International Journal on Document Analysis and Recognition (IJDAR)
https://doi.org/10.1007/s10032-022-00411-w

SPECIAL ISSUE PAPER

# Domain adaptation for staff-region retrieval of music score images

Francisco J. Castellanos[1] · Antonio Javier Gallego[1] · Jorge Calvo-Zaragoza[1] · Ichiro Fujinaga[2]

Received: 15 March 2022 / Revised: 15 June 2022 / Accepted: 22 August 2022

## Abstract

Optical music recognition (OMR) is the field that studies how to automatically read music notation from score images. One of the relevant steps within the OMR workflow is the staff-region retrieval. This process is a key step because any undetected staff will not be processed by the subsequent steps. This task has previously been addressed as a supervised learning problem in the literature; however, ground-truth data are not always available, so each new manuscript requires a preliminary manual annotation. This situation is one of the main bottlenecks in OMR, because of the countless number of existing manuscripts , and the associated manual labeling cost. With the aim of mitigating this issue, we propose the application of a domain adaptation technique, the so-called Domain-Adversarial Neural Network (DANN), based on a combination of a gradient reversal layer and a domain classifier in the inference neural architecture. The results from our experiments support the benefits of our proposed solution, obtaining improvements of approximately 29% in the F-score.

**Keywords** Unsupervised domain adaptation · Staff retrieval · Music score images · Optical music recognition

## 1 Introduction

Optical Music Recognition (OMR) is the field that studies how to automatically read music notation from document images [4,5]. Among the motivations to develop this research field, one of the main ones is to make this cultural heritage computationally accessible. The challenge, however, is that there are countless manuscripts with myriad of differences among them, such as the notation type—modern, mensural, neumatic or squared, among others—the historical era, the engraving mode, or even the degree of page degradation associated with physical formats. This situation hinders the proper digital transcription of existing document collections scattered all over the world and demands a scalable solution.

One of the processes involved in the traditional OMR workflow is document analysis . This process attempts to recognize those parts of the document image that are relevant for eventually extracting the music sequence. The staff regions contain most of musical information, thus, recognizing these regions are an important step for transcribing their content. Recent research trends apply deep learning techniques to retrieve these staff regions, specifically, those based on convolutional neural networks (CNNs), which are used to learn the proper features in a supervised manner in order to find these information areas [7,22,33]. In other words, these models are first trained with a part of the image collection to be processed, then they are used with the rest of the collection to automatically carry out the staff extraction. However, this strategy requires a labeled training set, which typically has to be manually obtained and involves a very costly process, which hinders the application of these methods in general. A potential solution to mitigate this issue is domain adaptation.

This paper is part of the I+D+i PID2020-118447RA-I00 (MultiScore) project funded by MCIN/AEI/10.13039/501100011033. The first author acknowledges support from the "Programa I+D+i de la Generalitat Valenciana" through grants ACIF/2019/042 and CIBEFP/2021/72. This work also draws on research supported by the Social Sciences and Humanities Research Council (895-2013-1012) and the Fonds de recherche du Québec-Société et Culture (2022-SE3-303927).

✉ Francisco J. Castellanos
fcastellanos@dlsi.ua.es

Antonio Javier Gallego
jgallego@dlsi.ua.es

Jorge Calvo-Zaragoza
jcalvo@dlsi.ua.es

Ichiro Fujinaga
ichiro.fujinaga@mcgill.ca

[1] Department of Software and Computing Systems, University of Alicante, Alicante, Spain

[2] Schulich School of Music, McGill University, Montreal, Canada

Published online: 10 September 2022

Domain adaptation techniques seek to adapt the knowledge learned with an annotated dataset—hereinafter referred to as *source*—to other datasets for which labeling is not available—referred to as *target* domains—thus enabling data processing regardless of the dataset's domain. In this way, in our context, with a few annotated pages from a manuscript, the model may learn features useful to process any target manuscript without manually annotating the target manuscript. Domain adaptation techniques have shown their benefits in other unsupervised scenarios [14,15] but their use has not been explored for the staff-retrieval task. Note that, these techniques are not intended to reduce the time of the training process, but to eliminate the need for creating annotations for each target, which is normally carried out by hand.

In this work, we study the applicability of existing domain adaptation techniques, and particularly the so-called Domain-Adversarial Neural Network (DANN) [13], based on the use of a gradient reversal layer (GRL)—a special layer that inverts the gradients during the training process—to retrieve staff regions within music scores in an unsupervised manner. In order to evaluate this unsupervised domain adaptation technique in our context, we considered a large selection of corpora with very different graphic characteristics to validate the benefits of DANN in adapting knowledge from one manuscript to another.

In our experiments, we report the results as the difference from the baseline case, where no adaptation is applied and only the source domain is used for training the model. We aim to demonstrate, therefore, that domain adaptation can be a promising strategy for solving this task with minimum human intervention.

The rest of the paper is organized as follows: Sect. 2 reviews the state-of-the-art methods for staff recognition; Sect. 3 explains the methodology; Sect. 4 states the experimental setup, including the corpora and metrics considered for validating the method, as well as the architecture specifications; Sect. 5 shows the results obtained; and finally, our conclusions and future work are addressed in Sect. 6.

## 2 Background

Recent work in OMR focuses on the development of end-to-end strategies for extracting the music sequence from each individual staff [2,20,35]. However, to accomplish this task, it is necessary to perform a preliminary process of staff retrieval, which is responsible for providing each staff region to the sequence recognition model. This step of finding staff regions is the focus of this work.

Traditionally, this process has been performed by heuristic methods [1,10,31]. However, although the results may have been satisfactory for the datasets considered in their experiments, these methods leverage specific features on the images that might not be present in other collections, such as ink color, staff line spacing or thickness, or different musical notation, among others. This limitation reduces the possibilities for reusing these strategies in practice.

With the emergence of machine learning and deep learning techniques, more generalizable and accurate methods were developed. For example, different strategies were proposed for the extraction of staff regions [3,21,24,34]. Recently, Castellanos et al. [9] reviewed general-purpose object detection models for the extraction of regions of interest with music score images. One conclusion was that a Selectional AutoEncoder (SAE) architecture showed very competitive results compared with other well-known models such as faster region-based convolutional neural network (Faster R-CNN) [25] and RetinaNet [19]. SAE also has been demonstrated to be useful for this task in a full-page OMR workflow [7] and it can be considered as the state of the art in this regard. This architecture processes the image to obtain a probabilistic map in which each pixel is assigned a value representing the confidence with which the model detected a specific class, such as music symbols, staff lines, and background. A similar SAE-based strategy was also used for staff-lines removal, a very related task to staff-regions retrieval [12]. This task aims to detect and remove those pixels belonging to staff lines within the image [11,32]. Despite the similarity between the two tasks, note that the objective of staff-line removal differs from that of staff-region retrieval, as the latter aims to obtain a bounding box that surrounds the entire staff region, which usually includes music symbols that extend beyond the region defined by the staff lines alone.

Nevertheless, these methods only work in supervised cases, where staff regions must be manually annotated to learn features. This situation hinders the use of these techniques in practice because of the heterogeneity and the large number of manuscripts that need to be processed. A potential solution for this fact may be the use of unsupervised domain adaptation, which leverages the knowledge extracted from annotated data to process new data from a different domain for which the labeling is not available.

