Signer-Independent Sign Language Recognition with Adversarial Neural Networks

Pedro M. Ferreira, Diogo Pernes, Ana Rebelo, and Jaime S. Cardoso

***Abstract*—Sign Language Recognition (SLR) has become an appealing topic in modern societies because such technology can ideally be used to bridge the gap between deaf and hearing people. Although important steps have been made towards the development of real-world SLR systems, signer-independent SLR is still one of the bottleneck problems of this research field. In this regard, we propose a deep neural network along with an adversarial training objective, specifically designed to address the signer-independent problem. Specifically, the proposed model consists of an *encoder*, mapping from input images to latent representations, and two classifiers operating on these underlying representations: (i) the *sign-classifier*, for predicting the class/sign labels, and (ii) the *signer-classifier*, for predicting their signer identities. During the learning stage, the *encoder* is simultaneously trained to help the *sign-classifier* as much as possible while trying to fool the *signer-classifier*. This adversarial training procedure allows learning signer-invariant latent representations that are in fact highly discriminative for sign recognition. Experimental results demonstrate the effectiveness of the proposed model and its capability of dealing with the large inter-signer variations.**

***Index Terms*—Sign language recognition, gesture recognition, adversarial neural networks, deep learning.**

I. Introduction

Sign languages are the naturally occurring linguistic systems that arise within a Deaf community and, currently, considered the standard education method of deaf people worldwide. Sign language communication is expressed by means of articulated hand gestures (i.e., manual signs) along with facial expressions to convey meaning. Contrary to the popular belief, sign language is not universal and, just like spoken languages, it has its own lexicon, syntax and grammar. This is why most of hearing people are unfamiliar with sign language, which obviously creates a serious communication barrier between deaf communities and the hearing majority.

As a key technology to help bridging the gap between deaf and hearing people, Sign Language Recognition (SLR) has become one of the most active research topics in the human-computer interaction field. Its main purpose is to automatically translate the signs, from video or images, into the corresponding text or speech. Although recent SLR methods have demonstrated remarkable performances in signer-dependent scenarios, i.e. when training and test data come from the same signers, their recognition rates typically decrease significantly when the signer is new to the system. This performance drop is the result of the large inter-signer variability in the manual signing process of sign languages (see Fig. 1). However, a practical SLR system must operate in a signer-independent scenario, which means that the signer of the probe must not be seen during the training routine of the models. Therefore, signer-independent SLR has become one of the bottleneck problems for the development of a real-world and practical SLR system.

Fig. 1. Inter-signer variability: it is possible to observe not only phonological variations (i.e., different handshapes, palm orientations, and sign locations) but also a large physical variability (i.e., different hand sizes) when six signers are performing the same sign.

Borrowing from recent works on adversarial neural networks [1], [2] and domain transfer [3], we introduce a deep neural network along with a novel adversarial training objective to specifically tackle the signer-independent SLR problem. The underlying idea is to preserve as much information as possible about the signs, while discarding the signer-specific information that is implicitly present in the manual signing process. For this purpose, the proposed deep model is composed by an *encoder* network, which maps from the input images to latent representations, as well as two discriminative classifiers operating on top of these underlying representations, namely the *sign-classifier* network and the *signer-classifier* network. While the *sign-classifier* is trained to predict the sign labels, the *signer-classifier* is trained to discriminate their signer identities. In addition, the parameters of the *encoder* network are optimized to minimize the loss of the *sign-classifier* while trying to fool the *signer-classifier* network. This adversarial

Manuscript received December 1, 2019; revised August 20, 2020. This work was partially financed by the ERDF – European Regional Development Fund through the Operational Program for Competitiveness and Internationalization - COMPETE 2020 Program and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project "POCI-01-0145-FEDER-030707"', and in part by Ph.D. under grants SFRH/BD/102177/2014 and SFRH/BD/129600/2017.

The authors are with Centre for Telecommunications and Multimedia, INESC TEC, 4200-465 Porto, Portugal, and also with Faculdade de Engenharia da Universidade do Porto, 4200-465 Porto, Portugal, and also with Faculdade de Ciências da Universidade do Porto, 4169-007 Porto, Portugal, and also with Universidade Portucalense, 4200-072 Porto, Portugal (Corresponding author: Pedro M. Ferreira, e-mail: pmmf@inesctec.pt).

doi: 10.18178/ijmlc.2021.11.2.1024

and competitive training scheme encourages the learned representations to be signer-invariant and highly discriminative for the sign classification task. To further constrain the latent representations to be signer-invariant, we introduce an additional training objective that operates on the hidden representations of the *encoder* network in order to enforce the latent distributions of different signers to be as similar as possible.

Although this adversarial training framework is similar to those initially introduced by Ganin *et al* [3], in the context of domain adaptation, and then by Feutry *et al* [2] to learn anonymized representations, our main contributions on top of these works are two-fold:

1) The application of the adversarial training concept to the signer-independent SLR problem;
2) A novel adversarial training objective that differs from the ones of Ganin *et al* [3] and Feutry *et al* [2] in two ways. First, our training objective is minimum if and only if the adversarial classifier, which in our case corresponds to the *signer-classifier*, produces a uniform distribution over the signer identities, meaning that our model is completely invariant to the signer identity of the training data. Second, we introduce an additional term to the adversarial training objective that further discourages the learned representations of retaining any signer-specific information, by explicitly imposing similarity in the latent distributions of different signers.

This paper is an extension of our conference paper [4]. The new contributions of this paper are summarized as follows:

1) The introduction of a transfer learning strategy in the proposed adversarial training objective, yielding an overall improvement in the sign recognition performance. Concretely speaking, instead of training all the network components from scratch, the *encoder* network is initialized with the first 10 layers of VGG-19 [5], pre-trained on the ImageNet [6], and then finetuned to our specific task.
2) An extended experimental section to further demonstrate the effectiveness of the proposed model. Specifically, the experimental evaluation of the proposed model is extended to an additional SLR database. Moreover, we introduce a quantitative analysis of the produced latent representations and an analysis of the training behavior of the proposed model.

The remainder of the paper is organized as follows. Section II presents the related work. The proposed model along with its adversarial training scheme are fully described in Section III. Experimental results and conclusions are reported in Sections IV and V, respectively.

II. Related Work

According to the amount of data required from the test signers, previous signer-independent SLR works can be roughly classified into two main groups, namely (i) signer adaptation approaches, where a previously trained model is adapted to a new test signer by using a small amount of signer specific data, and (ii) truly signer-independent methods, in which a generic model robust for new test signers is built without using data of those test signers.

