted in Fig. 2.

After experimentation, we choose the values of 1 for γ_S and 0.001 for γ_U. Training is performed using the Adam optimiser with an initial learning rate of 0.001, which is reduced by a factor of 0.25 if the supervised validation loss $\mathcal{L}_\mathrm{S}$ stops improving for 50 consecutive epochs, and if no improvement happens in 200 epochs the training is stopped. The separation quality is evaluated using the BSS_eval [38] set of objective metrics that are largely used by the MSS community.

TABLE I
OBJECTIVE EVALUATION OF HPSS ON MUSDB18 AND TAP & FIDDLE. THE VALUES ARE IN DB AND REPRESENT THE MEDIAN OF METRICS OVER TRACKS IN EACH TEST SET. IBM IS THE IDEAL BINARY MASKING AND IRM REPRESENTS THE IDEAL RATIO MASKING ORACLE METHODS

Method (Training Set)	Test Set												Type of Data	
	MUSDB18 (Domain $\mathcal{A}$)						Tap & Fiddle (Domain $\mathcal{B}$)							
	Percussive			Harmonic			Percussive			Harmonic				
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	$\mathcal{A}$	$\mathcal{B}$
HPSS_MUSDB ($\mathcal{A}$)	4.5	13.0	5.0	10.0	13.4	12.3	1.3	15.8	0.3	22.0	23.0	29.7	labelled	—
HPSS_T&F ($\mathcal{B}$)	−0.2	0.3	10.5	3.1	16.5	5.2	10.2	16.9	12.7	35.0	36.4	34.3	—	labelled
SDA_joint ($\mathcal{A}+\mathcal{B}$)	4.8	13.3	5.1	10.2	13.9	12.1	4.6	18.1	6.4	27.5	28.9	30.2	labelled	labelled
SDA_tuned ($\mathcal{A}\rightarrow\mathcal{B}$)	2.9	8.6	3.3	7.1	9.3	10.5	12.1	18.8	12.6	35.3	37.1	35.6	labelled	labelled
UDA_small	4.8	12.2	5.1	10.0	13.4	11.8	3.4	13.0	2.9	25.0	25.9	30.8	labelled	unlabelled
UDA_large	4.6	12.9	4.9	10.1	14.1	12.0	7.4	18.0	8.4	29.2	30.6	33.1	labelled	unlabelled
OpenUnmix [2]	5.2	11.2	6.0	10.1	17.7	10.7	6.7	7.0	5.1	28.6	36.8	25.9	labelled	—
IBM	7.8	16.4	7.9	11.9	17.9	13.2	13.5	20.8	13.7	37.8	41.3	37.7	—	—
IRM	8.0	12.4	9.7	12.2	15.8	15.0	13.4	19.5	13.8	37.2	42.0	37.2	—	—

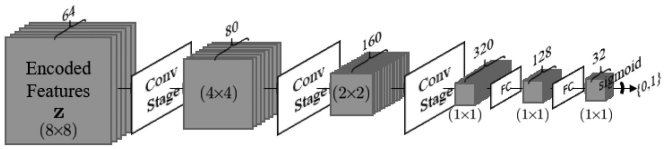

Fig. 2. Architecture of the domain discriminator. Each "Conv Stage" is a 3×3 convolutional layer followed by 2×2 max pooling. "FC" is a fully connected layer.

VI. RESULTS

Recordings from MUSDB18 represent domain $\mathcal{A}$ while recordings from the T&F dataset represent domain $\mathcal{B}$. We aim to investigate how different training scenarios perform across the two domains. We compare our UDA proposal to traditional supervised HPSS approaches that use only labelled data from one of the domains, to SDA frameworks, which include joint training using labelled data from both datasets and fine-tuning over samples from T&F after training on MUSDB18, and to another state-of-the-art DNN for MSS named OpenUnmix [2]. This method was previously trained on an augmented version of MUSDB18 and serves as a baseline in our comparison.

In addition to the mixtures in the T&F dataset, we have a collection of 50 new recordings of Scandinavian fiddle tunes with accompanying foot-tapping. This collection is also part of domain $\mathcal{B}$ and although no labels are available, it can also be used by our UDA method. We then test two versions of our approach: HPSS_UDA_small, which uses the mixtures on the train set of T&F for performing the adaptation to domain $\mathcal{B}$, and HPSS_UDA_large, which uses the larger set of mixtures from our internal collection. Results are shown in Table I.

By inspecting Table I, we can readily note that models that were trained only with samples from one dataset had poor performance on the other, which makes it possible to conclude that MUSDB18 and T&F have very different priors over the data. This fact is also reflected in the performance of OpenUnmix, which is much lower on T&F if compared with the performance provided by the ideal masking methods. Moreover, as expected, the joint trained model, SDA_joint, achieved relatively good performance overall because it uses supervised data from both domains. The SDA_tune model, which is the HPSS_MUSDB model fine-tuned for T&F, was indeed greatly improved when evaluated over this domain, but, as a trade-off, it lost a lot of its original performance on the original MUSDB18 dataset. On the other hand, both versions of the proposed UDA approach got a boost in performance on all 3 of the metrics on T&F without losing any considerable performance on MUSDB18. This means that our proposed UDA approach can perform HPSS on both domains successfully, even though the labelled data used for training came only from domain $\mathcal{A}$.

The quantity of unlabelled data from domain $\mathcal{B}$ also impacted the performance of the proposed method. Even though the results of UDA_large are similar to UDA_small over domain $\mathcal{A}$, the former performs much better over samples from domain $\mathcal{B}$ than the latter due to the fact that it uses more than double the amount of mixtures from this particular domain during training to perform domain adaptation. Another interesting result is that UDA_large, which is a semi-supervised framework, had similar performance over MUSDB18, but much better over T&F if compared to SDA_joint, which is a fully supervised method. This means that UDA using large amounts of unlabelled data can be much more promising than joint training using a smaller amount of labelled data.