In other contexts, there are multiple issues facing unsupervised learning. For example, multiple strategies have been proposed for obtaining a domain-invariant feature representation while minimizing a measure of divergence between source and target domains [29,30,36]. Other methods try to obtain a common representation for both involved domains—source and target—in order to use the same model to process both , such as Deep Reconstruction Classification Network [14], or the proposal by Isola et al. [16], which transforms one domain into the another by means of a conditional generative adversarial network (GAN). Adversarial training also has been a relevant key for adapting the two domains by means of GAN [15] or GRL [13] , a special layer that inverts the gradients in the training process that was originally used by the DANN strategy for classification tasks. This layer was connected to a domain classifier in order to force the inference model to learn domain-invariant features. The latter strategy was successfully extended for pixel-level layout analysis [8], a related task to the one at issue in this work.

In the next section, we show how we adapted the method based on GRL in combination with SAE, which is the object detection model with the best results in supervised experiments for extracting staff regions of a non-annotated manuscript [9].

## 3 Method

Let $\mathcal{S}$ be the source domain, a collection of images with the respective annotations paired in the form $(\mathcal{S}^i_\mathcal{I}, \mathcal{S}^i_\mathcal{A})$, where $\mathcal{S}^i_\mathcal{I} = [0, 255]^{w^s_i \times h^s_i}$ is the $i$-th image with a size of $w^s_i$ pixels of width, $h^s_i$ pixels of height and $\mathcal{S}^i_\mathcal{A} = [\mathcal{B}^i_1, \mathcal{B}^i_2, ..., \mathcal{B}^i_b]$ represents the list of the $b$ bounding boxes of staves within $\mathcal{S}^i_\mathcal{I}$. Note that, each bounding box $\mathcal{B}^i$ could be represented as rectangles or polygons, depending on the limitations of the tool used for annotating them.[1] Let also $\mathcal{T}$ be the target domain, a collection of images from which its annotations are not available. Thus, given $\mathcal{T}^j_\mathcal{I} = [0, 255]^{w^t_j \times h^t_j}$, the $j$-th image in $\mathcal{T}$ with $w^t_j$ pixels of width and $h^t_j$ pixels of height, the goal of the method is to automatically obtain its respective staff bounding-box annotations $\mathcal{T}^j_\mathcal{A}$. Note that, the images could have different resolutions even within the same collection. In addition, we considered grayscale images, but other criteria could also be applied.

The method proposed for the unsupervised staff retrieval is based on two existing approaches to obtain the annotations from the target domain $\mathcal{T}^j_\mathcal{A}$: the SAE architecture, which was the state of the art in this task but for the supervised case [7], and the GRL, since it was successfully used for binarization, another common and related image processing task used in OMR [8]. Figure 1 shows the scheme of the proposed approach for the combination of these methods.

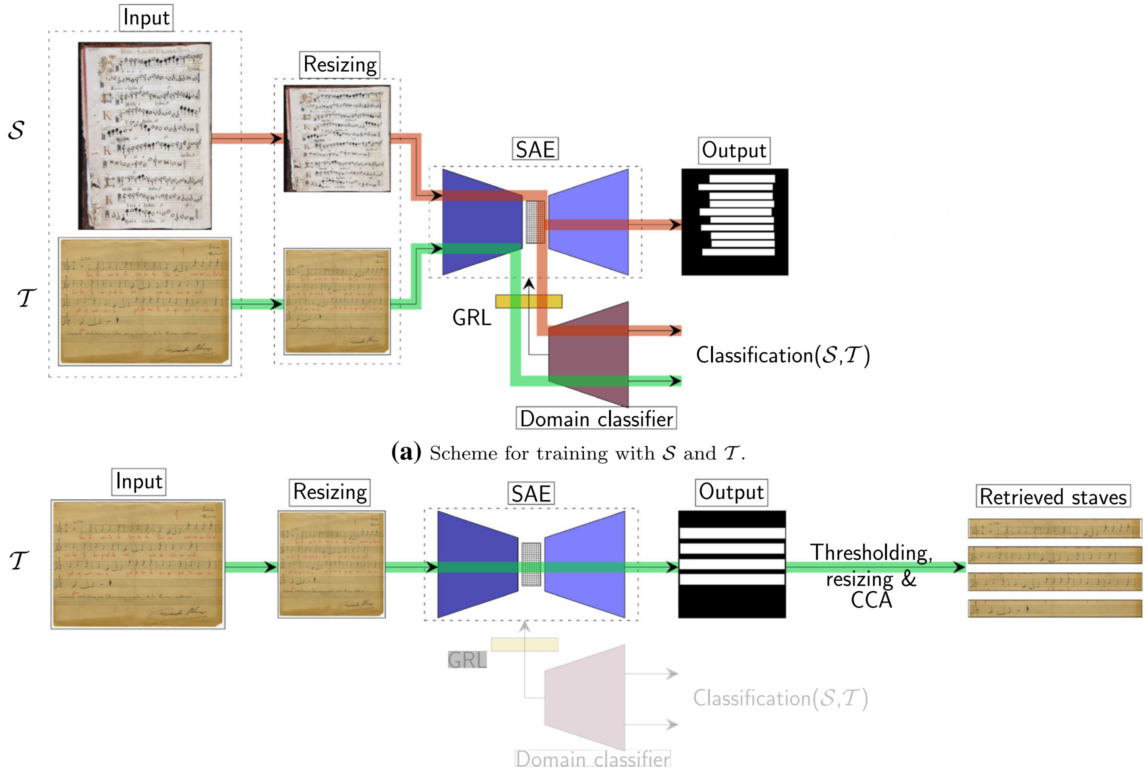
As can be seen in Fig. 1, the SAE architecture contains two main parts: an encoder and a decoder. The encoder processes the input image to obtain a representative feature vector, so-called latent code, by means of a series of consecutive blocks of convolution and down-sampling operations, whereas the decoder inverts these operations to obtain a result with the same size of the input from the latent code. After training, given an image to be processed, this model obtains a probabilistic map $\mathcal{P} = [0, 1]^{w \times h}$, with size $w \times h$, which represents the probability of each pixel belonging to a specified class—in our case, staff or background. This probability has to be converted to a decision for selecting the class to which each pixel belongs. This decision can be carried out by means of a global threshold $\rho \in [0, 1]$ to obtain a binary mask $\mathcal{M} = \{0, 1\}^{w \times h}$. In our experiments, we considered $\rho = 0.5$, similar to [7], in which the influence of this threshold was studied. After that, the bounding boxes of the staves can be retrieved by performing a connected-component analysis (CCA) over $\mathcal{M}$. However, it should be noted that for this task, it is not necessary to obtain a high-detail level of predictions, so that the input image can be resized to a smaller spatial resolution $(w \times h)$ before being processed by the neural network, and then apply a reverse resizing on $\mathcal{M}$ to recover the original resolution. This allows for less stringent resource requirements for training the model. The specific configuration considered for this work is described in Sect. 4.3. After this reverse resizing process, the CCA can obtain the bounding boxes of the staves with respect to the original size of the image.

The GRL is a special layer that inverts the gradients during the training process. It is connected to the inference model—the SAE model in this case—and a domain classifier. Its goal is to determine the association of given images to a particular domain—source or target—thus finding the useful features that differentiate them. However, since the domain classifier is connected to the GRL, the search of these features is inverted, with the idea of learning common features between the involved domains instead, i.e., domain-invariant features. The hypothesis is that these domain-invariant features that allow detection of staff regions for the source domain will also be useful to perform this task in the target domain. This technique makes use of adversarial training to adapt learning of the SAE model to indistinctly deal with images from $\mathcal{S}$ and $\mathcal{T}$. Nevertheless, the contribution of the domain classifier to the learning process can be tuned by means of a hyper-parameter $\lambda$, whose optimization depends on both the position in which the GRL is connected to the SAE and the architecture of the domain classifier. Section 4.3 details the configuration used for this approach.