Greatly inspired by speaker adaptation methods from the speech recognition research, Von Agris *et al* [7] proposed the combination of the eigenvoice (EV) approach [8] with maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) estimation to adapt trained Hidden Markov Models (HMMs) to new signers. More recently, Kim *et al* [9] investigated the potential of different deep neural network adaptation strategies for the signer-independence problem. Yin *et al* [10] proposed an interesting weakly-supervised signer adaptation approach, in which the adaptation data from the new signer has not to be labeled. Specifically, a generic metric is first learnt from the available labeled data of several different signers and, then, adapted to the new signer by considering clustering and manifold constraints along with the collected unlabeled data. Although signer adaptation is a reasonable approach, in practice, collecting enough training data from each new signer to retrain and adapt the model may not be feasible. In this regard, several works focused on the development of truly-signer independent models that do not require any data from the new signers [11]-[17]. Most of them involved a huge feature engineering effort in order to build normalized hand-crafted feature descriptors robust to the large inter-signer variations. A major weakness across all the aforementioned methods is related to the fact that representation and metric learning is not jointly performed. Motivated by the inherent difficulty of designing reliable handcrafted features to the large inter-signer variability, recent SLR systems are mostly based on deep neural networks [18]-[22]. It is well-known that deep neural networks are remarkably good in figuring out reliable high-level feature representations from the data. However, in previous deep SLR methodologies, the learned representations are not explicitly constrained to be signer-invariant. Therefore, there is nothing to prevent the learned representations of different signers and the same class of being far apart in the representation space and, hence, signer invariance is not ensured.

This paper presents a novel adversarial training objective, based on representation learning and deep neural networks, specifically designed to address the signer-independent SLR problem. Different from the aforementioned methodologies, our model jointly learns the representation and the classifier from the data, while explicitly imposing signer invariance in the high-level representations for a robust and truly signer-invariant sign recognition.

III. Proposed Method

The ultimate goal of our model is to learn signer-invariant latent representations that preserve the relevant part of the information about the signs while discarding the signer-specific traits that may hamper the sign classification task. To accomplish this purpose, we introduce a deep neural network along with an adversarial training scheme that is able to learn feature representations that combine both sign discriminativeness and signer-invariance.

More specifically, let $\mathbb{X} = \{X_i, y_i, s_i\}_{i=1}^{N}$ denote a labeled dataset of N samples, where X_i represents the i-th colour image, and y_i and s_i denote the corresponding class (sign) label and signer identity, respectively. To induce the model

to learn signer-invariant representations, the proposed model comprises three distinct sub-networks:

1) An *encoder* network, which aims at learning an encoding function $h(\mathrm{X};\ \theta_h)$, parameterized by θ_h, that maps from an input image X to a latent representation h;
2) A *sign-classifier* network, which operates on top of this underlying latent representation h to learn our task-specific function $f(h;\ \theta_f)$, parameterized by θ_f, that maps from h to the predicted probabilities $p(\mathrm{y}|h;\ \theta_f)$ of each sign class.
3) A *signer-classifier* network, with the purpose of learning a signer-specific function $g(h;\ \theta_g)$, parameterized by θ_g, that maps the same hidden representation h to the predicted probabilities $p(\mathrm{s}|h;\ \theta_g)$ of each signer identity.

During the learning stage, the parameters of both classifiers are optimized in order to minimize their errors on their specific tasks on the training set. In addition, the parameters of the *encoder* network are optimized in order to minimize the loss of the *sign-classifier* network while forcing the *signer-classifier* to be a random guessing predictor. In the course of this adversarial training procedure, the learned latent representations h are encouraged to be signer-invariant and highly discriminative for sign classification. To further discourage the latent representations of retaining any signer-specific traits, we introduce an additional training objective that enforces the latent distributions of different signers to be as similar as possible. The result is a truly signer-independent model robust to new test signers.

A. Architecture

As illustrated in Fig. 2, the architecture of the proposed model is composed by three main sub-networks or blocks, i.e. an *encoder*, a *sign-classifier* and a *signer-classifier*.

The *encoder* network attempts to learn a mapping from an input image **X** to a latent representation h. It simply consists of a sequence of L_e pairs of consecutive 3×3 convolutional layers with Rectified Linear Units (ReLUs) as non-linearities. For downsampling, the last convolutional layer of each pair has a stride of 2. On top of that, there is a fully-connected layer, also with a ReLU, representing the desired signer-invariant latent representations h.

Taking the latent representations h as input, the *sign-classifier* block is composed by a sequence of L_s fully-connected layers, with ReLUs as the non-linear functions, for predicting the sign class $\hat{y} = \arg\max f(h;\ \theta_f)$. Therefore, the last fully-connected layer has a softmax activation function which outputs the probabilities for each sign class.

The *signer-classifier* network has exactly the same topology as the *sign-classifier* net. However, it maps the latent representations h to the predicted signer identity $\hat{s} = \arg\max g(h;\ \theta_g)$. Therefore, the number of nodes of the output layer is defined according to the number of signers in the training set.

B. Adversarial Training

By definition, signer-invariant representations discard all signer-specific information and, as such, no function (i.e., classifier) exists that maps such representations into the correct signer identity. This naturally leads to an adversarial problem, in which: (i) a *signer-classifier* network $g(\cdot;\ \theta_g)$ receives latent representations $h = h(X;\ \theta_h)$ from an *encoder* network $h(\cdot;\ \theta_h)$ and tries to predict the signer identity s corresponding to image X and (ii) the *encoder* network tries to fool the *signer-classifier* network while still providing good representations for the *sign-classifier* network $f(\cdot;\ \theta_f)$, which in turn receives the same representations h and aims to predict the sign label y corresponding to image X.