More information about our work can be found in the letter's supplementary document and supplementary webpage.[1]

VII. CONCLUSIONS

In this work we presented an adversarial UDA model for HPSS. Our proposal is a semi-supervised framework that is able to exploit unlabelled mixtures from a target domain in order to improve HPSS generalisation to samples from this particular domain. Results showed that our framework improves separation performance on the target domain without losing considerable performance on the source domain.

As future work, we plan to investigate how the utilisation of small amounts of labelled samples from the target domain affect domain adaptation performance. We believe that this "few-shot" approach can be useful in improving source separation performance in the absence of many data samples.

[1] http://c4dm.eecs.qmul.ac.uk/auda-hpss

REFERENCES

[1] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation*, 1st ed. New York, NY, USA: Academic, 2010.

[2] F. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix—A reference implementation for music source separation," *J. Open Source Softw.*, vol. 4, no. 41, pp. 1667, 2019.

[3] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, vol. 16, Sep. 2018, pp. 106–110.

[4] Z. Rafii, A. Liutkus, F. Stöter, S. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[5] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2019, pp. 45–49.

[6] S. Uhlich *et al.*, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 261–265.

[7] A. Cohen-Hadria, A. Roebel, and G. Peeters, "Improving singing voice separation using deep U-Net and Wave-U-Net with data augmentation," in *Proc. IEEE Eur. Signal Process. Conf.*, Sep. 2019, pp. 1–5.

[8] J. Quionero-Candela, A. Sugiyama, Masashi Schwaighofer, and N. Lawrence, Eds., *Dataset Shift in Machine Learning*. Cambridge, MA, USA: MIT Press, 2008.

[9] X. Peng, Q. Bai, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*, Seoul, South Korea, 2019, pp. 1406–1415.

[10] S. Gharib, K. Drossos, E. Çakir, D. Serdyuk, and T. Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," in *Proc. Detect. Classification Acoust. Scenes and Events Workshop*, Surrey, U.K., Nov. 2018, pp. 138–142.

[11] W. Wei, H. Zhu, E. Benetos, and Y. Wang, "A-CRNN: A domain adaptation model for sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2020, pp. 276–280.

[12] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, Sep. 2017.

[13] Z. Meng, J. Li, Y. Gong, and B.-H. Juang, "Adversarial feature-mapping for speech enhancement," in *Proc. Interspeech*. Hyderabad, India: ISCA, Sep. 2018, pp. 3259–3263.

[14] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. Eur. Signal Process. Conf.*, Lausanne, Switzerland, Aug. 2008, pp 1–4.

[15] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. Int. Conf. Digit. Audio Effects*, Graz, Austria, vol. 13, Sep. 2010, pp. 246–253.

[16] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Oct. 2014, vol. 15, pp. 611–616.

[17] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4298–4310, Aug. 2014.

[18] G. Roma, O. Green, and P. A. Tremblay, "Stationary/transient audio separation using convolutional autoencoders," in *Proc. Int. Conf. Digit. Audio Effects*, Sep. 2018, pp. 65–71.

[19] W. Lim and T. Lee, "Harmonic and percussive source separation using a convolutional auto encoder," in *Proc. IEEE Eur. Signal Process. Conf.*, Sep. 2017, vol. 25, pp. 1804–1808.

[20] K. Drossos, P. Magron, S. I. Mimilakis, and T. Virtanen, "Harmonic-percussive source separation with deep neural networks and phase recovery," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2018, vol. 16, pp. 421–425.

[21] C. Lordelo, E. Benetos, S. Dixon, and S. Ahlbäck, "Investigating kernel shapes and skip connections for deep learning-based harmonic-percussive separation," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2019, pp. 40–44.

[22] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Conf. Neural Inf. Process. Syst.*, Montreal, Canada, Dec. 2014, vol. 28, pp. 2672–2680.

[23] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 2391–2395.

[24] Z. C. Fan, Y. L. Lai, and J. S. Jang, "SVSGAN: Singing voice separation via generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 726–730.

[25] Y. Z. Ong, C. K. Chui, and H. Yang, "CASS: Cross adversarial source separation via autoencoder," May 2019, *arXiv e-print:1905.09877*.

[26] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Sep. 2014, pp. 1717–1724.

[27] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, "Transferring GANs: Generating images from limited data," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 220–236.

[28] M. Maciejewski, G. Sell, L. P. Garcia-Perera, S. Watanabe, and S. Khudanpur, "Building corpora for single-channel speech separation across multiple domains," Nov. 2018, *arXiv e-print:1811.02641*.

[29] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Statist. Plan. Inference*, vol. 90, no. 2, pp. 227–244, Oct. 2000.

[30] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf, "Correcting sample selection bias by unlabeled data," in *Proc. Conf. Neural Inf. Process. Syst.*, Dec. 2006, pp. 601–608.

[31] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2013, pp. 222–230.

[32] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2960–2967.

[33] P. Xiao, B. Du, J. Wu, L. Zhang, R. Hu, and X. Li, "TLR: Transfer latent representation for unsupervised domain adaptation," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2018, pp. 1–6.

[34] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[35] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, Jan. 2016.

[36] C. Lordelo, S. Ahlbäck, E. Benetos, S. Dixon, and P. Ohlsson, "Tap & Fiddle: A dataset with Scandinavian fiddle tunes with accompanying foot-tapping," *Zenodo*, Dec. 2020. [Online]. Available: http://doi.org/10.5281/zenodo.4308731

[37] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel music separation with deep neural networks," in *Proc. Eur. Signal Process. Conf.*, Aug. 2016, pp. 1748–1752.

[38] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.