## 4 Experimental setup

In this section, we describe the setup of our experiments, including the corpora, metrics and neural specifications considered. The code and instructions for use can be found in https://github.com/fjcastellanos/domain_adaptation_staff_retrieval.git.

[1] We used the MuRET tool [26] for annotating the staff regions as rectangular bounding boxes.



**(a)** Scheme for training with $\mathcal{S}$ and $\mathcal{T}$.

**(b)** Scheme for processing $\mathcal{T}$.

**Fig. 1** Scheme of the model proposed for performing unsupervised domain adaptation of staff-region retrieval. Figure 1a represents the scheme during the training process, whereas Fig. 1b indicates the manner of using the approach for evaluating the target images. Note that, only the ground truth of $\mathcal{S}$ is available for training the model

### 4.1 Corpora

For the experimentation, we considered several scanned manuscripts of different specifications. Note that, the significant differences between them will validate the application of the method in more realistic environments. Table 1 summarizes the details of each corpus considered.

- Capitan: collection of 96 images from a complete *Missa* of the second half of the 17th century [6] in mensural notation.
- Seils: dataset of 150 scanned typeset pages, written in mensural notation, of the '*Il Lauro Secco*' manuscript [23] belonging to an anthology of 16th-century Italian madrigals.
- Fmt: the '*Fondo de Música Tradicional IMF-CSIC*' corpus [27] consists of popular Spanish songs transcribed by musicologists between 1944 and 1960. As there are images with different graphic features, we split it into two different datasets. We selected a portion, specifically those 80 scanned pages under the name of '*c14*,' henceforth **Fmt-C**, and other portion which combines two sub-collections that depict similar features—'*M16*' and '*M38*'—with a total of 372 images, to which we will refer from now as **Fmt-M**.
- Patriarca: a dataset of 41 scanned pages preserved on '*Archivo Real Colegio Seminario de Corpus Christi*' under the code '*VAcp-Mus*,' and obtained with the respective ground-truth staff regions using MuRET [26].
- Guatemala: a collection of 384 images belonging to a repertoire of Guatemalan choirbooks, also extracted with the ground-truth data from MuRET.

We divided the corpora into three partitions for training, validation, and testing, with 60%, 20%, and 20% of the images, respectively. For comparison reasons, we considered the evaluation of a fixed portion of the images for each collection, used for the assessment of the supervised and the unsupervised approaches. The training set is used for training the model, whereas the validation portion will evaluate the model for each epoch to keep the model that optimizes the results.

Examples of each manuscript can be found in Fig. 2. Note that, all datasets are handwritten except Seils, while Fmt uses preprinted staff lines. Note also that the ground truth of several examples, shown in Fig. 2b, e and f, presents a

**Table 1** Details of the corpora considered for the experimentation

| | Pages | Avg. size (px.) | Staves |
|---|---|---|---|
| CAPITAN | 96 | 2 109 × 3 047 | 711 |
| SEILS | 150 | 813 × 1 200 | 1 141 |
| FMT- M | 372 | 699 × 939 | 1 352 |
| FMT- C | 80 | 4 036 × 3 161 | 257 |
| PATRIARCA | 41 | 3 496 × 4 120 | 379 |
| GUATEMALA | 384 | 2 000 × 1 335 | 3 300 |

certain degree of overlap. Since SAE is affected by this characteristic, we applied the same strategy proposed in [7]: For mitigating the ground truth overlaps, we reduced the height of each bounding box by a factor $\delta = 20\%$, applying 20% top and bottom trims to train the model, and then performed the reverse operation on the prediction stage.

In our experiments, the images were used in grayscale format through an OpenCV codec internal conversion, since in previous experiments, it was determined that the use of color did improve neither the supervised nor the unsupervised case.

## 4.2 Metrics

For evaluation, we included metrics used in object detection, as well as other additional metrics for extending the analysis and discussion of the results.

Intersection over Union (IoU) is often used to evaluate the quality of the retrieved bounding boxes. It measures the overlap between the ground-truth and the predicted regions by means of area comparison. With this metric, we will correlate this value with the number of bounding boxes properly retrieved, as well as the miss-detected ones. This relationship will give us a guide to determine if the unsupervised method is able to improve the quality of the predictions with respect to the supervised case.

For the sake of analysis, we include also an evaluation considering the precision P and the recall R metrics, which are mathematically defined as follows:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

where TP represents *True Positives* or correctly detected staff regions, FP stands for *False Positives* or those predictions that do not match a ground-truth bounding box, and FN the *False Negatives* that indicates the staff regions that have not been detected. For additional information, we also calculate their harmonic mean F-score ($\mathrm{F}_1$), which is defined as follows:

$$\mathrm{F}_1 = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P}+\mathrm{R}}.$$

Note that, the positive class, in this case, is represented by the staff areas. To complement these metrics, we will also observe the accuracy rates Acc with the aim of analyzing in more detail the obtained results. In our context, in which bounding boxes of staves are considered, this metric can be computed as follows:

$$\mathrm{Acc} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}.$$

## 4.3 Neural specifications

We considered a SAE architecture based on the one used previously for the supervised approach to staff-region recognition [9]. This type of architecture is a fully convolutional network (FCN) used also in other contexts [18,28].

The input image $\mathcal{I}$ is resized to 512 × 512 pixels and then normalized to facilitate the convergence of the training process. The normalization is performed with

$$\mathcal{I}_n = \frac{255 - \mathcal{I}}{255},$$

where $\mathcal{I}_n$ is the normalized image, whose pixels contain values between 0 and 1, used as input of the neural network. Note that, this normalization should be made for images from both $\mathcal{S}$ and $\mathcal{T}$.

Table 2 shows the details of the architecture considered. The model is trained for 300 epochs by means of Stochastic Gradient Descent [17] with a learning rate of 0.01. The validation set was used to determine the best model, which is later used for evaluation in the experiments carried out.

Concerning the domain adaptation technique based on GRL [13], we connected this special layer after the second last convolution layer (see Table 2). This layer was directly connected to a domain classifier that, given an image, has to predict whether the image belongs to $\mathcal{S}$ or $\mathcal{T}$. This classifier keeps the same architecture of the SAE model from the point in which the GRL is connected. In this way, similar to the configuration used in [8], the two outputs of the approach—one for the SAE part and the other one for the domain classifier—have the same size.

In addition, as described before, this technique includes a hyper-parameter $\lambda$ that controls the contribution of the domain classifier to the SAE weight updating. In our case, we selected $\lambda = 0.01$ with increments of 0.001 per epoch. Note also that the SAE part of the model was initially pre-trained for 50 epochs (of the 300 considered) using only $\mathcal{S}$, in order to start the domain adaptation step with more adequate weights for the task. The aforementioned considerations were decided after preliminary experiments.