Therefore, the *signer-classifier* network shall be trained to minimize the negative log-likelihood of correct signer predictions:

$$\min_{\theta_g} \mathcal{L}_{\text{signer}}(\theta_h, \theta_g) = -\frac{1}{N}\sum_{i=1}^{N} \log p(s_i|h(X_i;\theta_h);\ \theta_g) \quad (1)$$

In the perspective of the *encoder*, the predictions of the *sign-classifier* should be as accurate as possible and the predictions of the *signer-classifier* should be kept close to uniform, meaning that this latter classifier is not capable of doing better than random guessing the signer identity. Formally, this may be translated into the following constrained objective:

$$\min_{\theta_h,\theta_f} \mathcal{L}_{\text{sign}}(\theta_h, \theta_f) = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i|h(X;\ \theta_h);\ \theta_f) \quad (2)$$

$$\text{subject to} \frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}(\mathcal{U}_{\mathbb{S}}(\mathrm{s}) \,||\, p(\mathrm{s}|h(X_i;\ \theta_h);\ \theta_g) \leq \epsilon, \quad (3)$$

where $\mathrm{D_{KL}}$ is the Kullback-Leibler (KL) divergence and $\mathcal{U}_{\mathbb{S}}(\mathrm{s})$ denotes the discrete uniform distribution on the random variable s, defined over the set of identities S in the training set. Here, $\epsilon \geq 0$ determines how far from uniform the *signer-classifier* predictions are allowed to be (as measured by the KL divergence). The choice of the uniform distribution implies the underlying assumption that the training set is balanced relatively to the number of examples per signer (which should be true for most practical datasets). When this is not the case, the empirical distribution of signer identities in the training set may be used instead.

The constraint inequality (3) may be rewritten as:

$$\mathcal{L}_{\text{adv}}(\theta_h, \theta_g) = \frac{1}{N|\mathbb{S}|}\sum_{i=1}^{N}\sum_{\mathrm{s}\in\mathbb{S}} \log p(\mathrm{s}|h(X_i;\ \theta_g);\ \theta_g) \leq \epsilon + \\ + \log|\mathbb{S}| \quad (4)$$

and the constrained optimization problem may be equivalently formulated as:

$$\min_{\theta_h,\theta_f} \mathcal{L}(\theta_h, \theta_f, \theta_g) = \mathcal{L}_{\text{sign}}(\theta_h, \theta_f) + \lambda\mathcal{L}_{\text{adv}}(\theta_h, \theta_g), \quad (5)$$

where $\lambda \geq 0$ depends on ϵ and $\mathcal{L}_{adv}$ plays the role of an adversarial loss with respect to the signer classification loss L_{signer}.

This objective and the structure of our model are similar to those used in [3], in the context of domain adaptation, and in [2], to learn anonymized representations for privacy purposes. However, the former uses the negative signer classification loss as the adversarial term (i.e., $\mathcal{L}_{adv} \leftarrow -\mathcal{L}_{signer}$), which is not lower bounded, leading to high gradients and difficult optimization. The latter addresses this

problem by replacing this term with the absolute difference between the adversarial loss as defined in equation (4) and the signer classification loss (i.e., $\mathcal{L}_{adv} \leftarrow |\mathcal{L}_{adv} - \mathcal{L}_{signer}|$. This option has a nice information theoretic interpretation as being an empirical upper bound for the mutual information between the distribution of signer identities and the distribution of latent representations. Nonetheless, this loss vanishes for infinitely many (non-uniform) distributions. Our choice, besides being clearly lower bounded by the entropy of the uniform distribution, $\log|\mathbb{S}|$, is minimum if and only if $p(s|h(X_i;\theta_h);\theta_g) \equiv \mathcal{U}_{\mathbb{S}}(s), \forall i$, meaning that the *signer-classifier* block is completely agnostic relatively to the signer identities of the training samples.

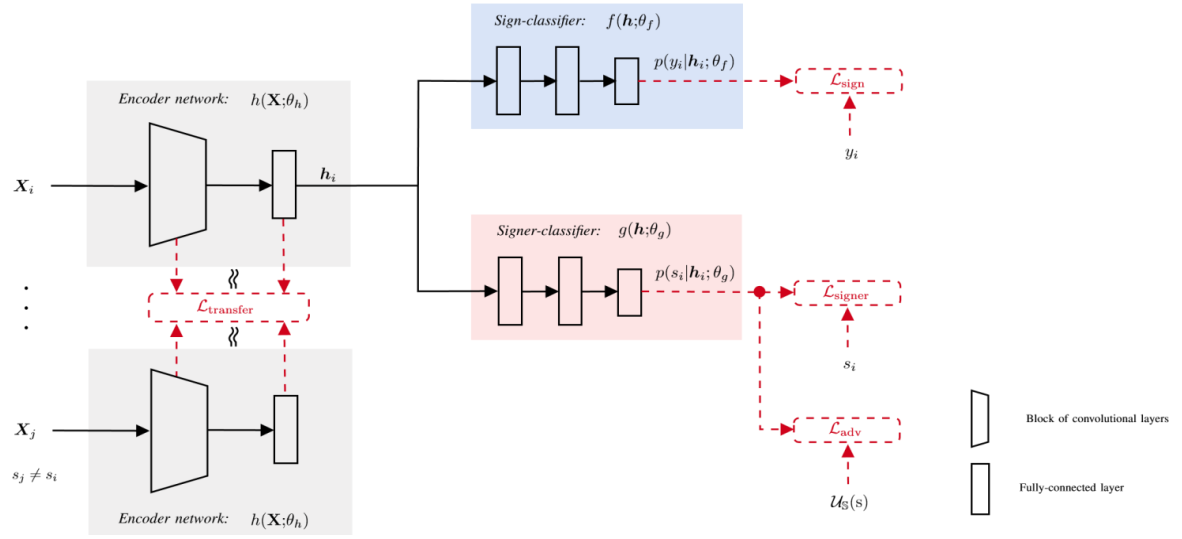


Fig. 2. Architecture of the proposed signer-invariant neural network. It comprises three main sub-networks or blocks: an *encoder*, a *sign-classifier* and a *signer-classifier*.

