

# Adversarial Feature Augmentation for Unsupervised Domain Adaptation

Riccardo Volpi<sup>1</sup>, Pietro Morerio<sup>1</sup>, Silvio Savarese<sup>2</sup>, Vittorio Murino<sup>1,3</sup>

{riccardo.volpi, pietro.morerio, vittorio.murino}@iit.it, ssilvio@stanford.edu

<sup>1</sup>Pattern Analysis & Computer Vision - Istituto Italiano di Tecnologia

<sup>2</sup>Stanford Vision and Learning Lab - Stanford University

<sup>3</sup>Computer Science Department - Università di Verona

## Abstract

Recent works showed that Generative Adversarial Networks (GANs) can be successfully applied in unsupervised domain adaptation, where, given a labeled source dataset and an unlabeled target dataset, the goal is to train powerful classifiers for the target samples. In particular, it was shown that a GAN objective function can be used to learn target features indistinguishable from the source ones. In this work, we extend this framework by (i) forcing the learned feature extractor to be domain-invariant, and (ii) training it through data augmentation in the feature space, namely performing feature augmentation. While data augmentation in the image space is a well established technique in deep learning, feature augmentation has not yet received the same level of attention. We accomplish it by means of a feature generator trained by playing the GAN minimax game against source features. Results show that both enforcing domain-invariance and performing feature augmentation lead to superior or comparable performance to state-of-the-art results in several unsupervised domain adaptation benchmarks.

images from one style to another [32, 18, 1, 17].

Furthermore, GANs have been exploited in the context of unsupervised domain adaptation. Here, a source (labeled) dataset and a target (unlabeled) dataset are considered, which are separated by the so-called domain shift [33], i.e., they are drawn from two different data distributions. Unsupervised domain adaptation aims at building models that are able to correctly classify target samples, despite the domain shift. In this framework, adversarial training has been used (i) to learn feature extractors that map target samples in a feature space indistinguishable from the one where source samples are mapped [8, 34], and (ii) to develop image-to-image translation algorithms [32, 18, 1, 17] aimed at converting source images in a style that resembles that of the target image domain.

In this paper, we build on the work by Tzeng et al. [34], which proposes to use a GAN objective to learn target features that are indistinguishable from the source ones, leading to a pair of feature extractors, one for the source and one for the target samples. We extend this approach in two directions: (a) we force domain-invariance in a *single* feature extractor trained through GANs, and (b) we perform data augmentation in the feature space (i.e., *feature augmentation*), by defining a more complex minimax game. More specifically, we perform feature augmentation by devising a feature generator trained with a Conditional GAN (CGAN [21]). The minimax game is here played with features instead of images, allowing to generate features conditioned to the desired classes. The CGAN generator is thus able to learn the class distribution in the feature space, and therefore to generate an arbitrary number of labeled feature vectors. Our results show that forcing domain-invariance and augmenting features are both valuable approaches in the unsupervised domain adaptation setting, leading to higher classification accuracies.

In summary, the *main contributions* of this paper are the following:

## 1. Introduction

Generative adversarial networks (GANs [10]) are models capable of mapping noise vectors into realistic samples from a data distribution. GANs are defined by two neural networks, a generator and a discriminator, and the training procedure is a minimax game where the generator is optimized to fool the discriminator, and the discriminator is optimized to correctly classify generated samples from actual training samples. Recently, this framework proved to be able to generate images with impressive accuracy [24], to generate videos from static frames [37], and to translate

1. Introducing for the first time the use of GANs to perform data augmentation in the feature space.
2. Proposing a new method for unsupervised domain adaptation, based on feature augmentation and (source/target) feature domain-invariance.
3. Evaluating the proposed method on unsupervised domain adaptation benchmarks (cross-dataset digit classification and cross-modal object classification), obtaining results which are superior or comparable to current state-of-the-art in most of the addressed tasks.

The remaining of the paper is organized as follows. Section 2 is dedicated to the related work. The models and the training procedure are presented in Section 3. In Section 4, the datasets used for the analysis and method’s validation are described. The experiments and associated results are detailed in Section 5. Finally, conclusive remarks are drawn in Section 6.

## 2. Related work

The work related to our proposed method is focused on GAN research and on modern domain adaptation techniques (*i.e.*, based on deep learning).

**Generative adversarial networks.** In the original formulation by Goodfellow et al. [10], a GAN model is trained through a minimax game between a generator, that maps noise vectors in the image space, and a discriminator, trained to discriminate generated images from real ones. Several other papers address ways to control what GANs generate [21, 3, 26]. In particular, CGANs [21] allow to condition on the desired classes, from which samples are generated. Other works [7, 6] propose to learn inference by playing a minimax game against features. In these works, trained models are feature extractors that map images into the feature space, not feature generators, which is *our primary goal*.

Performing feature augmentation through GANs is one of the original aspects of our approach. We propose a generator able to generate features from noise vectors and label codes, via a CGAN [21] framework, playing a minimax game with features extracted from a pre-trained model instead of images.

**Unsupervised domain adaptation.** Ganin and Lempitsky [8] propose a neural network (Domain-Adversarial Neural Network, DANN) where a ConvNet-based [16] feature extractor is optimized to both correctly classify source samples and have domain-invariant features, through adversarial training. Different works [35, 19] aim at minimizing the Maximum Mean Discrepancy [11] between features extracted from source and target samples, training

a classifier to correctly classify source samples while minimizing this measure. Bousmalis et al. [2] propose to learn image representations divided in two components, one shared across domains and one private, following the hypothesis that modeling unique elements in each domain can help to extract features which are domain-invariant. Tzeng et al. [34] use GANs to train an encoder for target samples, by making the features extracted with this model indistinguishable from the ones extracted through an encoder trained with source samples. The last layer of the latter can then be used for both encoders to infer labels. Saito et al. [28] propose an asymmetric tri-training where pseudo-labels are inferred and exploited for target samples during training. In particular, two networks are trained to assign labels to target samples and one to obtain target-discriminative features. Haeusser et al. [13] propose to exploit associations between source and target features during training, to maximize the domain-invariance of the learned features while minimizing the error on source samples.

Recently, several image-to-image translation methods have been proposed to solve unsupervised domain adaptation tasks. Taigman et al. [32] propose the Domain Transfer Network (DTN), that allows to translate images from a source domain to a target one, under a  $f$ -constancy constraint, where  $f$  is a generic function that maps images in a feature space. Translated images result portrayed in the target images’ style, while maintaining the content of the images fed in input. Liu and Tuzel [18] introduce Coupled GAN (CoGAN), an extension of GAN that allows to model a joint distribution  $P(X, Y)$  and to generate couples of images from noise vectors, one belonging to  $P(X)$  and one to  $P(Y)$ . This model can be applied to image-to-image translation tasks: fixing one image, the noise vector that most likely could have generated that picture can be inferred and, feeding it to the model, the second image is generated. Bousmalis et al. [1] propose to train an image-to-image translation network relying on both a GAN loss and a *task-specific* loss (and in problems with prior knowledge, also a *content-specific* loss). The resulting network takes in input both an image and a noise vector, that allows to generate a potentially infinite number of target images. Liu et al. [17] propose UNIT, an extension of CoGAN that relies on both GANs and Variational Auto-Encoders, and makes the assumption of a shared latent space. Image-to-image translation methods [32, 18, 1, 17] are applied to unsupervised domain adaptation by generating target images and training classifiers directly on them.