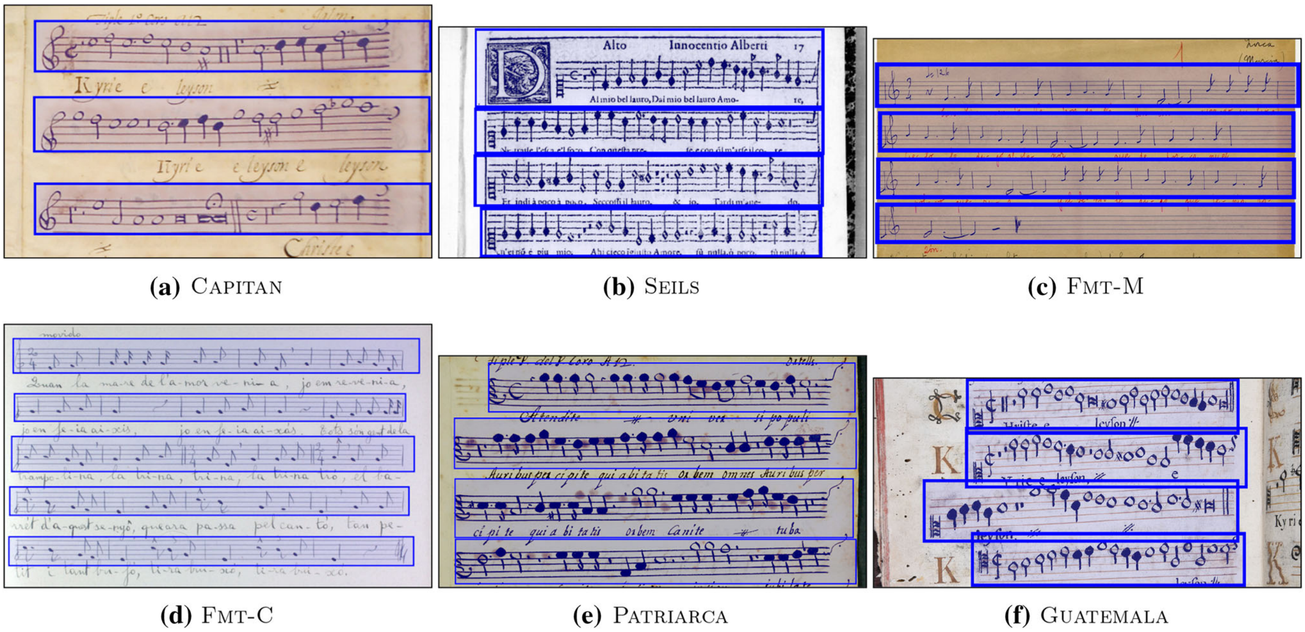

**(a)** Capitan **(b)** Seils **(c)** Fmt-M
**(d)** Fmt-C **(e)** Patriarca **(f)** Guatemala

**Fig. 2** Samples of the corpora considered for the experimentation with the respective ground truth

**Table 2** Detailed description of the selected SAE architecture, implemented as a FCN. 'ReLU' and 'sigmoid' denote the Rectifier Linear Unit and Sigmoid activations, respectively

| Input | Encoder | Decoder | Output |
|---|---|---|---|
| $[0, 1]^{512\times512}$ | Conv2D(32, 3 × 3, 'ReLU')<br>MaxPool(2 × 2) | Conv2D(32, 3 × 3, 'ReLU')<br>UpSamp(2 × 2) | $[0, 1]^{512\times512}$ |
| | Conv2D(32, 3 × 3, 'ReLU')<br>MaxPool(2 × 2) | Conv2D(32, 3 × 3, 'ReLU')<br>UpSamp(2 × 2) | |
| | Conv2D(32, 3 × 3, 'ReLU')<br>MaxPool(2 × 2) | Conv2D(32, 3 × 3, 'ReLU')<br>UpSamp(2 × 2) | |
| | | Conv2D(1, 3 × 3, 'sigmoid') | |

## 5 Results

In this section, we present the results obtained in the experimentation. To assess the benefits of domain adaptation for the staff recognition process, we compare our proposal—DANN—with the model without any adaptation—SAE. Since the objective is to process a manuscript different from the one used for training by means of unsupervised domain adaptation, the results, shown in Table 3, are organized according to the domains involved in each experiment. Note that, the metrics considered (described in Sect. 4.2) require determining when a predicted bounding box is correctly obtained. For this, a threshold is typically applied to the IoU value in order to determine when a prediction should be considered TP. According to previous work [9], this threshold, hereafter represented as $\alpha$, should be at least 0.7 to extract staff regions with sufficient quality to be eventually processed by an end-to-end approach and obtain the respective music sequence. In addition to this value, we also considered another typical threshold used in object detection problems, $\alpha = 0.5$, to analyze the results in more detail.

Table 3 shows the results obtained for all combinations of pairs of datasets, one as $\mathcal{S}$ and the other as $\mathcal{T}$, denoted in the rest of the work as $\mathcal{S}\rightarrow\mathcal{T}$. $Thirty$ experiments were performed for evaluating the benefits of the domain adaptation technique proposed, of which more than half, specifically 19, reported gain for both $\mathrm{F}_1$ values computed with different thresholds ($\alpha = 0.5$ and $\alpha = 0.7$). Also, there are 6 cases in which the improvements were not clear or the results were equal for both models, and only $five$ cases with a loss of performance for both $\mathrm{F}_1$ values.

Concerning the successful cases from the point of view of DANN, there are examples with a huge improvement in $\mathrm{F}_1$, such as the case of Seils→Capitan, with $\mathrm{F}_1{}^{\alpha=0.5}$ from 55.6 to 95.2% and $\mathrm{F}_1{}^{\alpha=0.7}$ from 40 to 78.6%, or the case of Guatemala→Fmt- M, in which the DANN model obtains 97.1% and 85.3%, against the results provided by the SAE model with 37.1% and 27.1%, for $\alpha = 0.5$ and

**Table 3** Results of our experimentation with all possible combinations of pairs of $\mathcal{S}$ and $\mathcal{T}$ with the corpora considered. Note that, as a reference, the table also includes the results of the SAE model in supervised experiments, i.e., when $\mathcal{S} = \mathcal{T}$, for the respective testing partitions. The last column summarizes each row according to the $F_1$ results: ✓ indicates that DANN is better than SAE for both values of $\alpha$ considered; = points out both those cases in which the same $F_1$ is obtained for the two considered thresholds and those where DANN improves the $F_1$ for one of the thresholds, but not for both; finally, ✗ stands for the scenarios in which DANN obtains worse results for both values of $\alpha$