C. Signer-Transfer Training Objective

To further encourage the latent representations h to be signer-invariant, we introduce an additional term in objective (5), the so-called signer-transfer loss $\mathcal{L}_{\text{transfer}}$. The core idea of $\mathcal{L}_{\text{transfer}}$ is to enforce the latent distributions of different signers to be as similar as possible. In practice, this is achieved by minimizing the difference between the hidden representations
of different signers, at each layer of the *encoder* network. To measure the signers' distribution difference at the m-th layer, $m \in \{1, 2, \ldots, M\}$, we compute a distance $D^{(m)}$ between the hidden representations $h^{(m)}(\cdot;\theta_h)$ of two signers s and t at the output of that layer, as:

$$\mathcal{D}^{(m)}(s,t;\theta_h) = \left\| \frac{1}{N_s}\sum_{i:s_i=s} h^{(m)}(X_i;\theta_h) - \frac{1}{N_t}\sum_{j:s_j=t} h^{(m)}(X_j;\theta_h) \right\|_2^2, \quad (6)$$

where $\|\cdot\|_2$ is the ℓ-2 norm, and N_s and N_t denote the number of training examples of signers s and t, respectively. Accordingly, the signer-transfer loss at the m-th layer is the sum of the pairwise distances between all signers, i.e.:

$$\mathcal{L}^{(m)}_{\text{transfer}}(\theta_h) = \sum_{s\in\mathbb{S}}\sum_{t\in\mathbb{S}\wedge t\neq s} \mathcal{D}^{(m)}(s,t;\theta_h). \quad (7)$$

The overall signer-transfer loss $\mathcal{L}_{\text{transfer}}$ is then a weighted sum of the losses computed at each layer of the *encoder* network, such that:

$$\mathcal{L}_{\text{transfer}}(\theta_h) = \sum_{m=1}^{M} \beta^{(m)} \mathcal{L}^{(m)}_{transfer}(\theta_h), \quad (8)$$

where $\beta^{(m)}$ is a hyperparameter that controls the relative importance of the loss obtained at the m-th layer. By combining (5) and (8), the *encoder* and *sign-classifier* networks are trained to minimize the following loss function:

$$\min_{\theta_h,\theta_f} \mathcal{L}(\theta_h,\theta_f,\theta_g) = \mathcal{L}_{\text{sign}}(\theta_h,\theta_f) + \lambda\mathcal{L}_{\text{adv}}(\theta_h,\theta_g), \quad (9)$$

where $\gamma \geq 0$ is the weight that controls the relative importance of the signer-transfer term.

Summing up, the adversarial training procedure is organized by alternatively either training both the *encoder* and the *sign-classifier* in order to minimize objective (9) or training the *signer-classifier* in order to minimize objective (1).

IV. Experimental Evaluation

The experimental evaluation of the proposed model was performed using three publicly available SLR databases: the Jochen-Triesch database [23], the Microsoft Kinect and Leap Motion American sign language (MKLM) database [24], [25], and the Portuguese Sign Language and Expressiveness Recognition (SI-PSL) database [26]. Jochen-Triesch [23] is a dataset of 10 hand signs performed by 24 signers against three different types of backgrounds: uniform light, uniform dark and complex. Experiments on Jochen-Triesch were conducted using the standard evaluation protocol of this dataset [27], in which 8 signers are used for the training and the remaining 16 signers are used for the test. MKLM [24], [25] contains a total of 10 signs, each one repeated 10 times by 14 different signers. In this dataset, the performance of the models is assessed using 5 random splits, created with signer-independence, yielding at each split a training set of 10 signers, a validation set of 2 signers and a test set of 2 signers. The SI-PSL database contains 31 isolated signs, representing the alphabet and the cardinal numbers 0 to 9 of the Portuguese sign language. All

the signs were performed three times by 11 native signers, in a free and natural signing environment, without any clothing restriction but with a slightly controlled uniform background. SI-PSL has a well-defined standard evaluation protocol, which consists of 6 signers for training, 1 signer for validation and the remaining 4 signers are used for testing.

A. Implementation Details

In order to extract the manual signs from the noisy background of the images, the automatic hand detection algorithm [28] is used as a pre-processing step. The images are then cropped, resized to the average sign size of the training set, and normalized to be in the range [−1, 1]. Throughout this section, the proposed model is compared with state-of-the-art methods for each dataset [15], [16], [24], [27], [28]. Nevertheless, to further attest the robustness of the proposed model, two different baselines are also implemented:

1) (Baseline 1) A CNN trained from scratch with ℓ-2 regularization. For a fair comparison, the architecture of the baseline CNN corresponds to the architecture of the *encoder* network followed by the *sign-classifier* network of the proposed model.
2) (Baseline 2) A CNN with the baseline 1 topology, but trained with the triplet loss [29].

TABLE I: HYPERPARAMETERS SETS

Hyperparameters	Acronym	Set
Leaning rate	-	{1e−04,1e−03}
ℓ-2 norm coefficient	-	{1e−05,1e−04}
$\mathcal{L}_{\text{triplet}}$ weight	ρ	{0.1,0.5,1,5,10}
$\mathcal{L}_{\text{adv}}$ weight	λ	{0.1,0.5,0.8,1,3}
$\mathcal{L}_{\text{transfer}}$weight	γ	{1.5e−04,2e−04,4e−04,1e−03}

Here, the triplet loss concept is explored in order to impose signer-independence in the representation space and, hence, build up a more robust baseline. The underlying idea is to minimize the distance between an *anchor* and a *positive* latent representation, h_{y_i,s_i} and h_{y_p,s_p}, respectively; while maximizing the distance between the *anchor* h_{y_i,s_i} and a *negative* representation h_{y_n,s_n}. It is important to note that while *anchor* and *positive* latent representations have to be from the same sign class, their signer identity may or not change. On the other hand, *anchor* and *negative* representations are from different sign classes, whereas their signer identity may also change. In order to train baseline 2 in an end-to-end fashion for sign classification, the overall loss function to be minimized is a trade-off between the triplet loss $\mathcal{L}_{\text{triplet}}$ and a classification loss $\mathcal{L}_{\text{sign}}$, such that:

$$\mathcal{L} = \mathcal{L}_{\text{sign}} + \frac{\rho}{N}\sum_{i=1}^{N}\left\|h_{y_i,s_i} - h_{y_p,s_p}\right\|_2^2 - \left\|h_{y_i,s_i} - h_{y_n,s_n}\right\|_2^2, \quad (10)$$

where $\mathcal{L}_{\text{sign}}$corresponds to the categorical cross-entropy as defined in equation (2). The second term denotes the L_{triplet}, where $y_p = y_i$and$y_n \neq y_i$, and $\rho \geq 0$ is a hyperparameter controlling its relative importance. The margin enforced between *positive* and *negative* pairs was fixed as $\alpha = 1$. In addition, following [29], an *online* triplet generation strategy, by selecting the hardest *positive*/*negative* samples within every mini-batch, was adopted.