The domain-invariant feature extractor we designed is inspired by Tzeng et al. [34], with two main differences. First, we play the minimax game against features which are *generated* by a pre-trained model, thus performing *feature augmentation*. Second, we train the feature extractor

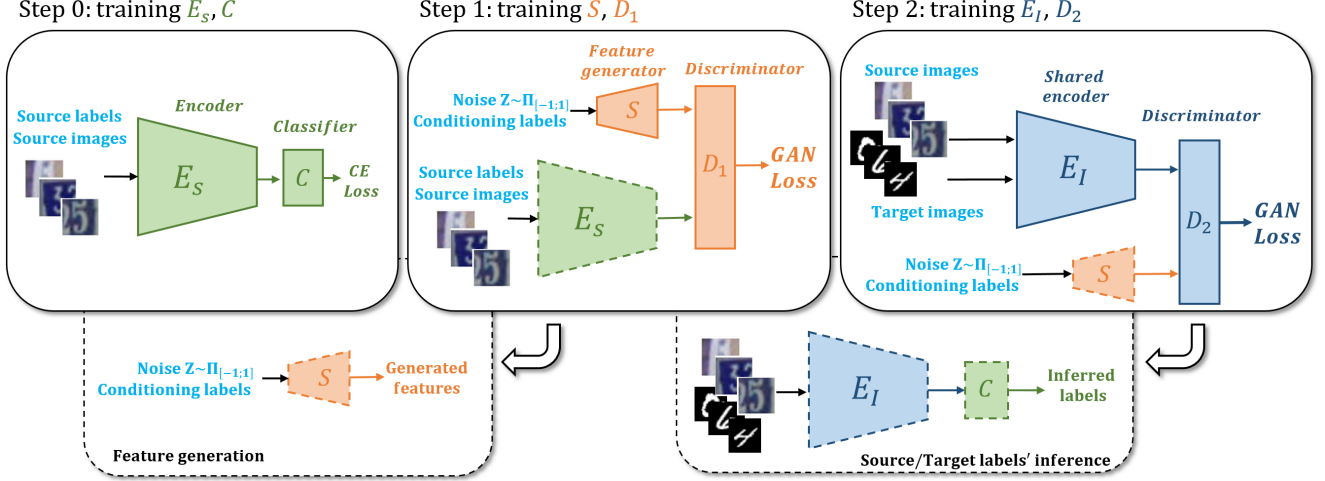


Figure 1. Training procedure, representing the steps described in Section 3.1. Solid lines indicate that the module is being trained, dashed lines indicate that the module is already trained (from previous steps). All modules are neural networks, whose architectures are detailed in Section 5.1. Smaller, dashed panels in the bottom indicate how to generate features (left) and how to infer source or target labels (right).

in order to make it work for both source and target samples (thus achieving *domain-invariance*), avoiding catastrophic forgetting. Both modifications lead to higher accuracies in classifying target samples, as we will show in Section 5. Domain-invariance also allows to use the same feature extractor for both source and target samples, while in Tzeng et al. [34] two different encoders are required.

### 3. Model

Our goal is to train a domain-invariant feature extractor ( $E_I$ ), whose training procedure is made more robust by data augmentation in the space of source features. The training procedure we designed to accomplish our intent is based on three different steps, depicted in Figure 1. First, we need to train a feature extractor on source data ( $C \circ E_s$ ). This step is necessary because we need a reference feature space and a reference classifier that performs well on it. Secondly, we need to train a feature generator ( $S$ ) to perform data augmentation in the source feature space. We can train it by playing a GAN minimax game against features extracted through  $E_s$ . Finally, we can train a domain-invariant feature extractor ( $E_I$ ) by playing a GAN minimax game against features generated through  $S$ . This module can then be combined with the softmax layer previously trained ( $C \circ E_I$ ) to perform inference on both source and target samples. All modules are neural networks trained by backpropagation [27]. In the following sections, we detail how each Step is performed, how new features can be generated, and how source/target labels can be inferred.

#### 3.1. Training

**Step 0.** The model  $C \circ E_s$  is trained to classify source samples.  $E_s$  represents a ConvNet feature extractor and  $C$  represents a fully connected softmax layer, with a size that depends on the problem. The optimization problem consists in the minimization of the following cross-entropy loss (*CE Loss* in Figure 1):

$$\min_{\theta_{E_s}, \theta_C} \ell_0 = \mathbb{E}_{(x_i, y_i) \sim (X_s, Y_s)} H(C \circ E_s(x_i), y_i), \quad (1)$$

where  $\theta_{E_s}$  and  $\theta_C$  indicate the parameters of  $E_s$  and  $C$ , respectively,  $X_s, Y_s$  are the distributions of source samples ( $x_i$ ) and source labels ( $y_i$ ), respectively, and  $H$  represents the softmax cross-entropy function.

**Step 1.** The model  $S$  is trained to generate feature samples that resemble the source features. Exploiting the CGAN framework, the following minimax game is defined:

$$\begin{aligned} \min_{\theta_S} \max_{\theta_{D_1}} \ell_1 = & \mathbb{E}_{(z, y_i) \sim (p_z(z), Y_s)} \|D_1(S(z||y_i)||y_i) - 1\|^2 \\ & + \mathbb{E}_{(x_i, y_i) \sim (X_s, Y_s)} \|D_1(E_s(x_i)||y_i)\|^2, \end{aligned} \quad (2)$$

where  $\theta_S$  and  $\theta_{D_1}$  indicate the parameters of  $S$  and  $D_1$ , respectively,  $p_z(z)$  is the distribution<sup>1</sup> from which noise samples are drawn, and  $\|$  denotes a concatenation operation. In this and the following steps, we relied on Least Squares GANs [20] since we observed more stability during training.

**Feature generation.** In order to generate an arbitrary

<sup>1</sup>Uniform in the range  $[-1, 1]$  throughout this work.

number of new feature samples, we only need  $S$ , which takes as input the concatenation of a noise vector and a *one-hot* label code, and outputs a feature vector from the desired class:

$$F(z|y) = S(z||y) \quad (3)$$

where  $z \sim p_z(z)$  and  $F$  is a feature vector belonging to the class label associated with  $y$  (dashed box in Figure 1, left).