| Scenario ($\mathcal{S}\rightarrow\mathcal{T}$) | | $F_1^{\alpha=0.5}$ (%) | | $F_1^{\alpha=0.7}$ (%) | | IoU (%) | | |
|---|---|---|---|---|---|---|---|---|
| | | SAE | DANN | SAE | DANN | SAE | DANN | |
| $\mathcal{S}$ = Capitan | | | | | | | | |
| $\mathcal{T}$ | Capitan | 95.1 | – | 93.5 | – | 75.9 | – | |
| | Seils | 24.6 | 33.9 | 18.9 | 20.9 | 26.0 | 32.5 | ✓ |
| | Fmt- C | 72.3 | 81.9 | 43.9 | 69.9 | 53.4 | 66.7 | ✓ |
| | Fmt- M | 94.6 | 96.3 | 77.4 | 80.3 | 72.5 | 74.5 | ✓ |
| | Patriarca | 89.0 | 90.7 | 75.9 | 72.7 | 69.1 | 70.2 | = |
| | Guatemala | 6.5 | 4.6 | 2.8 | 1.5 | 19.0 | 19.2 | ✗ |
| $\mathcal{S}$ = Seils | | | | | | | | |
| $\mathcal{T}$ | Seils | 98.1 | – | 92.9 | – | 80.0 | – | |
| | Capitan | 55.6 | 95.2 | 40.0 | 78.6 | 41.9 | 71.1 | ✓ |
| | Fmt- C | 0.0 | 0.0 | 0.0 | 0.0 | 1.8 | 13.3 | = |
| | Fmt- M | 0.5 | 91.9 | 0.0 | 82.8 | 6.9 | 75.2 | ✓ |
| | Patriarca | 83.0 | 86.0 | 56.1 | 60.2 | 62.0 | 64.1 | ✓ |
| | Guatemala | 42.5 | 34.1 | 19.3 | 17.5 | 37.6 | 30.8 | ✗ |
| $\mathcal{S}$ = Fmt- C | | | | | | | | |
| $\mathcal{T}$ | Fmt- C | 61.8 | – | 21.7 | – | 51.4 | – | |
| | Capitan | 2.2 | 62.1 | 0.0 | 27.6 | 13.1 | 42.8 | ✓ |
| | Seils | 4.5 | 17.2 | 0.0 | 3.9 | 13.9 | 25.0 | ✓ |
| | Fmt- M | 1.8 | 25.0 | 0.4 | 13.2 | 8.9 | 30.6 | ✓ |
| | Patriarca | 9.7 | 9.3 | 0.0 | 0.0 | 24.0 | 19.0 | = |
| | Guatemala | 23.0 | 16.5 | 4.8 | 1.6 | 31.7 | 25.9 | ✗ |
| $\mathcal{S}$ = Fmt- M | | | | | | | | |
| $\mathcal{T}$ | Fmt- M | 80.5 | – | 66.7 | – | 65.3 | – | |
| | Capitan | 61.5 | 76.1 | 4.4 | 16.8 | 46.1 | 52.0 | ✓ |
| | Seils | 7.0 | 30.0 | 0.0 | 3.3 | 26.5 | 35.7 | ✓ |
| | Fmt- C | 0.0 | 0.0 | 0.0 | 0.0 | 15.4 | 16.8 | = |
| | Patriarca | 12.4 | 15.0 | 0.0 | 4.2 | 27.1 | 13.3 | ✓ |
| | Guatemala | 0.0 | 2.2 | 0.0 | 0.0 | 2.6 | 17.6 | ✓ |
| $\mathcal{S}$ = Patriarca | | | | | | | | |
| $\mathcal{T}$ | Patriarca | 45.3 | – | 31.2 | – | 35.2 | – | |
| | Capitan | 40.8 | 33.8 | 10.0 | 1.5 | 32.1 | 35.6 | ✗ |
| | Seils | 0.7 | 26.2 | 0.0 | 7.8 | 4.7 | 30.0 | ✓ |
| | Fmt- C | 0.0 | 0.0 | 0.0 | 0.0 | 3.6 | 0.0 | = |
| | Fmt- M | 7.8 | 15.2 | 6.3 | 14.6 | 28.2 | 40.5 | ✓ |
| | Guatemala | 0.0 | 0.0 | 0.0 | 0.0 | 12.7 | 3.3 | = |
| $\mathcal{S}$ = Guatemala | | | | | | | | |
| $\mathcal{T}$ | Guatemala | 74.5 | – | 63.1 | – | 55.4 | – | |
| | Capitan | 16.9 | 68.9 | 1.2 | 64.2 | 21.8 | 60.3 | ✓ |
| | Seils | 76.1 | 68.7 | 42.3 | 29.3 | 53.6 | 47.3 | ✗ |
| | Fmt- C | 95.2 | 99.0 | 59.1 | 88.5 | 66.9 | 73.7 | ✓ |
| | Fmt- M | 37.1 | 97.1 | 27.1 | 85.3 | 48.2 | 75.9 | ✓ |
| | Patriarca | 64.7 | 93.1 | 45.1 | 68.5 | 43.0 | 66.0 | ✓ |

$\alpha = 0.7$, respectively. Also, an extreme scenario can be found in SEILS→FMT- M, in which SAE hardly estimates staff regions (0.5% and 0.0%), whereas DANN increases them up to 91.9% and 82.8% of $F_1$ for both values of $\alpha$. Note that, the architecture of SAE and DANN is the same, except that DANN includes the GRL with the domain classifier. Note also that the $\mathcal{T}$ domain is not annotated in the training process, so that, the SAE model is trained only with $\mathcal{S}$ and, therefore, it does not use images from the target. However, DANN uses $\mathcal{S}$ and $\mathcal{T}$ for training the model, being able to learn domain-invariant features from both domains involved even when $\mathcal{T}$ does not include ground-truth information. That is, with only one annotated domain, this approach is able to process other non-annotated ones without the need of annotating new data, and saving the time and efforts that would take to manually perform this laborious task.

With respect to the six cases in which there is no improvement, such as CAPITAN→PATRIARCA, we can see that a slight improvement is obtained in the quality of the staff retrieval by DANN, from 89% to 90.7% if we consider $\alpha = 0.5$, but shows a reduction in the performance from 75.9 to 72.7% when $\alpha = 0.7$. Note that, $\alpha$ is not a hyperparameter, but another metric to analyze the number of staves obtained with different quality in terms of IoU. Also, there are cases in which neither of the two models obtains staves with enough IoU to be considered as TP, such as SEILS→FMT- C or PATRIARCA→GUATEMALA. These are examples of failed staff extraction for both models that could be attributed to the fact of existing overlapping in the ground truth of several manuscripts. For example, SEILS→FMT- C does not obtain proper staff areas, but CAPITAN→FMT- C yields excellent results for the same target corpus, increasing ${F_1}^{\alpha=0.5}$ from 72.3 to 81.9% and ${F_1}^{\alpha=0.7}$ from 43.9 to 69.9%. This means that, depending on the source domain used, the target could experience improvements regardless of the specific manuscript. Another example is FMT- C→PATRIARCA, with results very low and similar regardless the model used, but when the source is changed to another one, such as the scenario GUATEMALA→PATRIARCA, being the same target, DANN improves the results from 64.7 to 93.1% for $\alpha = 0.5$ and 45.1% to 68.5% for $\alpha = 0.7$.

Moreover, there are $five$ cases in which the domain adaptation is clearly detrimental in the experiments. SEILS→GUATEMALA is an example, where the $F_1$ is decreased from 42.5 to 34.1% for $\alpha = 0.5$ and from 19.3 to 17.5% for $\alpha = 0.7$, or PATRIARCA→CAPITAN, with figures that goes from 40.8 to 33.8% and from 10 to 1.5% for both values of $\alpha$, respectively. However, as mentioned previously, the key to obtaining good results is in the selection of the domain used as $\mathcal{S}$. For example, although PATRIARCA→CAPITAN gets worst results for CAPITAN by the DANN approach, if the source used is SEILS or FMT- C (i.e., SEILS→CAPITAN and FMT- C→CAPITAN), the results are clearly improved by the domain adaptation strategy. This reinforces the idea that, even if a particular source manuscript does not provide good results for a target domain, another source may provide a better staff extraction for the target.