All deep models were implemented in PyTorch and trained with the Adam optimization algorithm using a batch size of 32 samples. For reproducibility purposes, the source code as well as the weights of the trained models are publicly available online[1]. The hyperparameters that are common to all the implemented models (i.e., learning rate and ℓ -2 regularization weight) as well as some hyperparameters that are specific to the proposed model (i.e., λ and γ) and to the implemented baseline 2 (i.e., ρ) were optimized by means of a grid search approach and cross-validation on the training set (see Table I for more details). The signer-transfer penalty $\mathcal{L}_{\text{transfer}}$ is applied to the last two layers of the *encoder* network with a relative weight of 1. Regarding the model's architecture, the number of consecutive convolutional layers pairs L_e was set to 3, which results in a total of 6 convolutional layers. The number of filters starts as 32, which is then doubled after each convolutional pair. The dense layer on top of the *encoder* network has 128 neurons. The number of dense layers of both classifiers L_s was set to 3, and the number of nodes of each hidden layer was set as 128.

TABLE II: JOCHEN-TRIESCH EXPERIMENTAL RESULTS. RESULTS ARE REPORTED IN TERMS OF AVERAGE CLASSIFICATION ACCURACY. THE FIRST BLOCK OF THE TABLE PRESENTS THE RESULTS OF STATE-OF-THE-ART METHODS. THE SECOND BLOCK DEPICTS THE RESULTS OF THE PROPOSED MODEL AND OF BOTH IMPLEMENTED BASELINES

Method	Classification accuracy (%)		
	Background		
	Uniform	*Complex*	*Both*
Just *et al* [27]	92.79	81.25	87.92
Kelly *et al*. [15]	91.80	-	-
Dahmani *et al* [16]	93.10	-	-
CNN (Baseline 1)	97.50	74.38	89.79
CNN with Triplet loss (Baseline 2)	98.13	75.63	90.63
Proposed method	98.75	91.25	96.25
CNN (Baseline 1) with T.L.	100.00	98.75	99.58
CNN with Triplet loss (Baseline 2) with T.L.	99.69	97.50	98.96
Proposed method with T.L.	100.00	99.38	99.79

TABLE III: MKLM EXPERIMENTAL RESULTS. THE RESULTS ARE REPORTED IN TERMS OF AVERAGE CLASSIFICATION ACCURACY. THE FIRST BLOCK OF THE TABLE PRESENTS THE RESULTS OF STATE-OF-THE-ART METHODS. THE SECOND BLOCK DEPICTS THE RESULTS OF THE PROPOSED MODEL AND OF BOTH IMPLENTED MODELS WITH TRANSFER LEARNING

Method	Classification accuracy (%)		
	average (std)	min	max
Marin *et al* [24]	89.71 (-)	-	-
Ferreira *et al* [28]	93.17 (-)	-	-
CNN (Baseline 1)	89.90 (8.81)	73.00	98.00
CNN with Triplet loss (Baseline 2)	91.40 (3.93)	86.50	96.50
Proposed method	94.80 (3.53)	89.50	100.00
CNN (Baseline 1) with T.L.	97.30 (1.91)	94.50	100.00
CNN with Triplet loss (Baseline 2) with T.L.	98.50 (1.48)	96.00	100.00
Proposed method with T.L.	99.30 (0.60)	98.50	100.00

B. Results and Discussion

Experiments on Jochen-Triesch, MKLM, and SI-PSL databases are summarized in Tables II, III, and IV respectively. The results on the Jochen-Triesch database are presented in terms of average classification accuracy in the overall test set as well as against each specific background type (i.e., uniform and complex). For the MKLM database, Table III depicts the average classification accuracy computed across all the 5 test splits, as well as the minimum and maximum accuracy value achieved by each method. As

[1] https://github.com/pmmf/SI-SLR

the SI-PSL database is clearly the most challenging one and contains a large number of sign classes (i.e., 31), the results are presented in terms of top-1, top-3 and top-5 classification accuracy (see Table IV).

TABLE IV: SI-PSL EXPERIMENTAL RESULTS. THE RESULTS ARE REPORTED IN TERMS OF TOP-1, TOP-3 AND TOP-5 CLASSIFICATION ACCURACY. THE FIRST BLOCK DEPICTS THE RESULTS OF THE PROPOSED MODEL AND OF BOTH IMPLEMENTED BASELINES. THE SECOND BLOCK PRESENTS THE RESULTS OF THE IMPLEMENTED MODELS WITH TRANSFER LEARNING

Method	Classification accuracy (%)		
	Top-1	Top-3	Top-5
CNN (Baseline 1)	45.97	74.73	85.75
CNN with Triplet loss (Baseline 2)	42.74	72.31	81.99
Proposed method	49.13	76.01	85.19
CNN (Baseline 1) with T.L.	67.74	91.13	94.89
CNN with Triplet loss (Baseline 2) with T.L.	75.81	92.20	95.43
Proposed method with T.L.	76.08	94.89	98.12

The most interesting observation is the superior performance of the proposed model. Specifically, the proposed model provides the best overall classification accuracy across all the SLR databases, clearly outperforming both implemented baselines and all the previous state-of-the-art models. In complex scenarios, as reported in Table II, the proposed model surpasses all the other methods by a large margin (i.e., 91.25% against 81.25%, 74.38% and 75.63%). In addition, by analyzing the standard deviation as well as the minimum and maximum accuracy values, it possible to observe that the proposed model is the method with the lowest variability, yielding consistently high accuracy rates across all test splits of the MKLM dataset (see Table III). These results attest the robustness of the proposed model and its capability of better dealing with the large inter-signer variability that exists in the manual signing process of sign languages. Interestingly, the obtained results also reveal that the implemented baselines are in fact fairly strong models, both of them outperforming most of the state-of-the-art methods on both datasets. Finally, it is worth mentioning the superiority of the proposed model in the most challenging database (i.e., the SI-PSL). As shown in Table IV, the proposed model outperformed both the implemented baselines in all the three classification metrics.

C. Transfer Learning

To further improve the performance of the proposed model, we introduce a transfer learning strategy in the proposed adversarial training objective. Transfer learning aims to extract knowledge from one or multiple source tasks (or domains) and, then, use this prior knowledge when learning a model for a new target task [30]. Transfer learning techniques are particularly useful when we have to deal with limited sized training sets, as it happens in most available SLR databases. In this work, we applied a conventional transfer learning strategy that can be summarized as follows:

- The *encoder* network is initialized with the first 10 layers of VGG-19 [5], pre-trained on the ImageNet [6] database;
- During the first training epochs (≈ 30), the optimization algorithm is defined so that only the parameters of both classifiers are updated;
- In the remaining training epochs, the *encoder* network is fine-tuned for our particular task, which means that all the model parameters are updated.