**Step 2.** The domain-invariant encoder  $E_I$  is trained via the following minimax game, after being initialized with weights optimized on Step 0 (note that  $E_S$  and  $E_I$  have the same architecture), a requirement to reach optimal convergence:

$$\min_{\theta_{E_I}} \max_{\theta_{D_2}} \ell_2 = \mathbb{E}_{x_i \sim X_s \cup X_t} \|D_2(E_I(x_i)) - 1\|^2 \quad (4) \\ + \mathbb{E}_{(z, y_i) \sim (p_z(z), Y_s)} \|D_2(S(z||y_i))\|^2,$$

where  $\theta_{E_I}$  and  $\theta_{D_2}$  indicate the parameters of  $E_I$  and  $D_2$ , respectively. Since the model  $E_I$  is trained using both source and target domains, the feature extractor results domain-invariant. In particular, it maps both source and target samples in a common feature space, where features are indistinguishable from the ones generated through  $S$ . Being the latter trained to produce features indistinguishable from the source ones, the feature extractor  $E_I$  can be combined with the classification layer of Step 0 ( $C$ ) and used for inference (as in Tzeng et al. [34]):

$$\tilde{y}_i = C \circ E_I(x_i), \quad (5)$$

where  $x_i$  is a generic image from the source or the target data distribution and  $\tilde{y}_i$  is the inferred label (dashed box in Figure 1, right).

## 4. Datasets

To evaluate our approach, we used several benchmark splits of public source/target datasets adopted in domain adaptation.

**MNIST  $\leftrightarrow$  USPS.** Both datasets consist of white digits on a solid black background. We tested two different protocols: the first one (P1) consists in sampling 2,000 MNIST [16] images and 1,800 USPS [5] images. The second one (P2) consists in using the whole MNIST training set, 50,000 images, and dividing USPS in 6,562 images for training, 2,007 for testing, and 729 for validation. For P1, we tested the two directions of the split (MNIST  $\rightarrow$  USPS and MNIST  $\leftarrow$  USPS). For P2, we tested only MNIST  $\rightarrow$  USPS, and we avoided to use the validation set in this case, too. In both experimental protocols, we resized USPS digits to  $28 \times 28$  pixels, which is the MNIST images' size.

**SVHN  $\rightarrow$  MNIST.** SVHN [23] is built with real images of Street View House Numbers. We used the whole training sets of both datasets, following the standard protocol for unsupervised domain adaptation (SVHN training set contains 73,257 images), and tested on MNIST test set. We resized MNIST images to  $32 \times 32$  pixels and converted SVHN to grayscale. We did not use the extra set of SVHN.

**SYN DIGITS  $\rightarrow$  SVHN.** This split represents a synthetic-to-real domain adaptation problem, of great interest for research in computer vision since, quite often, generating labeled synthetic data requires less effort than obtaining large labeled dataset with real examples. SYN DIGITS [8] contains 500,000 images belonging to the same SVHN classes. We tested on SVHN test set.

**NYUD (RGB  $\rightarrow$  D).** This modality adaptation problem was proposed by Tzeng et al. [34]. The dataset is gathered by cropping out tight bounding boxes around instances of 19 object classes present in the NYUD [29] dataset. It comprises 2,186 labeled source (RGB) images and 2,401 unlabeled target (HHA-encoded [12]) depth images. Note that these are obtained from two different splits of the original dataset, to ensure that the same instance is not seen in both domains. The adaptation task is extremely challenging, due to the very different domains, the limited number of examples (especially for some classes), and the low resolution of the cropped bounding boxes.

## 5. Experiments

In this section, we evaluate our approach. First, we show that our model  $S$  is able to generate consistent and discriminant feature vectors conditioned on the desired classes. Second, we report an ablation study to figure out the benefits brought by the different steps that compose our approach. Finally, we compare our method with competing algorithms on unsupervised domain adaptation tasks.

### 5.1. Architectures

A detailed description of architectures and hyperparameters used (learning rate, batch sizes, *etc.*) is reported in the Supplementary Material. We provide here the details necessary for a basic understanding of the experiments.<sup>2</sup>

$S$  is built by the repetition of two blocks, each defined by a fully connected layer, a Batch Normalization layer [14], and a Dropout layer [31], followed by a fully connected layer with *tanh* activation functions.  $D_1$  is a one-hidden-layer neural network, with a *sigmoid* hidden unit as output layer. We defined  $E_S$  and  $E_I$  following standard architectures used in unsupervised domain adaptation [8]. In partic-

<sup>2</sup>Models were implemented using Tensorflow, and training procedures were performed on a NVIDIA Titan X GPU. Code: <https://github.com/ricvolpi/adversarial-feature-augmentation>