It should be also noted that there are cases in which the results of a supervised experiment outperform the unsupervised one tested on the same $\mathcal{T}$, even when considering SAE (i.e., without applying an adaptation mechanism). For example, focusing on FMT- C as $\mathcal{T}$, when SAE is supervisedly trained, it obtains ${F_1}^{\alpha=0.5} = 61.8\%$ and ${F_1}^{\alpha=0.7} = 21.7\%$, but in case that SAE is trained with GUATEMALA and evaluated fon FMT- C, we observe ${F_1}^{\alpha=0.5} = 95.2\%$ and ${F_1}^{\alpha=0.7} = 59.1\%$. Here, the supervised SAE has lower performance, which may be attributed to several factors: The high density of staves within a page, the high resolution of the images, and the low contrast between ink and background could be the main obstacles in this case. The combination of the first two factors could be affected by the size of the image given to the neural network, fixed to 512 × 512 pixels. This may be improved by increasing this size, although more computational resources will be needed. The third factor could be affected by the threshold $\rho = 0.5$, which is applied to determine in each pixel the presence of an area of interest, as described in Sect.3. For these numbers, we followed the considerations from the state of the art [7].

Similar results can be found in the case of PATRIARCA, since, the SAE model trained and evaluated with PATRIARCA obtains $F_1$ values of 45.3% and 31.2%, but, when we use CAPITAN, SEILS or GUATEMALA as $\mathcal{S}$, the results of SAE are considerably improved, even so without the need of domain adaptation mechanisms. In this case, there are two factors that may cause this phenomenon: the difference in the number of pages, since PATRIARCA is the corpus with fewer pages among considered corpora, and the presence of overlap in many bounding boxes. The lack of reference data combined with the fact of that SAE is not designed for overlapping may be key factors that can be detrimental to the staff retrieval task. Note that, in these cases, SAE also improves the unsupervised results. For example, in the GUATEMALA→PATRIARCA scenario, SAE increases to 64.7% for $\alpha = 0.5$ and 45.1% for $\alpha = 0.7$, signifying relative improvements of 43% and 44%, respectively. However, DANN provides ${F_1}^{\alpha=0.5} = 93.1\%$ and ${F_1}^{\alpha=0.7} = 68.5\%$, which are relative increases of 105% and 119%, respectively. This example shows that, although the SAE performs well, the DANN is able to improve it.

For a more generalizable analysis, Table 4 summarizes the average results obtained for all unsupervised cases ($\mathcal{S} \rightarrow \mathcal{T}$) for SAE and DANN. In order to simplify the table, we show the metrics computed with $\alpha = 0.7$, which, according to Castellanos et al. [9], is the minimum IoU value for obtaining reasonable results in the transcription process within OMR. On average, the results clearly show a notable improvement for all metrics considered. First, $F_1$ increases from 17.8 to

**Table 4** Comparison of average results between the model without adaptation (SAE) and the model with adaptation (DANN), considering only the combinations of domains (from Table 1) whose $\mathcal{S}$ and $\mathcal{T}$ are different ($\mathcal{S}\rightarrow\mathcal{T}$). As reference, the SAE model tested with the same domain used for training is also included as supervised learning reference ($\mathrm{SAE}_{\mathcal{S}\rightarrow\mathcal{S}}$). Bold cells highlight the best values in the cases with $\mathcal{S} \rightarrow \mathcal{T}$. We considered $F_1$, P, R, Acc and IoU metrics with $\alpha = 0.7$

| Model | $F_1$(%) | P (%) | R (%) | Acc (%) | IoU (%) |
|---|---|---|---|---|---|
| $\mathrm{SAE}_{(\mathcal{S}\rightarrow\mathcal{T})}$ | 17.8 | 13.5 | 29.1 | 13.4 | 30.5 |
| $\mathrm{DANN}_{(\mathcal{S}\rightarrow\mathcal{T})}$ | **30.5** | **25.4** | **41.9** | **25.4** | **41.0** |
| $\mathrm{SAE}_{(\mathcal{T}\rightarrow\mathcal{T})}$ (reference) | 61.5 | 52.3 | 81.2 | 52.0 | 60.5 |

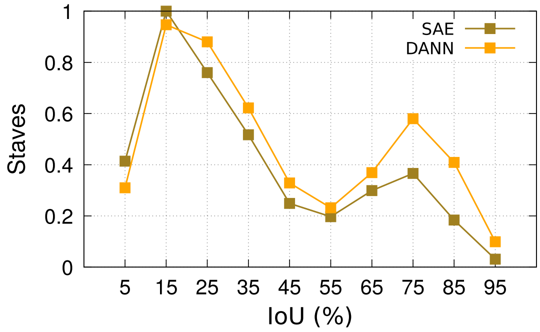


**Fig. 3** Normalized distribution of predicted staves according to the IoU for both the SAE and the DANN models

30.5%. Both P and R also experience great average benefits, from 13.5 to 25.4% and from 29.1 to 41.9%, respectively. The Acc metric also indicates that the number of retrieved staves is, in general, greater, increasing from 13.4 to 25.4%, whereas the IoU figures correlates also with the rest of metrics, increasing from 30.5 to 41.0%. Note that all metrics are improved and get closer to values obtained by the reference model, which is the supervised case in which the SAE model is trained and evaluated with $\mathcal{T}$.

The experiments reveal that, although there are several cases in which the domain adaptation does not improve the staff detection, DANN is, on average, definitely beneficial for the task. These results demonstrate that this technique is a potential solution for unsupervised scenarios.

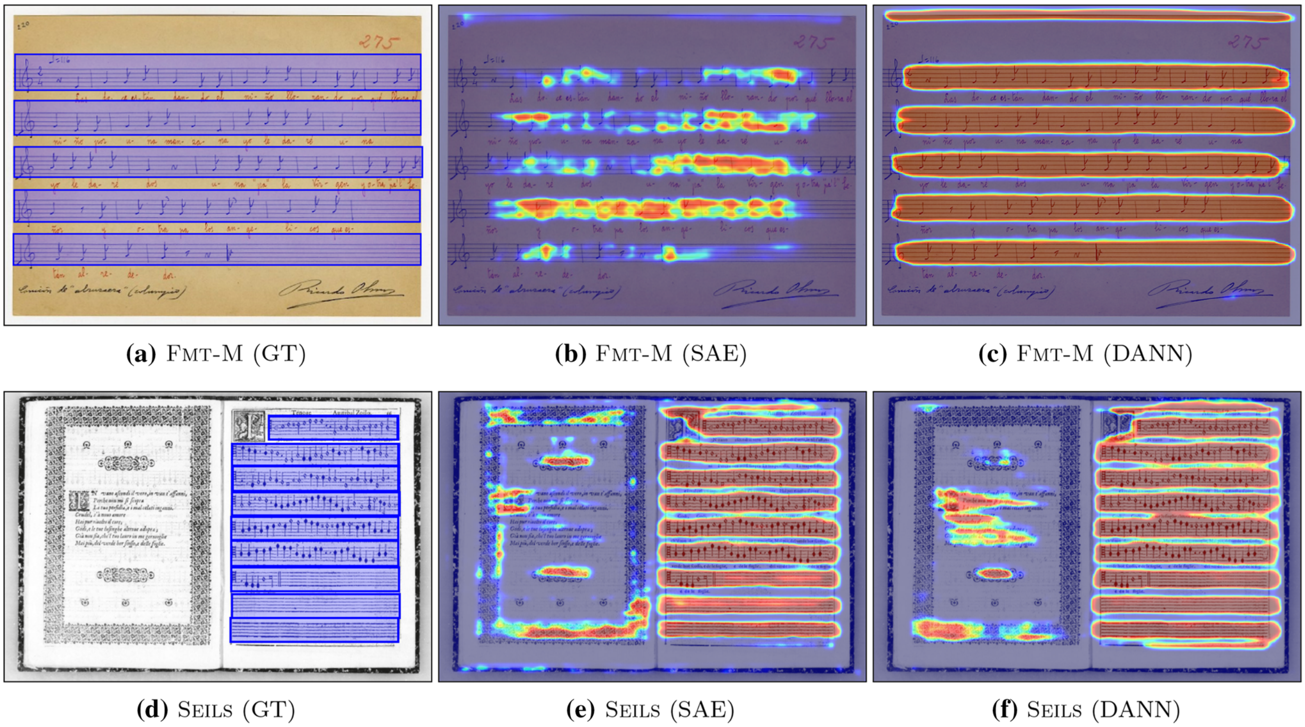
To analyze these results in more detail, Fig. 3 shows the normalized distribution of the predicted staves matched with the corresponding real staves according to the IoU. We can observe that both SAE and DANN have similar trends throughout the entire range of IoU. However, DANN slightly reduces the number of predicted staves with 15% or less of IoU with respect to the ground truth, while the number of predicted staves is increased when this metric is augmented. Indeed, we observe an important improvement in the number of predicted staves by DANN when IoU is higher than 65%, and it is particularly beneficial in the range between 75 and 85%, demonstrating, thus, that DANN can generally improve the prediction of staves. Note that, as aforementioned, 70% of IoU ($\alpha = 0.7$) is the minimum value stated in the literature to obtain reasonable quality in the eventual music transcription, and, therefore, the most interesting predictions in our context are those with IoU $\geq$ 70%. However, for IoU $\geq$ 60%, the graph also shows a significant increase.

