Turk J Elec Eng & Comp Sci
() : –

Signer-Independent Sign Language Recognition with Feature Disentanglement

İpek Erdoğan[1], **İnci M. Baytaş**[2*]
[1,2]Department of Computer Engineering, Faculty of Engineering, Boğaziçi University, İstanbul, Turkey,
ORCID iD: https://orcid.org/0000-0003-4765-2615

Received: .202 • **Accepted/Published Online:** .202 • **Final Version:** ..202

Abstract: Learning a robust and invariant representation of various unwanted factors in Sign Language Recognition applications is essential. One of the factors that might degrade the sign recognition performance is the lack of signer diversity in the training datasets, causing a dependence on the singer's identity during representation learning. Consequently, capturing signer-specific features hinders the generalizability of sign language recognition systems. This study proposes a feature disentanglement framework comprising a Convolutional Neural Network and a Long Short-Term Memory network based on adversarial training to learn a signer-independent sign language representation that might enhance the recognition of signs. We aim to improve the feature representations by incorporating various regularization techniques to facilitate feature disentanglement. Particularly, Kullback Leibler divergence between uniform distribution and output of a signer classifier is employed to reduce the effect of signer identity on spatial embeddings. Similarly, the Optimal Transport distance and Mean Square Error are investigated to minimize the disparity between the spatial and temporal representations of the same signs performed by different signers. The proposed framework is evaluated on two Turkish isolated sign language datasets constituting varying characteristics and challenges. The qualitative results show that the proposed feature disentanglement framework helps reduce the influence of the signer's identity on the sign representations. According to the quantitative analyses, the best performances of 94% and 89% classification accuracy are obtained for two Turkish sign language benchmark datasets, BosphorusSign22k and Ankara University Turkish Sign Language (AUTSL) datasets, respectively.

Key words: Sign language recognition, adversarial learning, disentangled representation learning, deep neural networks

Nomenclature

$AUTSL$ Ankara University Turkish Sign Language Dataset

CNN Convolutional Neural Network

GAN Generative Adversarial Network

GCA Gloss Classification Accuracy

GCN Graph Convolutional Network

KL Kullback-Leibler

$LSTM$ Long Short-Term Memory

MSE Mean Squared Error

*Correspondence: inci.baytas@bogazici.edu.tr

OT Optimal Transport

SLR Sign Language Recognition

VAE Variational Auto-Encoder

1. Introduction

Sign language enables primary communication with the Deaf community. Although there are more than 300 sign languages around the world [1], accessing educational material for sign languages and assistance is still a challenge [2]. Therefore, sign languages are considered under-resourced languages [3] which lack digital resources [4]. Automatic Sign Language Recognition (SLR) facilitates designing tools and generating digital material to improve sign language education and applications, such as sign language translation [5, 6] and animation [7]. For a successful SLR framework, powerful representations should be learned from sign videos containing the signer's face, body, and hand gestures. Here, we can define a powerful representation of sign video in terms of capturing the distinct actions of a sign, discarding the signer identity contaminating the representation.

Recognizing sign languages requires learning a unified representation for the gestures of multiple body parts, such as torso, face, and hands [8], ignoring the signers' environmental factors and identity. On the other hand, the number of unique signers in the SLR frameworks is often insufficient to attain signer independence and generalization. Although the goal is not to identify signers, their identities, outfits, and accessories might be captured by the SLR frameworks. As a result, a significant gap between the training and evaluation performances might occur under a singer-independent evaluation protocol, where videos of test signers are not included in the training set. Therefore, reducing the signer dependency in SLR frameworks is crucial.

One of the causes of singer dependency in the SLR frameworks is the lack of signer variations in the training dataset. The limited number of signers leads to overfitting, a prevalent issue in Convolutional Neural Network (CNN) based SLR frameworks. Some studies consider skeletal joint data to reduce the signer characteristics in the learned sign representations [9, 10]. However, sign language datasets do not necessarily contain ground-truth annotations for key joints. In such cases, whole-body, face, and hand joints should be extracted first. Moreover, skeleton-based SLR models constitute several challenges, such as difficulty in capturing the high-level spatial structure [11]. Therefore, when an RGB-based SLR framework is preferred, ensuring its invariance to signer characteristics is critical.