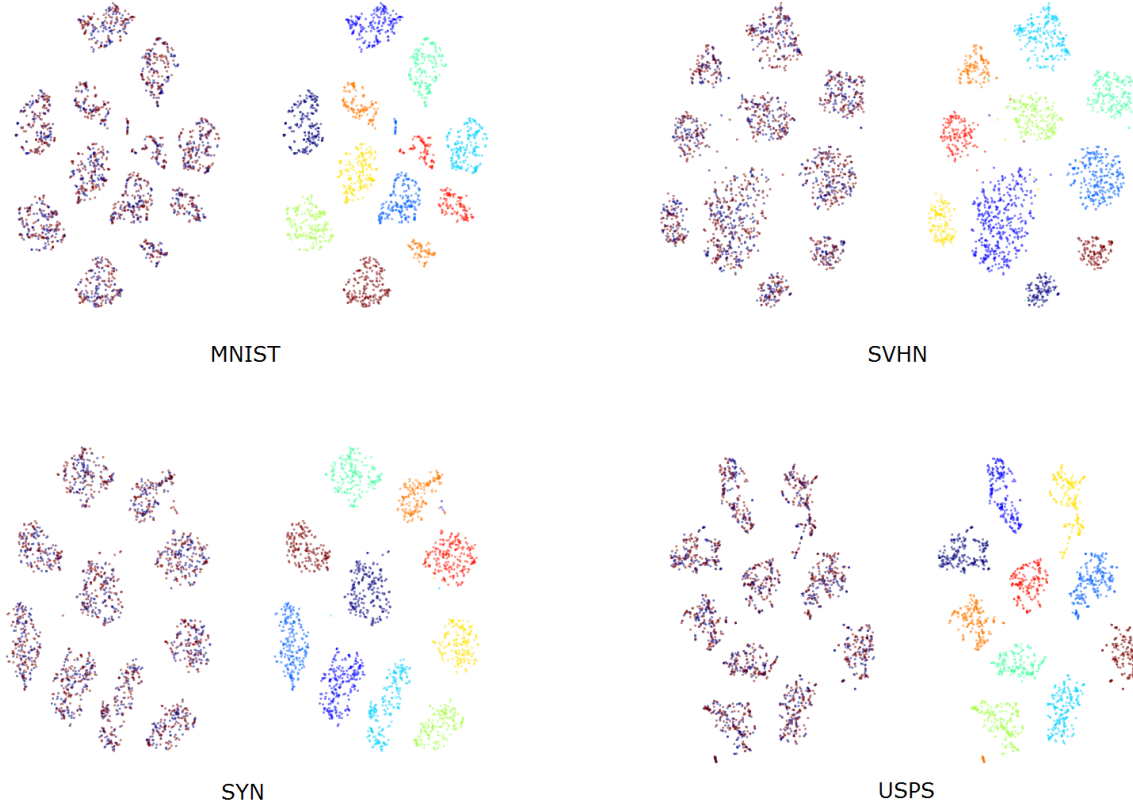


Figure 2. t-SNE plots of features associated with different adopted datasets (MNIST, SVHN, SYN, USPS). For each dataset, in the *left* part of the panels, red and blue dots indicate real and generated features, respectively. In the *right* part of the panels, different colors indicate different classes.

ular, for  $SVHN \rightarrow MNIST$ ,  $MNIST \rightarrow USPS$  and  $USPS \rightarrow MNIST$ , we defined the network as conv-pool-conv-pool-fc-fc-softmax (with Dropout [31] on fully connected layers for  $MNIST \leftrightarrow USPS$  experiments). For  $SVHN \rightarrow SYN$ , conv-pool-conv-pool-conv-fc-fc-softmax. For the NYUD experiment, in order to be comparable with [34], we used a VGG-16 [30] pretrained on ImageNet [4]. The final feature dimensionality (*e.g.*, the size of the feature vector fed to the softmax layer) was set to 128 for all experiments, except for  $SVHN \rightarrow SYN$  (256).  $D_2$  is built with two or three fully connected layers (depending on the experiment) with a *sigmoid* unit on top. Note that for the NYUD experiment we used three hidden layers, while Tzeng et al. [34] built the discriminator with two, since our method requires an additional one to reach convergence. For all our experiments, we used Adam optimizer [15] with momentum set to 0.95. *ReLU* [22] units were used throughout the architectures, except for last layers of discriminators, defined as *sigmoid* units, last layer of  $S$ , whose activation functions are *tanh*, and  $D_2$ , which was built with *Leaky ReLU* units, in agreement with the findings of Radford et al. [24].

## 5.2. Generating features

We qualitatively show with t-SNE [36] that we can generate feature vectors from the desired classes, after having trained  $S$  as described in Section 3.1. Figure 2 shows comparisons between real and generated features for different datasets. For each dataset, two identical point clouds are represented: the *bi-color* side (at the left of each panel), highlights real and generated samples (in red and blue, respectively); the *multi-color* side (at the right of each panel) highlights instead the different classes. From a

Table 1. *Second* column: number of activation patterns (APs) among the features extracted from training data. *Third* column: number of APs that  $S$  is able to generate. *Fourth* column: classification accuracy of the generated features, accordingly to given labels.

Dataset	#APs $E_S(x)$	#APs $S(z  y)$	Accuracy
SVHN	69,625	$\sim 10^6$	0.974
USPS	1,422		0.998
MNIST	1,910	$\sim 10^3$	0.995
NYUD	19		0.998

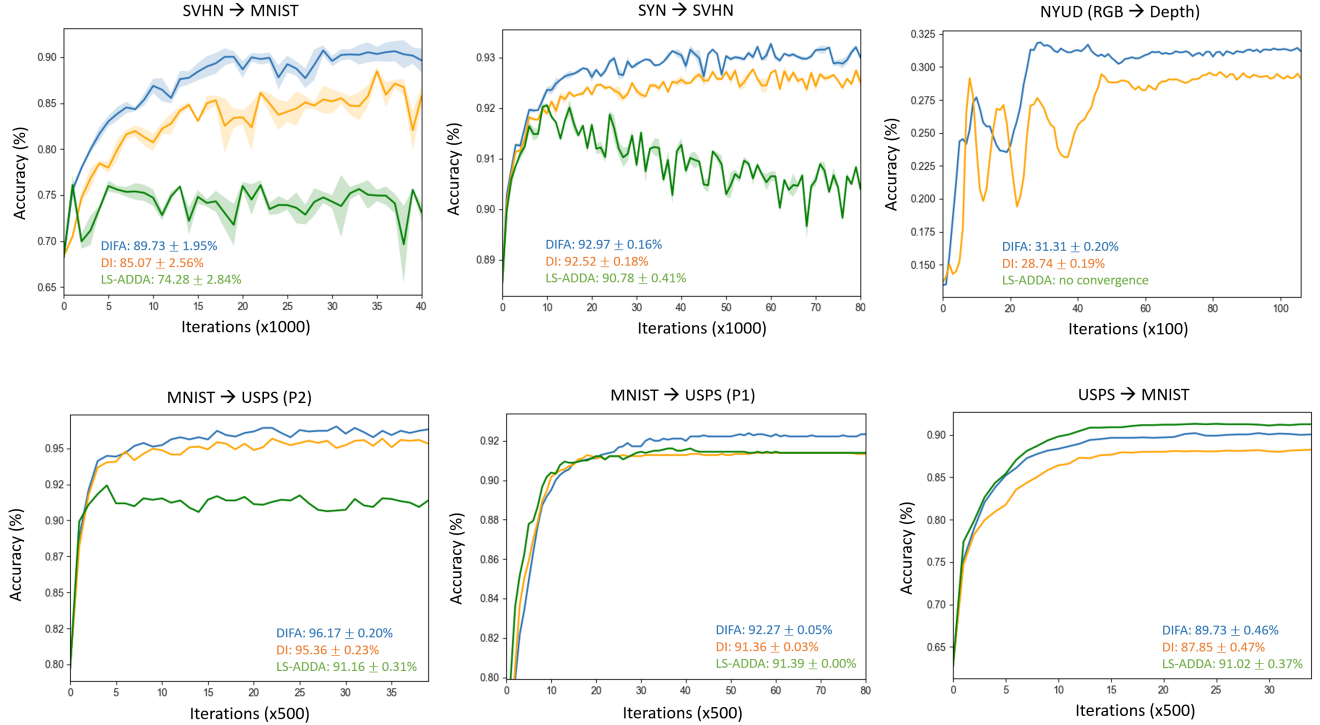


Figure 3. Accuracies on target samples evaluated throughout the training of the feature extractors of *LS-ADDA* (green), *DI* (orange) and *DIFA* (blue). Inference was performed by combining the feature extractor being learned with  $C$  of Step 0, Section 3.1. In the NYUD experiment the green curve is missing due to non-convergence of *LS-ADDA*. SVHN → MNIST and SYN → SVHN plots were obtained averaging over three different runs; confidence bands are portrayed.

qualitative point of view, real and generate features appear indistinguishable, and class structure is preserved. To quantitatively measure the quality of the features generated, we fed them to the classifier  $C$  trained with the original samples for class estimation. Table 1 (fourth column) shows that such features are also quantitatively reliable, and this is valid for all the datasets considered.