As shown above, the domain adaptation technique based on GRL is able to improve the performance of the staff retrieval step in the OMR context. Most of the scenarios considered have shown the benefits of this strategy for the task. In general, the performance obtained by DANN in unsupervised experiments clearly surpasses the SAE model, making the adaptation a success. Although there are a few cases in which the adaptation is not suitable, they can be improved by using another manuscript as the source domain, making this strategy convenient for staff recognition.

To complement the analysis above, Fig. 4 shows two examples of staff recognition by means of the SAE model trained with source and evaluated in target (see Fig. 4b and e) and the results obtained using the domain adaptation approach DANN (see Fig. 4c and f).

The first example is GUATEMALA→FMT- M, in which the SAE model illustrates deficient estimation of staff zones (see Fig. 4b). Although the model detects the staves with a high degree of certainty, several parts of them have been spread apart and are inconsistent, these results are insufficient for correctly obtaining the bounding boxes. On the other hand, as it can be observed in Fig. 4c, DANN obtains a more stable result for all existing staves within the image and, as it can be seen in Fig. 4a, is closer to the expected results for the example, i.e., the ground truth. Note that, this qualitative example is consistent with the $F_1$ obtained in Table 3, in which SAE obtained 37.1% and 27.1%, according to the value of $\alpha$ analyzed (0.5 or 0.7, respectively), whereas DANN increases these figures to 91.1% and 85.3%.

Concerning the second example in Fig. 4, which corresponds to GUATEMALA→SEILS, at a first glance to Fig. 4e and f, it can be seen that the predictions on the staff areas are quite similar for SAE and DANN. On the left of the figures, there is an area without real staves where both models make mistakes, predicting staff areas where they do not exist. Despite this, DANN is able to reduce the amount of staff predictions in that area, improving thus the false positive errors. However, in the zone in which there are real staves, that is


**(a)** Fmt-M (GT) **(b)** Fmt-M (SAE) **(c)** Fmt-M (DANN)

**(d)** Seils (GT) **(e)** Seils (SAE) **(f)** Seils (DANN)

**Fig. 4** Examples of resulting probability maps obtained by the models SAE and DANN represented as heat maps compared with the respective ground-truth data (GT). The first row shows the results in Fmt- M for the adaptation scenario Guatemala→Fmt- M; whereas the second row is the case evaluated on Seils with the same source as the one used in the first example, i.e., Guatemala→Seils. Note that, both examples are unsupervisedly evaluated

on the right of the figures, although both detect staves in the correct areas, DANN worsens the results because of the overlapping between consecutive staves. This means that, when the CCA is performed, several staves will be predicted as a single staff, missing true positives in this process. Note, however, that this situation could be mitigated if image post-processing is performed to reduce the overlapping of the staff predictions.

## 6 Conclusions

In this work, an unsupervised domain adaptation technique is proposed for staff-region recognition of music score images. The main goal of the staff retrieval process is to obtain the bounding boxes of the staves, so that they can be processed individually to find the music sequence within them. Thus, the full-page transcription is possible by combining the sequences obtained for all staves within the image. Staff recognition has been addressed in the literature as a supervised problem, which assumes the availability of ground truth for a part of the image collection that has to be processed. However, these strategies require partial annotations, made usually by hand, with the associated cost in terms of human resources. With the countless number of manuscripts remaining to be transcribed, the cost of creating training sets is one of the main bottlenecks for OMR using supervised learning.

To reduce the cost, we propose the use of a domain adaptation technique, the so-called DANN, which is based on the use of GRL. This is a special layer that inverts the gradients, and, combined with a domain classifier, can adjust the neural weights to find domain-invariant features between two datasets: one annotated corpus, or source, and another non-annotated one, or target, for which staff regions have to be extracted. This technique is adapted to an existing staff-region retrieval model (SAE) that demonstrated good results for this object detection context in the supervised case.

It should be noted that, time-wise, the objective of our unsupervised approach is not to reduce the training time but to eliminate the need of having annotated data for each new corpus to process, which is one of the biggest obstacles to the practical application of OMR. Both in the supervised case and in our proposal, it is necessary to carry out a training process, although in the latter case, it is carried out without supervision. Therefore, our approach allows us to process images from a new target corpus without human intervention, performing a process that only uses the new images to be processed and not their labeling.

To evaluate the DANN-based approach, we considered six corpora of different provenances and characteristics, combined in pairs as the source and the target domains with a total of 30 unsupervised scenarios, with the aim of extracting more generalizable conclusions. Of the 30 cases, 19 demonstrated the benefits of DANN against the non-adaptation model (SAE), whereas six of the experiments yielded similar results for both models and five showed a decrease in performance. Although a target manuscript may show inefficient results by using a specific source domain, another manuscript used as the source may improve this situation without the need for annotating part of the target images. Even though there are a few cases that do not show a favorable adaptation between the involved domains, on average, the results show an improvement in the quality of the staff retrieval, with $F_1$ values from 19.8% obtained by the SAE model to 30.5% reached by DANN. Note that, as a reference, in supervised experiments, this metric only obtained 61.5% on average. Having this value as a possible upper bound, the use of the DANN approach resulted in a relative improvement of approximately 29%. The rest of the metrics considered—precision, recall, accuracy, and IoU—also supported the benefits of DANN for the task at hand.

In addition, it is demonstrated that the number of staves predicted on average for all unsupervised experiments is increased for those staff regions that have obtained high IoU, and particularly favorable for the cases with IoU greater than 60%, 70% being the minimum value stated in the literature to obtain reasonable results in the subsequent music sequence transcription.