It is important to note that for a fair comparison, we have also employed the same transfer learning strategy to both implemented baselines. The performance of the models with transfer learning is reported in the bottom blocks of Tables II, III, and IV. As it is possible to observe, transfer learning has brought substantial gains for all the models. Besides, the most important observation is that the proposed model remains the best method by a large margin.

D. Ablation Study

Table V depicts an ablation study of the proposed model, in which it is possible to assess the effect of each proposed training scheme. For this purpose, the proposed model was trained either (i) with just the adversarial procedure, without the signer-transfer $\mathcal{L}_{\text{transfer}}$loss, or (ii) with just the $\mathcal{L}_{\text{transfer}}$ penalty on the *encoder* network, without adversarial training. The results clearly demonstrate the complementary effect between the two training procedures, as their combination provides the best overall classification accuracy. Interestingly, each training scheme outperforms on its own both baselines and state-of-the-art methods.

TABLE V: THE EFFECT OF EACH TRAINING PROCEDURE IN THE PROPOSED MODEL. THE RESULTS IN THE LAST COLUMN ARE REPLICATED FROM TABLES II, III AND IV AS THEY INCLUDE BOTH TRAINING PROCEDURES

Dataset	Classification accuracy (%)		
	Only adversarial training	Only $\mathcal{L}_{\text{transfer}}$ penalty	Both
Jochen-Triesch	95.21	94.38	96.25
MKLM	94.00	94.10	94.80
SI-PSL	48.56	39.25	49.13

E. Latent Space Visualization

To further demonstrate the effectiveness of the proposed model in promoting signer-invariant latent representation spaces, we have performed a visual inspection of the latent representations through the t-distributed stochastic neighbor embedding (t-SNE) [31] (see Fig. 3). These plots clearly demonstrate the better capability of the proposed model of imposing signer-independence in the latent representations. The proposed model yields a latent representation space in which representations of different signers and same class are close to each other and well mixed, while it keeps latent representations of different classes far apart. By analyzing the t-SNE plot of baseline 1, it is possible to observe that the latent representations of different signers and the same class tend to be far apart in the latent space. In addition, there is some overlapping between clusters of different classes. Although baseline 2 (CNN with the triplet loss) promoted slightly improvements over the standard baseline CNN, the proposed model achieved by far the best signer-invariance and class separability.

F. Cluster Analysis in the Latent Space

In order to obtain an objective quality assessment of the produced latent representations, we have evaluated how well the model is able to cluster the different sign classes (and thus ignore the signer identity) in the latent space. For this purpose, we use two cluster validation metrics: the average Silhouette coefficient [32] per cluster and the Dunn's index

[33] per cluster.

The Silhouette coefficient for an observation i is computed as follows. Let C_i be the cluster (sign class) associated with the observation i. The average intra-cluster distance a_i and the minimum average inter-cluster distance b_i for the observation i are obtained as follows:

$$a_i = \frac{1}{|C_i|-1}\sum_{j\in C_i} d(i,j), \quad (11)$$

$$b_i = \min_{C\neq C_i} \frac{1}{|C|}\sum_{j\in C} d(i,j), \quad (12)$$

where $|C_i|$ denotes the number of observations in the cluster C_i and $d(i,j)$ is the Euclidean distance between the observations i and j. Then, the Silhouette index S_i for the observation i is defined as:

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}. \quad (13)$$

Clearly, $-1 \leq S_i \leq 1$. Intuitively, clusters are desirably compact (small a_i) and well separated (large b_i), so a larger value of S_i indicates better clustering. However, this metric is defined per observation. Hence, in order to have a global measure of clustering quality, we compute the average Silhouette coefficient for each cluster.

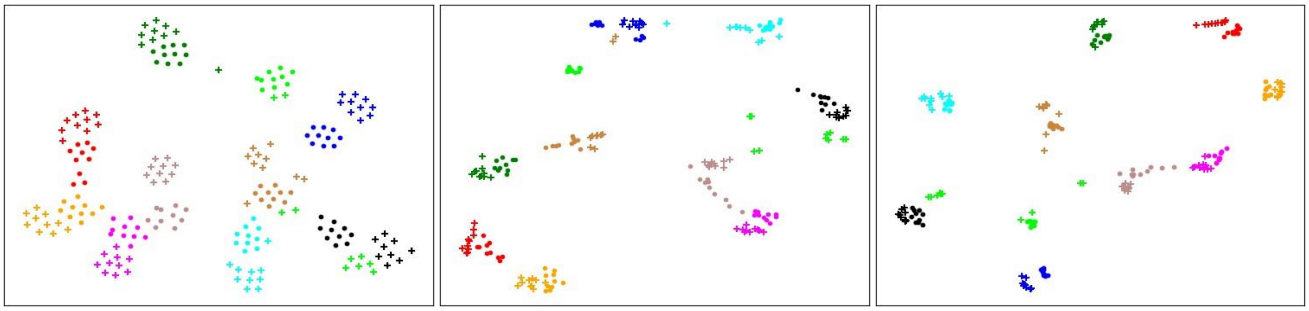
Dunn's index follows a similar idea of measuring cluster compactness versus separation, but uses minimum and maximum distances instead of average distances, and is more sensitive to extreme and occasional errors. Specifically, Dunn's index D_C for a cluster C is defined as the ratio between the minimum inter-cluster distance δ_C from C to all other clusters (which measures cluster separation) and the maximum intra-cluster distance Δ_C for the cluster C (which measures cluster compactness):

$$\delta_C = \min_{i\in C, j\notin C} d(i,j), \quad (14)$$

$$\Delta_C = \max_{i,j\in C} d(i,j), \quad (15)$$

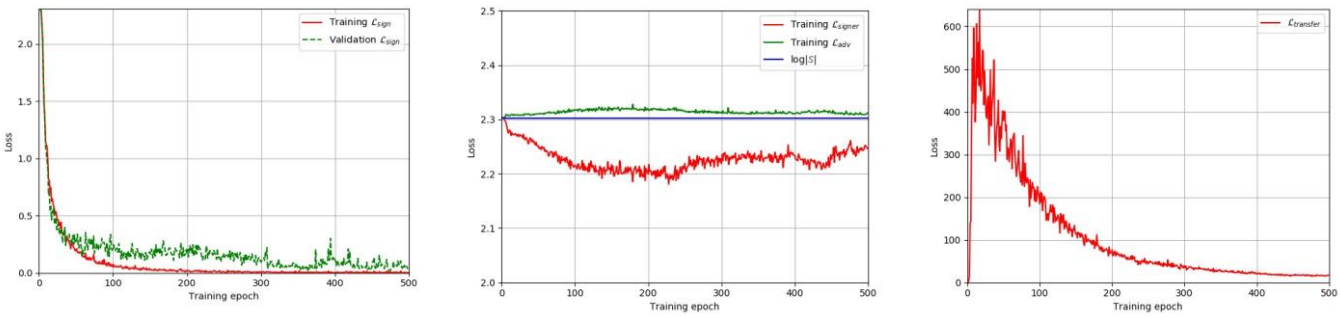
$$D_C = \frac{\delta_C}{\Delta_C}. \quad (16)$$