**Feature augmentation.** Finally, we are interested in evaluating the variability of the features generated through  $S$  to figure out whether (i) the model is memorizing the features from the training set, and (ii) it is realistic to assume that we are performing data augmentation in the feature space. To shed light on these two questions, we decided to perform the following empirical test: we counted the number of activation patterns (APs) that  $S$  is able to generate, and compared it with the ones intrinsically available in the original dataset. An activation pattern is defined by thresholding the output of the activation functions of the hidden state of a network. Raghu et al. [25] defined this concept for *ReLU*s [22], where values greater than zero are set to one, the others to zero. For our purposes, we can apply the same rule even if we are using

*tanh* activation functions. For example, SVHN has 73,257 samples that - with the feature extractor we used for our experiments - correspond to 69,625 activation patterns.  $S$  can instead generate a number of activation patterns in the order of  $10^6$  (counted empirically, feeding noise to  $S$  till saturation), indistinguishable from the original ones due to the training procedure defined in Section 3.1. Table 1 reports the results associated with the other datasets considered. Interestingly, activation patterns associated with the 2,186 source samples of NYUD are only 19: each pattern is associated with a different class. This is most likely due to overfitting: the network is already explicitly encoding classes at feature level. However, the generator  $S$  can enrich the feature set to a broad extent.

### 5.3. Ablation study

We carried out an ablation study to evaluate the benefit brought by the introduced modifications to the current way of using GAN objectives in unsupervised domain adaptation. Since the Least Squares GAN [20] framework is required to solve Step 1 and Step 2 of our method (Section 3.1), we re-designed the ADDA algorithm [34] in this framework as a baseline, and from this point we imple-

Table 2. Comparison of our method with competing algorithms. The row *LS-ADDA* lists results obtained by our implementation of Least Squares ADDA. The row *Ours (DI)* refers to our approach in which only domain-invariance is imposed. The row *Ours (DIFA)* refers to our full proposed method, which includes feature augmentation. (\*) DTN [32] and UNIT [17] use extra SVHN data (531, 131 images). (\*\*) Protocols P1 and P2 are mixed in the results section of Bousmalis et al. [1]. Convergence not reached is indicated as *no conv.*

	SVHN→MNIST	MNIST→USPS <sub>P1</sub>	MNIST→USPS <sub>P2</sub>	USPS→MNIST	SVHN→SYN	NYUD
Source	0.682	0.723	0.797	0.627	0.885	0.139
DANN [8, 9]	0.739	0.771 ± 0.018 [34]	-	0.730 ± 0.020 [34]	0.911	-
DDC [34]	0.681 ± 0.003	0.791 ± 0.005	-	0.665 ± 0.033	-	-
DSN [2]	0.827	-	-	-	0.912	-
ADDA [34]	0.760 ± 0.018	0.894 ± 0.002	-	<b>0.901 ± 0.008</b>	-	0.211
Tri [28]	0.862	-	-	-	<b>0.931</b>	-
DTN [32]	0.844*	-	-	-	-	-
PixelDA** [1]	-	-	0.959	-	-	-
UNIT [17]	<b>0.905*</b>	-	0.960	-	-	-
CoGANs [18]	no conv. [34]	0.912 ± 0.008	0.957 [17]	0.891 ± 0.008	-	-
<i>LS-ADDA</i>	0.743 ± 0.028	0.914 ± 0.000	0.912 ± 0.003	<b>0.910 ± 0.004</b>	0.908 ± 0.004	no conv.
<i>Ours (DI)</i>	0.851 ± 0.026	0.914 ± 0.000	0.954 ± 0.002	0.879 ± 0.005	0.925 ± 0.002	0.287 ± 0.002
<i>Ours (DIFA)</i>	<b>0.897 ± 0.020</b>	<b>0.923 ± 0.001</b>	0.962 ± 0.002	0.897 ± 0.005	<b>0.930 ± 0.002</b>	<b>0.313 ± 0.002</b>
Target	0.992	0.999	0.999	0.975	0.913	0.468 [34]

mented our peculiar contributions, showing that each one favourably concurs to improve performance. We term it *LS-ADDA*, and it is defined by the following minimax game:

$$\min_{\theta_{E_t}} \max_{\theta_D} \ell = \mathbb{E}_{x_i \sim X_t} \|D(E_t(x_i)) - 1\|^2 \quad (6)$$

$$+ \mathbb{E}_{x_i \sim X_s} \|D(E_s(x_i))\|^2,$$

where  $E_s$  is the feature extractor trained on source samples (as the one pre-trained in Step 0, Figure 1), and  $E_t$  is the encoder for the target samples that is being trained.  $D$  is the discriminator, as those described in this work.

The second analysis stage lies in imposing domain-invariance, and this is carried out by solving the following minimax problem:

$$\min_{\theta_{E_I}} \max_{\theta_D} \ell = \mathbb{E}_{x_i \sim X_s \cup X_t} \|D(E_I(x_i)) - 1\|^2 \quad (7)$$

$$+ \mathbb{E}_{x_i \sim X_s} \|D(E_s(x_i))\|^2,$$

where  $E_I$  is the *shared* encoder for the source and target samples that is being trained, and the rest of the modules are the same described above. This represents our first notable contribution, which we call *DI* (short for *DI LS-ADDA*, as this architecture introduces domain-invariance to *LS-ADDA*). Finally, the third analysis stage is constituted by our complete proposed approach, in which the minimax game also embeds the feature augmentation procedure (described in Step 2 of Section 3.1). We term it *DIFA* (*Domain-Invariance + Feature Augmentation*). For each of the three architectures proposed in this ablation study, we finally end up with an encoder that can be combined with the module  $C$  trained in Step 0 (see Figure 1). We tested these algorithms on the benchmark splits detailed in Section 4. Figure 3 shows the evolution of the performance of these three

Table 3. Difference in accuracy between training and test *source* data, by classifying with  $C \circ E_S$  and  $C \circ E_I$ . Source test data is not provided for NYUD [34].  $E_I$  does not experience catastrophic forgetting and generalizes well on unseen source data (test).