To complement the analysis, qualitative results also show the benefits and limitations of DANN. Concerning the benefits, SAE often obtains deficient staff detection because it is trained with only the source domain and evaluated with a different domain. This situation is solved by our proposal, enabling the automatic staff retrieval of music score images from the target domain with more accurate and stable results. In some cases, the model without adaptation is able to detect the staves with high certainty and the use of adaptation only worsens the retrieval task. In these cases, DANN fails, likely because it modifies the weights of the neural network that were already working well to extract features from the target domain. In other words, the adaptation can improve the results when the non-adaptation model is unable to properly find the staves, but the other way around may be detrimental, i.e., in the case in which SAE trained with $\mathcal{S}$ is able to correctly process images from $\mathcal{T}$.

Although the experiments indicated that DANN is generally better than the non-adaptation model, there were cases where the source domain did not provide the correct features to deal with the target domain. For future work, we plan to address this issue, by exploring other techniques of domain adaptation such as generative adversarial network, or multi-source domain adaptation, which could be the key to obtaining more generalizable and usable models for unsupervised scenarios.

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

1. Bainbridge, D., Bell, T.: The challenge of optical music recognition. Comput. Humanit. **35**(2), 95–121 (2001)
2. Baró, A., Badal, C., Fornés, A.: Handwritten historical music recognition by sequence-to-sequence with attention mechanism. In: 17th international conference on frontiers in handwriting recognition, ICFHR, Dortmund, Germany, September 8–10, 2020, pp. 205–210. IEEE (2020)
3. Bosch, V., Calvo-Zaragoza, J., Toselli, A.H., Vidal-Ruiz, E.: Sheet music statistical layout analysis. In: 15th international conference on frontiers in handwriting recognition, ICFHR, Shenzhen, China, October 23–26, 2016, pp. 313–8 (2016)
4. Byrd, D., Simonsen, J.G.: Towards a standard testbed for optical music recognition: Definitions, metrics, and page images. J. New Music Res. **44**(3), 169–195 (2015)
5. Calvo-Zaragoza, J., Hajic Jr, J., Pacha, A.: Understanding optical music recognition. ACM Comput. Surv. **53**(4), 77:1-77:35 (2020)
6. Calvo-Zaragoza, J., Toselli, A.H., Vidal, E.: Handwritten music recognition for mensural notation: formulation, data and baseline results. In: Proceedings of the 14th IAPR international conference on document analysis and recognition, ICDAR, Kyoto, Japan, November 9–15, 2017, pp. 1081–1086 (2017)
7. Castellanos, F.J., Calvo-Zaragoza, J., Iñesta, J.M.: A neural approach for full-page optical music recognition of mensural documents. In: Proceedings of the 21th international society for music information retrieval conference, ISMIR, Montreal, Canada, October 11–16, 2020, pp. 558–565 (2020)
8. Castellanos, F.J., Gallego, A., Calvo-Zaragoza, J.: Unsupervised domain adaptation for document analysis of music score images. In: Proceedings of the 22nd international society for music information retrieval conference, ISMIR, Online, November 7–12, 2021, pp. 81–87 (2021)

9. Castellanos, F.J., Garrido-Munoz, C., Ríos-Vila, A., Calvo-Zaragoza, J.: Region-based layout analysis of music score images. arXiv preprint arXiv:2201.04214 (2022)
10. Dalitz, C., Droettboom, M., Pranzas, B., Fujinaga, I.: A comparative study of staff removal algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **30**(5), 753–766 (2008)
11. Fornés, A., Dutta, A., Gordo, A., Llados, J.: The icdar 2011 music scores competition: Staff removal and writer identification. In: 2011 international conference on document analysis and recognition, pp. 1511–1515. IEEE (2011)
12. Gallego, A.J., Calvo-Zaragoza, J.: Staff-line removal with selectional auto-encoders. Expert Syst. Appl. **89**, 138–148 (2017)
13. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 1–35 (2016)
14. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: European conference on computer vision, pp. 597–613. Springer (2016)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Adv. Neural. Inf. Process. Syst. **27**, 2672–2680 (2014)
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134 (2017)
17. Ketkar, N.: Stochastic gradient descent. In: Deep learning with Python, pp. 113–132. Springer (2017)
18. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
19. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 318–327 (2020)
20. Liu, A., Zhang, L., Mei, Y., Han, B., Cai, Z., Zhu, Z., Xiao, J.: Residual recurrent CRNN for end-to-end optical music recognition on monophonic scores. In: Proceedings of the 2021 workshop on multi-modal pre-training for multimedia understanding, Taipei, Taiwan, August 21, 2021, pp. 23–27. ACM (2021)
21. Pacha, A.: Incremental supervised staff detection. In: Proceedings of the 2nd international workshop on reading music systems, pp. 16–20. Delft, The Netherlands (2019)
22. Pacha, A., Eidenberger, H.: Towards self-learning optical music recognition. In: 16th international conference on machine learning and applications, pp. 795–800 (2017)
23. Parada-Cabaleiro, E., Batliner, A., Schuller, B.W.: A Diplomatic edition of Il Lauro secco: ground truth for OMR of white mensural notation. In: Proceedings of the 20th international society for music information retrieval conference, ISMIR, Delft, The Netherlands, November 4–8, 2019, pp. 557–564 (2019)
24. Quirós, L., Toselli, A.H., Vidal, E.: Multi-task layout analysis of handwritten musical scores. In: Iberian conference on pattern recognition and image analysis, pp. 123–134. Springer (2019)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural. Inf. Process. Syst. **28**, 91–99 (2015)
26. Rizo, D., Calvo-Zaragoza, J., Iñesta, J.M.: MuRET: A music recognition, encoding, and transcription tool. In: Proceedings of the 5th international conference on digital libraries for musicology, DLfM, Paris, France, September 28, 2018, pp. 52–56 (2018)
27. Ros-Fábregas, E.: Codified Spanish music heritage through Verovio: the online platforms Fondo de Música tradicional IMF-CSIC and books of hispanic polyphony IMF-CSIC . In: Proceedings of the 9th music encoding conference, MEC, Alicante, Spain, July 19–22, 2021 (2021)
28. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017)
29. Shen, J., Qu, Y., Zhang, W., Yu, Y.: Wasserstein distance guided representation learning for domain adaptation. In: Thirty-second AAAI conference on artificial intelligence (2018)
30. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: Proceedings of the thirtieth AAAI conference on artificial intelligence, AAAI'16, p. 2058-2065 (2016)
31. Tardón, L.J., Sammartino, S., Barbancho, I., Gómez, V., Oliver, A.: Optical music recognition for scores written in white mensural notation. EURASIP J. Image Video Process. **2009**(1), 843401 (2009)
32. Visaniy, M., Kieu, V., Fornés, A., Journet, N.: Icdar 2013 music scores competition: Staff removal. In: 2013 12th International conference on document analysis and recognition, pp. 1407–1411. IEEE (2013)
33. Waloschek, S., Hadjakos, A., Pacha, A.: Identification and cross-document alignment of measures in music score images. In: Proceedings of the 20th international society for music information retrieval conference, ISMIR, Delft, The Netherlands, November 4–8, 2019, pp. 137–143 (2019)
34. Waloschek, S., Hadjakos, A., Pacha, A.: Identification and cross-document alignment of measures in music score images. In: Proceedings of the 20th international society for music information retrieval conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019, pp. 137–143 (2019)
35. van der Wel, E., Ullrich, K.: Optical music recognition with convolutional sequence-to-sequence models. In: Proceedings of the 18th international society for music information retrieval conference, ISMIR, Suzhou, China, October 23–27, 2017, pp. 731–737 (2017)
36. Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W.: Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR, Honolulu, HI, USA, July 21–26, 2017, pp. 945–954 (2017)