This study poses the signer independence in RGB-based SLR frameworks as a disentangled representation learning problem. The proposed SLR framework aims to disentangle singer features from the sign features to improve the generalizability of the SLR model. An isolated SLR task, where each video contains a single gloss. A gloss could be defined as the word associated with a sign. The spatial and temporal representations are learned using a CNN encoder and a Long Short Term Memory (LSTM) network. Complex architectures are not chosen to observe the contribution of the feature disentangling. For this purpose, an adversarial training framework is proposed with several regularization techniques to reinforce the feature disentanglement further. The contributions of the study are outlined below.

- Inspired by the unsupervised domain adaption technique [12], we train a signer classifier and the sign recognition module to disentangle the spatial representation of the sign gloss from the signer.
- We investigate the Kullback–Leibler (KL) divergence between the signer predictions and uniform distribution to reinforce feature disentanglement. Regularizing the adversarial training with the KL divergence intends

to compel the encoder to learn spatial representations confusing the signer classifier.

- We investigate the Optimal Transport (OT) distance between the signer predictions of the same glosses. The goal of the OT distance is to facilitate learning spatial representations with fewer variations for the same glosses due to different signers.

An extensive ablation study is conducted to explore the performance contributions of the different components in the proposed framework. The proposed approaches are evaluated on two publicly available Turkish sign language datasets, BosphorusSign22k [13] and Ankara University Turkish Sign Language Dataset (AUTSL) [14]. Both quantitative and qualitative analyses are provided. The rest of the manuscript is structured as follows: A literature review on SLR and feature disentanglement for closely related gait recognition is provided in Section 2. The proposed model, loss functions, and the training framework are introduced in Section 3. Datasets, experimental setup, and the results of quantitative and qualitative analyses are presented in Section 4. Finally, the conclusion and future work are discussed in Section 5.

2. Related Work

This study poses the signer-independent SLR as a disentangled representation learning problem. Disentangled representation learning has been popularly applied to recognition tasks. For instance, Tran et al. proposed an adversarial disentangled representation learning framework to alleviate the dependence on the pose in face recognition [15]. For this purpose, the authors trained an auto-encoder focusing only on identity. Meanwhile, a multi-task discriminator predicting identity and pose is trained along with a generator conditioned on the input face image [15]. Likewise, Liu et al. presented an encoder, a generator, and discriminators for disentangling the content and style of images [16]. In addition, Oldfield et al. [17] also preferred to train a Generative Adversarial Network (GAN) so that the latent representation of a facial attribute can be disentangled from the other attributes. The typical approach in the studies mentioned above [15, 16] is utilizing GAN loss to encourage disentangling a face image's pose, shape, and identity. On the other hand, the SLR problem also requires temporal modeling.

Feature disentangling is also essential for the gait recognition problem. Gait recognition is concerned with learning to represent a person's manner of walking rather than their outfit and accessories. In this regard, gait recognition and SLR have some similarities, such that both tasks require learning a representation of an individual's posture and movement. In gait recognition, the learned representation is desired to embed the individual's identity but to be invariant to external factors such as outfit. On the other hand, representation learned for a sign should not carry any information indicating the identity of the individual who performs it. Although they aim to be invariant to different factors, they could benefit from feature disentanglement. However, when a CNN-based framework is used, unwanted information due to variations in people's appearance might leak into the learned gait representation. Zhang et al. [18] addressed this issue by learning appearance invariant gait representations from walking videos using auto-encoders. Their proposed framework is regularized with the disparity between the gait representations of the same individual with different appearances resulting in feature disentanglement [18]. Li et al. focused on the re-identification of a subject without paying attention to their outfits with a Color Agnostic Shape Extraction Network (CASE-NET) that is trained with an adversarial loss [19]. Meanwhile, Yang and Yao aimed to disentangle the representation learned for hand pose from the background and camera viewpoint utilizing Variational Auto-Encoders (VAEs) [20].