Again, according to this metric, larger values indicate better clustering. As anticipated by the analysis of the two-dimensional t-SNE projection in Fig. 3, the results confirm that the proposed model produces the most compact and separated sign clusters, when compared with the remaining models. This observation supports the signer-invariance property of the representations produced by the proposed adversarial training framework: when exposed to images obtained from new signers, our model does a better job of grouping them according to the respective sign class only, ignoring the signer identity.



(a) CNN - baseline 1 (b) CNN with triplet loss - baseline 2 (c) Proposed model

Fig. 3. Two-dimensional projection of the latent representation space using the t-distributed stochastic neighbor embedding (t-SNE) [31]. Markers • and + represent 2 different test signers, while the different colors denote the 10 sign classes.



(a) (b) (c)

Fig. 4. Training behavior of the proposed model: (a) the evolution of both training and validation $\mathcal{L}_{\text{sign}}$ curves; (b) the evolution of both $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{signer}}$ loss terms; and (c) the evolution of the $\mathcal{L}_{\text{transfer}}$ loss term.

G. Training Behavior of the Proposed Model

The evolution of the loss values along the training iterations is presented in Fig. 4. On Fig. 4a, one observes a small gap between training and validation sign classification losses, proving that the model is being regularized properly. This regularization effect is promoted by the adopted adversarial training and signer-transfer objectives, whose loss functions are depicted in Fig. 4b and Fig. 4c.

The adversarial training dynamics in Fig. 4b are an immediate consequence of the min-max game played between the signer-classifier and the encoder networks. The former aims to minimize $\mathcal{L}_{\text{signer}}$, while the latter tries to maximize it (by minimizing the surrogate $\mathcal{L}_{\text{adv}}$). Note that,

at the beginning of training, both losses are equal to $\log|\mathbb{S}|$, which is the entropy of the uniform distribution over signer identities and is the minimum possible value of $\mathcal{L}_{\text{adv}}$. This results from the fact that the untrained signer-classifier is just a random predictor. As training progresses, this network starts learning to predict correct signer identities from the provided latent representations. Therefore, $\mathcal{L}_{\text{signer}}$ starts decreasing and, consequently, $\mathcal{L}_{\text{adv}}$increases. The min-max game eventually leads to a point where both losses become stable and fairly close to their initial value, $\log|\mathbb{S}|$. This implies that, at the ending of training, the latent representations produced by the encoder network exhibit high signer-invariance, as desired.

The signer-transfer objective exhibits a smooth evolution along the training epochs, as shown in Fig. 4c. The exception is the first few training iterations, where the corresponding loss $\mathcal{L}_{\text{transfer}}$ increases rapidly, as the network weights depart from their initial values (which are close to zero). After this short period, the distribution of the latent representations of different signers start becoming closer and the loss decreases almost monotonically, until it eventually plateaus at a low value.

TABLE VI: DUNN'S INDEX AND SILHOUETTE COEFFICIENT FOR THE SIGN CLASS CLUSTERS IN THE LATENT SPACE FOR THE TEST DATA. THESE METRICS WERE COMPUTED PER CLUSTER AND THE AVERAGE AND WORST RESULTS ARE REPORTED FOR EACH MODEL AND DATASET

Method	Jochen-Triesch				MKLM				SI-PSL			
	Dunn's index		Silhouette		Dunn's index		Silhouette		Dunn's index		Silhouette	
	Average	Worst	Average	Worst	Average	Worst	Average	Worst	Average	Worst	Average	Worst
CNN (Baseline 1)	0.297	0.184	0.669	0.618	0.718	0.185	0.673	0.386	0.329	0.197	0.362	0.223
CNN with Triplet loss (Baseline 2)	0.481	0.277	0.689	0.581	0.965	0.326	0.733	0.531	0.414	0.276	0.411	0.304
Proposed method	0.593	0.275	0.753	0.690	1.012	0.351	0.758	0.630	0.405	0.305	0.470	0.346

V. CONCLUSION

This paper presents a novel adversarial training objective, based on representation learning and deep neural networks, specifically designed to tackle the signer-independent SLR problem. The underlying idea is to learn signer-invariant latent representations that preserve as much information as possible about the signs, while discarding the signer-specific traits that are irrelevant for sign recognition. For this purpose, we introduce an adversarial training procedure for simultaneously training an *encoder* and a *sign-classifier* over the target sign variables, while preventing the latent representations of the *encoder* to be predictive of the signer identities. To further discourage the underlying representations of retaining any signer-specific information, we propose an additional training objective that enforces the latent distributions of different signers to be as similar as possible. Experimental results demonstrate the effectiveness of the proposed model in several SLR databases.

AUTHOR CONTRIBUTIONS

Pedro M. Ferreira and Diogo Pernes conceived the model and performed the experiments. Pedro M. Ferreira wrote most parts of the document. Ana Rebelo and Jaime S. Cardoso supervised the work.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds, vol. 27, pp. 2672–2680, 2014.

[2] C. Feutry, P. Piantanida, Y. Bengio, and P. Duhamel, "Learning anonymized representations with adversarial neural networks," *arXiv preprint arXiv:1802.09386*, 2018.

[3] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by back propagation," in *Proc. the 32nd International Conference on Machine Learning*, Lille, France, Jul. 2015, vol. 37, pp. 1180–1189.

[4] P. M. Ferreira, D. Pernes, A. Rebelo, and J. S. Cardoso, "Learning signer-invariant representations with adversarial training," in *Proc. the 12th International Conference on Machine Vision*, 2019.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 2014.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[7] U. von Agris, C. Blomer, and K. Kraiss, "Rapid signer adaptation for continuous sign language recognition using a combined approach of eigenvoices, mllr, and map," in *Proc. 2008 19th International Conference on Pattern Recognition*, Dec 2008, pp. 1–4.