Dataset	$E_S \rightarrow E_I(\text{training})$	$E_S \rightarrow E_I(\text{test})$
USPS	0.975 → 0.973	0.980 → 0.979
MNIST <sub>(P1)</sub>	1.000 → 0.997	0.960 → 0.961
MNIST <sub>(P2)</sub>	0.997 → 0.986	0.992 → 0.984
SVHN	0.982 → 0.883	0.905 → 0.856
SYN	0.998 → 0.996	0.995 → 0.994
NYUD	1.000 → 1.000	<i>test set n.a.</i>

frameworks throughout the minimax games: *green* curves are associated with *LS-ADDA*, *orange* curves are associated with *DI*, and *blue* curves are associated with *DIFA*. The values reported in the bottom part of the plots indicate the average and the standard deviation calculated over the final stages of training, *i.e.*, when the minimax game reaches a stability point, despite oscillations. For the splits SVHN → MNIST and SYN → SVHN, we averaged over three different runs, due to some instability in the equilibriums reached, that can be observed in Figure 3. The general trend is that enforcing domain-invariance (*DI*) brings a first improvement (except in the MNIST → USPS (P1) experiment), and feature augmentation (*DIFA*) adds a further increment. In NYUD, *LS-ADDA* cannot converge.

The only exception is USPS → MNIST, where *LS-ADDA* is the best performing method. Note that we did not report experiments related to embedding feature augmentation without domain-invariance because it performs poorly, due to high instability.

## 5.4. Comparisons with other methods

Table 2 reports our findings and results obtained by the other works in the literature. The first row reports accuracies on target data achieved with non-adapted classifiers trained on source data, and the last row reports accuracies on target data achieved with classifiers trained on target data (*oracle*). Our main contributions lie in forcing the domain-invariance in the GAN minimax game (*DI*) and further improving it with feature augmentation (*DIFA*). A difficulty in unsupervised domain adaptation is determine the fair accuracy reached by each method, since cross-validation is not feasible (target labels should be used only to evaluate the method at the end of the training procedure). We believe that a fair way is the one we proposed in the previous section (*mean*  $\pm$  *std* calculated over the last iterations), since choosing a single value would be arbitrary and unfair in stochastic training procedures (*e.g.*, see SVHN  $\rightarrow$  MNIST and SYN  $\rightarrow$  SVHN in Figure 3).

Results show that our approach based on domain-invariance and feature augmentation leads to accuracies comparable or higher to current state-of-the-art in several unsupervised domain adaptation benchmarks. Among the splits we tested, the only exception is USPS  $\rightarrow$  MNIST, where ADDA [34] and our implementation of it (*LS-ADDA*) perform better - with the drawback of having two different feature extractors for source/target samples. In SVHN  $\rightarrow$  MNIST, our approach gives results comparable to current state-of-the-art (UNIT [17]), but it must be noted that the latter was achieved by making use of extra SVHN set (531, 131 images), making the result difficult to interpret. In MNIST  $\rightarrow$  USPS (P2) we perform better or comparably to any other method that was tested on it. Also note that all those methods [1, 18, 17] rely on the generation of target images to perform adaptation, and that [1, 17] rely on additional hyperparameters - a severe drawback in unsupervised domain adaptation, where cross-validation is not applicable. In SYN  $\rightarrow$  SVHN, our method is statistically comparable with the one proposed by Saito et al. [28]. In this case, it is also worth noting that the adapted feature extractor performs better than a neural network trained on SVHN (target) training set (see Table 2, last row). This opens a wide range of possibility of using synthetic data, which are much easier to obtain than labeled, real data in real-world applications. In NYUD (RGB  $\rightarrow$  Depth), we perform better than ADDA [34] by a large margin. In particular, embedding both domain-invariance and feature augmentation leads to an improvement  $> 10\%$ . We did not include the work by Haeusser et al. [13] in Table 2 because it makes use of a much more powerful feature extractor (conv-conv-pool-conv-conv-pool-conv-conv-pool-fc-softmax), which makes their method hard to compare with other works.

Finally, Table 3 shows the difference of performance on classifying source samples using  $C \circ E_s$  or  $C \circ E_I$ . As it

can be observed, the encoder  $E_I$  (trained following Step 2) works well on source samples, too. This allows to use the same encoder for both target and source data, a very useful feature in an application setting where we might not know the source of the data. The worst results on source samples, achieved on SVHN dataset, are most likely due to the large difference between the source and the target domains.

## 5.5. Limitations

The main limit of the domain-invariant feature extractor we designed is the same that can be detected in the works by Tzeng et al. [34] and by Ganin and Lempitsky [8]. Practically, all these approaches encourage source and target features to be indistinguishable, but this does not guarantee that target samples will be mapped in the correct regions of the feature space. In our case and in ADDA’s one, this strongly depends on the feature extractor trained on source samples: if the representation is far from being good, the results will be sub-optimal.

## 6. Conclusions and future work

In this work, we proposed two techniques to improve the current usage of GAN objectives in the unsupervised domain adaptation framework. First, we induced domain-invariance through a straightforward extension of the original algorithm. Second, we proposed to perform data augmentation in the feature space through GANs [10], a novel application. An exhaustive evaluation was carried out on standard domain adaptation benchmarks, and results confirmed that both approaches lead to higher accuracies on target data. Also, we showed that the obtained feature extractors can be used on source data, too.

Results showed that our approach is comparable or superior to current state-of-the-art methods, with the exception of a single benchmark. In particular, we performed better than recent, more complex methods that rely on generating target images to tackle unsupervised domain adaptation tasks. This achievement re-opens the debate on the necessity of generating images belonging to the target distribution: recent results [1, 17] seemed to suggest it.

For future work, we plan to test our approach on more complex unsupervised domain adaptation problems, as well as investigate if feature augmentation can be applied to different frameworks, *e.g.*, the contexts where traditional data augmentation proved to be successful.

## Acknowledgments

The authors would like to thank Lyne P. Tchapmi for helpful suggestions on an early draft. The research reported in this publication was supported by funding from MURI (1186514-1-TBCJE).



## References

- [1] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 7, 8
- [2] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 343–351. Curran Associates, Inc., 2016. 2, 7
- [3] X. Chen, X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180. Curran Associates, Inc., 2016. 2
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 5, 11
- [5] J. S. Denker, W. R. Gardner, H. P. Graf, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, H. S. Baird, and I. Guyon. Advances in neural information processing systems 1. chapter Neural Network Recognizer for Hand-written Zip Code Digits, pages 323–331. 1989. 4
- [6] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *CoRR*, abs/1605.09782, 2016. 2
- [7] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. *CoRR*, abs/1606.00704, 2016. 2
- [8] Y. Ganin and V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1180–1189, 2015. 1, 2, 4, 7, 8, 11
- [9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1), Jan. 2016. 7
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 1, 2, 8, 11
- [11] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, Mar. 2012. 2
- [12] S. Gupta, R. Girshick, P. A. Aez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014. 4
- [13] P. Haeusser, T. Frerix, A. Mordvintsev, and D. Cremers. Associative domain adaptation. In *International Conference on Computer Vision (ICCV)*, 2017. 2, 8
- [14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. 4
- [15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5
- [16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998. 2, 4
- [17] M. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *CoRR*, abs/1703.00848, 2017. 1, 2, 7, 8
- [18] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 469–477. Curran Associates, Inc., 2016. 1, 2, 7, 8
- [19] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37, ICML’15*, pages 97–105, 2015. 2
- [20] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, and Z. Wang. Multi-class generative adversarial networks with the L2 loss function. *CoRR*, abs/1611.04076, 2016. 3, 6
- [21] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 1, 2
- [22] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Frnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814. Omnipress, 2010. 5, 6
- [23] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 4
- [24] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 1, 5
- [25] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. On the expressive power of deep neural networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2847–2854, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 6
- [26] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 217–225. Curran Associates, Inc., 2016. 2
- [27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Learning Internal Representations by Error Propagation, pages 318–362. MIT Press, Cambridge, MA, USA, 1986. 3
- [28] K. Saito, Y. Ushiku, and T. Harada. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the*

34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, pages 2988–2997, 2017. 2, 7, 8

- [29] N. Silberman, D. Hoiem, P. Kohli, , and R. Fergus. Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision (ECCV)*, 2012. 4
- [30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 5, 11
- [31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1), Jan. 2014. 4, 5, 12
- [32] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 1, 2, 7
- [33] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '11*, pages 1521–1528, Washington, DC, USA, 2011. IEEE Computer Society. 1
- [34] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 3, 4, 5, 6, 7, 8, 11
- [35] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014. 2
- [36] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 5
- [37] C. Vondrick, H. Pirsaviash, and A. Torralba. Generating videos with scene dynamics. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 613–621. Curran Associates, Inc., 2016. 1

## A. Architectures

We provide in this section a detailed description of the networks used for our experiments. For the digit datasets, the encoders follow the standard architectures commonly used in unsupervised domain adaptation [8].

Figure 4, *left*: architectures of  $E_S$  and  $E_I$  used for  $\text{MNIST} \leftrightarrow \text{USPS}$  and  $\text{SVHN} \rightarrow \text{MNIST}$ .

Figure 4, *right*: architectures of  $E_S$  and  $E_I$  used for  $\text{SYN} \rightarrow \text{SVHN}$ .

Figure 5, *left*: architecture of  $S$  used for all the experiments.

Figure 5, *right*: architecture of  $D_1$  used for all the experiments.

Figure 6, *left*: architecture of  $D_2$  used for  $\text{SVHN} \rightarrow \text{MNIST}$  and  $\text{SYN} \rightarrow \text{SVHN}$ .

Figure 6, *right*: architecture of  $D_2$  used for  $\text{MNIST} \leftrightarrow \text{USPS}$  and  $\text{NYUD} (\text{RGB} \rightarrow \text{D})$ .

Concerning  $E_S$  and  $E_I$  used in the NYUD experiment, we relied on a pretrained VGG-16 [30], following the protocol used by Tzeng et al. [34]. We cut it at *fc7*, which was shrunk to be 128-dim and modified with tanh activations. The classifier  $C$  consists in an additional 19-dimensional softmax layer.

We found out that  $D_2$  should be built with two or three hidden layers to stabilize the minimax game against  $E_I$  (whose structure must be the same as  $E_S$ ). We designed an  $S$  that proved to be reliable in all experiments; to play a balanced minimax game, we found out that a one-hidden-layer neural network as a discriminator ( $D_1$ ) is an optimal choice. The size of the hidden layer depends on the problem, and can be determined by observing the stability of the training procedure.

## B. Hyperparameters

We report in this section the hyperparameters used in the different Steps of the training procedures. Note that hyperparameters were set in order to reach the convergence of the GAN [10] minimax games, no cross-validation using target labels was performed.

### B.1. Digits

For each training Step, we used a batch size of 64 samples. The learning rate was set to  $3 \cdot 10^{-4}$  for Step 0,  $1 \cdot 10^{-4}$  for Step 1 and  $3 \cdot 10^{-5}$  for Step 2, in all experiments except  $\text{MNIST} \leftrightarrow \text{USPS}$ , where was set to  $3 \cdot 10^{-6}$ .

### B.2. NYUD

In Step 0, the network is not trained from scratches: following the protocol described in [34], we fully fine-tune a VGG-16 network [30] (pre-trained on ImageNet [4]) for 20.000 iterations, in order to have a comparable baseline model. Batch size is 32 (instead of 128) due to hardware limitations. The learning rate were  $10^{-4}$  for Step 0,  $10^{-5}$  for Step 1 and  $10^{-7}$  for Step 2.