[8] R. Kuhn, J., Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 278, no. 6, pp. 695–707, Nov. 2000.

[9] T. Kim, W. R. Wang, H. Tang, and K Livescu, "Signer-independent fingerspelling recognition with deep neural network adaptation," *CoRR*, abs/1602.04278, 2016.

[10] F. Yin, X. J. Chai, Y. Zhou, and X. L. Chen, "Weakly supervised metric learning towards signer adaptation for sign language recognition," in *Proc. the British Machine Vision Conference*, September 2015, vol. 35, pp. 1–12.

[11] J. Zieren and K.-F. Kraiss, "Robust person-independent visual sign language recognition," *Pattern Recognition and Image Analysis*, pp. 520–528, Berlin, Heidelberg, 2005.

[12] T. Shanableh and K. Assaleh, "User-independent recognition of arabic sign language for facilitating communication with the deaf community," *Digital Signal Processing*, vol. 2721, no. 4, pp. 535-542, 2011.

[13] U. von Agris, J. Zieren, U. Canzler, B. ¨ Bauer, and K.-F. Kraiss, "Recent developments in visual sign language recognition," *Universal Access in the Information Society*, vol. 276, no. 4, pp. 323–362, Feb 2008.

[14] W. W. Kong and S. Ranganath, "Towards subject independent continuous sign language recognition: A segment and merge approach," *Pattern Recognition*, vol. 2747, no. 3, pp. 1294–1308, 2014.

[15] D. Kelly, J. McDonald, and C. Markham, "A person independent system for recognition of hand postures used in sign language," *Pattern Recognition Letters*, vol. 2731, no. 11, pp. 1359–1368, 2010.

[16] D. Dahmani and S. Larabi, "User-independent system for sign language finger spelling recognition," *Journal of Visual Communication and Image Representation*, vol. 2725, no. 5, pp. 1240–1250, 2014.

[17] F. Yin, X. J. Chai, and X. L. Chen, "Iterative reference driven metric learning for signer independent isolated sign language recognition," in *Proc. Computer Vision – ECCV 2016*, Cham, 2016, pp. 434–450.

[18] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *Proc. Computer Vision - ECCV 2014 Workshops*, Cham, 2015, pp. 572–578.

[19] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3793–3802.

[20] D. Wu, L. Pigou, P. Kindermans, N. D. Le, L. Shao, J. Dambre, and J. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2738, no. 8, pp. 1583–1597, Aug 2016.

[21] N. Neverova, C. Wolf, G. Taylor, and F. N. Moddrop, "Adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38, no. 8, pp. 1692–1706, Aug 2016.

[22] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing*, vol. 27259, pp. 21–38, 2017.
[23] J. Triesch and C. von der Malsburg, "A system for person-independent hand posture recognition against complex backgrounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 2723, no. 12, pp. 1449–1453, December 2001.
[24] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," in *Proc. 2014 IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 1565–1569.
[25] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with jointly calibrated leap motion and depth sensor," *Multimedia Tools and Applications*, vol. 2775, no. 22, pp. 14991–15015, Nov 2016.
[26] Corsil. (2019). A portuguese sign language and expressiveness recognition database. [Online]. Available: https://github.com/pmmf/CorSiL
[27] A. Just, Y. Rodriguez, and S. Marcel, "Hand posture classification and recognition using the modified census transform," in *Proc. 7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, April 2006, pp. 351–356.
[28] P. M. Ferreira, J. S. Cardoso, and A. Rebelo, "On the role of multimodal learning in the recognition of sign language," *Multimedia Tools and Applications*, Sept. 2018.
[29] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.
[30] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 2722, no. 10, pp. 1345–1359, 2010.
[31] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[32] P. J Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 2720, pp. 53–65, 1987.
[33] J. C Dunn, *A Fuzzy Relative of the Isodata Process and Its Use in Detecting Compact Well-Separated Clusters*, 1973.

Pedro M. Ferreira received the degree in biomedical engineering from the Politécnico do Porto in 2009, and the M.Sc. degree in biomedical engineering from the Faculdade de Engenharia da Universidade do Porto (FEUP) in 2012. He is currently pursuing the Ph.D. degree enrolled in the Doctoral Program in electrical and computer engineering at FEUP. He is a Researcher at INESC TEC. His main research interests include computer vision, machine learning and artificial intelligence.

Diogo Pernes graduated in electrical and computers engineering with a 5-year master's degree at Universidade do Porto in 2014.

He is currently a Ph.D. candidate in computer science at the same university, working as a researcher at INESC TEC.

His main research topics are machine learning and computer vision.

Ana Rebelo was born in Porto, Portugal, in 1985. She received the degree in mathematics applied to technology from the School of Sciences, University of Porto, Portugal, in 2007, and the M.Sc. degree in mathematical engineering from the School of Sciences, University of Porto, Portugal, in 2008. She is currently pursuing the Ph.D. degree with the School of Engineering, University of Porto, Portugal. Since 2007, she has been a researcher at INESC TEC, an R&D Institute affiliated to University of Porto, Visual Computing and Machine Intelligence Group (VCMI). She was a project member of one FCT (Foundation of Science and Technology - Portugal) Research Project in the area of optical music recognition. She is currently an assistant professor with the Universidade Portucalense Infante D. Henrique. She is also a senior researcher at INESC TEC. Her main research interests include computer vision, image processing, biometrics, and document analysis.

Jaime S. Cardoso received the Licenciatura (5-year degree) in electrical and computer engineering in 1999, the M.Sc. degree in mathematical engineering in 2005, and the Ph.D. degree in computer vision in 2006, all from the University of Porto. He is currently an Associate Professor with Habilitation at the Faculty of Engineering of the University of Porto (FEUP) and also a coordinator of the Centre for Telecommunications and Multimedia, INESC TEC. He has co-authored more than 200 papers, over 60 of which in international journals, which attracted over 2800 citations, according to Google scholar. His research can be summed up in three major topics: computer vision, machine learning, and decision support systems. Image and video processing focuses on biometrics and video object tracking for applications such as surveillance and sports. The work on machine learning cares mostly with the adaptation of learning to the challenging conditions presented by visual data. The particular emphasis of the work in decision support systems goes to medical applications, always anchored on the automatic analysis of visual data.