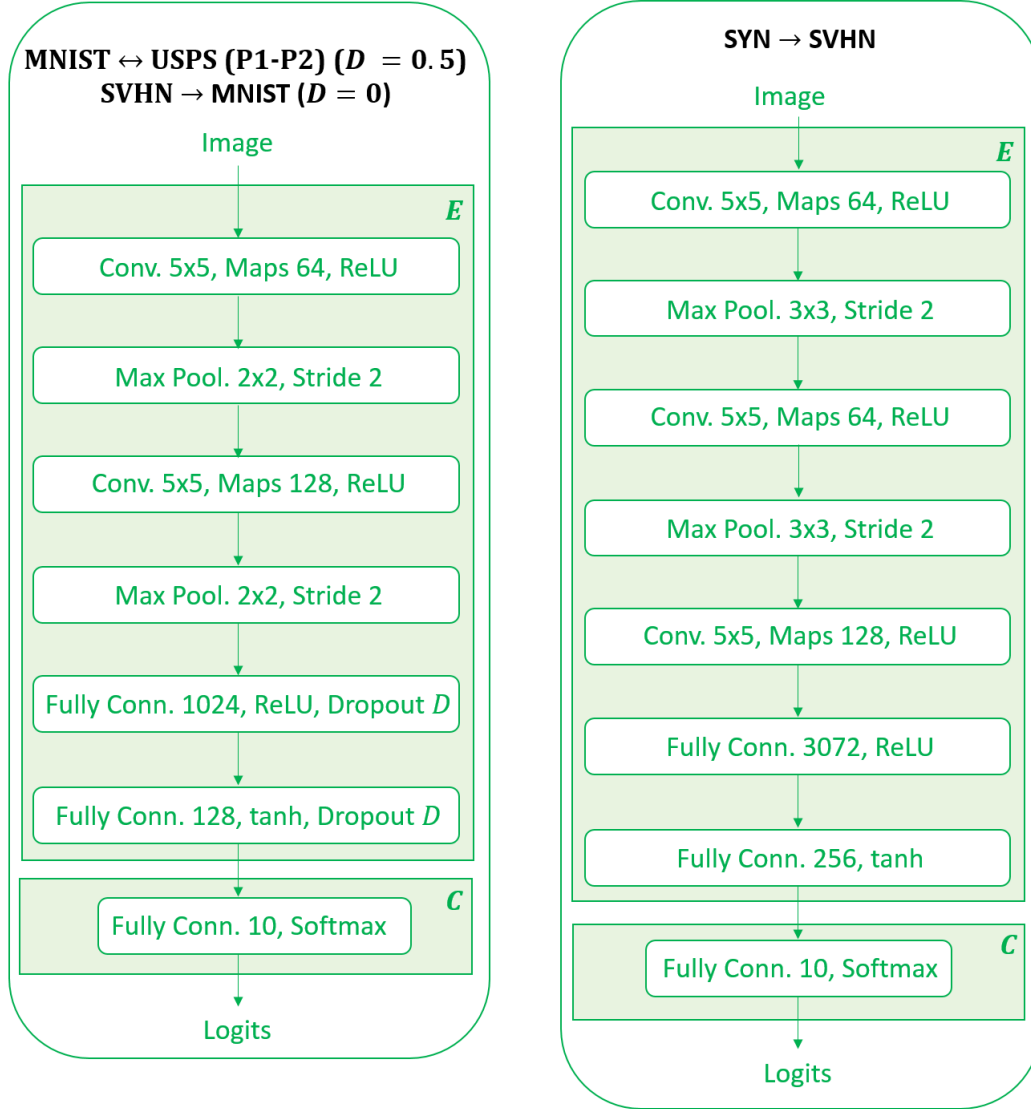


Figure 4. Architectures used for  $C \circ E_S$  and  $C \circ E_I$  ( $C \circ E$  for simplicity) in the MNIST ↔ USPS (P1-P2) and in the SVHN → MNIST (left) experiments, with the different values of Dropout [31] indicated ( $D$ ), and in the SYN → SVHN experiment (right). The classification module ( $C$ ) is a simple fully-connected + softmax layer.



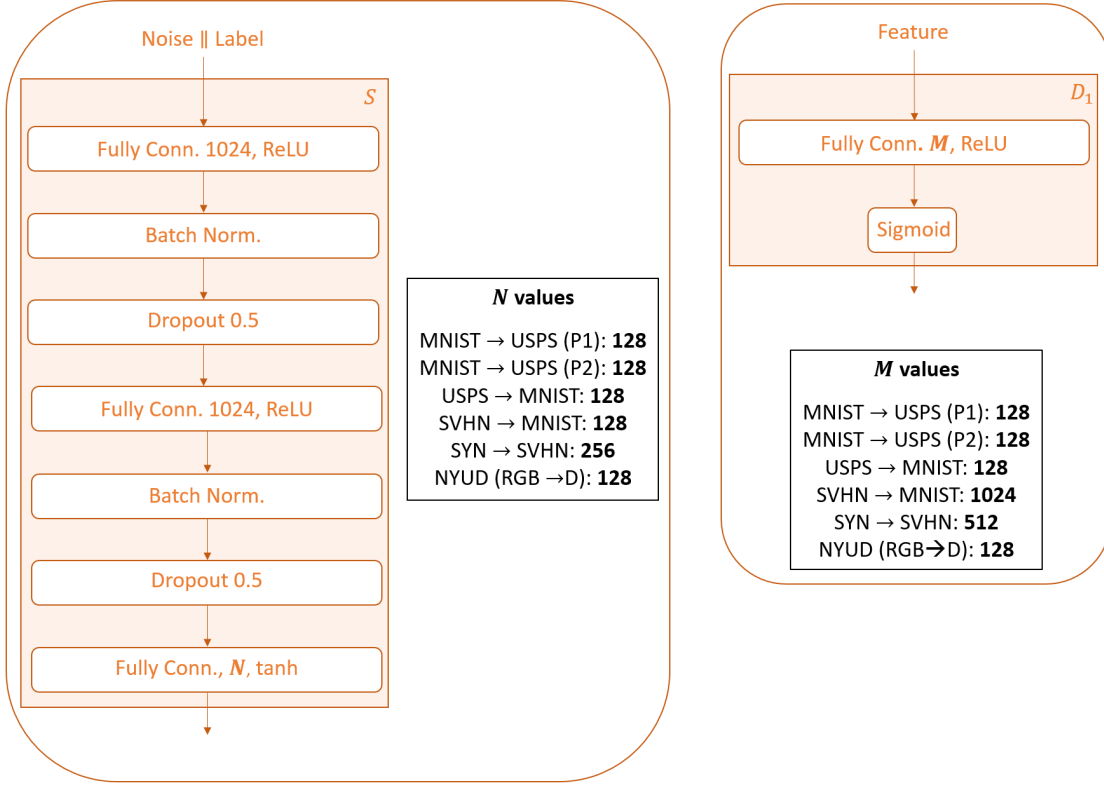


Figure 5. Architectures used for  $S$  (left) and for  $D_1$  (right), with the size of the features generated and of the hidden layer indicated, respectively.

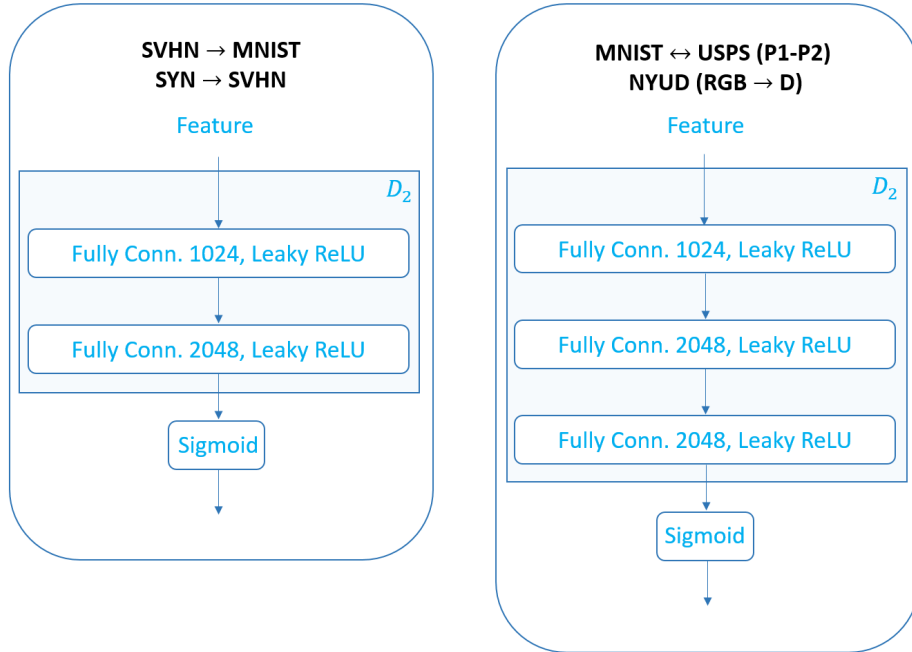


Figure 6. Architectures used for  $D_2$  in the NYUD and MNIST ↔ USPS experiments (right) and in all the others (left).