A few studies focus on disentangled representation in the SLR literature. For instance, Ferreira et al. utilized an adversarial loss for signer-independent SLR [21]. The authors employed Kullback-Liebler (KL)

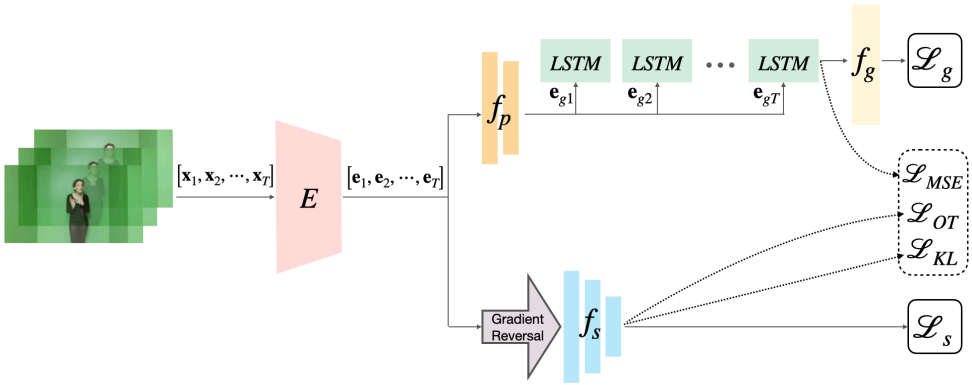


Figure 1: The proposed framework. The gradient reversal returns negative feedback from the signer classifier to update the spatial encoder. Dashed lines denote additional regularizers utilizing the encoder, LSTM, and signer classifier outputs. Depending on the dataset, one or a combination of the regularizers might be employed.

divergence to enforce a signer classifier's predictions to become uniform, encouraging the removal of the signer's characteristics from the sign representation. However, unlike this study, Ferreira et al. considered only hand images with a limited number of signs [21]. Zhang et al. focused on a similar task, speaker-independent lipreading for disentangled visual speech recognition [22]. Like Ferreira et al. [21], a speaker classifier with KL divergence is employed to disentangle speech features from the speaker's identity.

This study proposes an adversarial training framework for the signer-independent SLR task. The proposed architecture comprises spatial and temporal representation learning modules. A signer classifier and additional regularizers support disentanglement, boosting the disparity between sign and signer representations. The contribution of the proposed framework to signer-independent representation learning is investigated for shallow and deeper architectures with two publicly available Turkish sign language datasets of varying challenges.

3. Method

Isolated SLR is defined as a multi-class classification problem where learning a representation of a sign video's spatial and temporal patterns is essential. In this study, we propose the architecture in Figure 1 comprising a CNN encoder, an LSTM network, and fully connected projection layers. The entire framework is trained in an end-to-end manner.

3.1. Spatial and Temporal Modeling

Spatial embeddings for sign language video frames are obtained using a convolutional encoder. This study investigates the feature disentanglement with both shallow and deep CNN encoders. Frame embeddings are later projected into a gloss space as follows

$$\mathbf{e}_i = E\left(\mathbf{x}_i; \theta_E\right), \quad i = 1, \cdots, T \tag{1}$$

$$\mathbf{e}_{gi} = f_p\left(\mathbf{e}_i\right), \quad i = 1, \cdots, T \tag{2}$$

where $\mathbf{x}_i$ is the ith frame of the input video, $E\left(\cdot\right)$ denotes a convolutional encoder parameterized by θ_E, f_p denotes fully-connected layers to project the frame embeddings into the gloss space, and T is the total number of frames considered in the input video. Thus, the spatial embeddings of the sign video form a multivariate sequence that corresponds to a sign gloss. The sequential nature of the gloss video entails temporal modeling.

The LSTM network [23] is one of the most popular Recurrent Neural Network (RNN) architectures used for sequential modeling. In this study, we train an LSTM network with the embeddings in the gloss space, $\mathbf{e}_{gi}$. For each gloss, we obtain a sequence of hidden states of the LSTM network as follows:

$$\{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T\} = \text{LSTM}\left(\{\mathbf{e}_{g1}, \mathbf{e}_{g2}, \cdots, \mathbf{e}_{gT}\}; \theta_{\text{LSTM}}\right) \tag{3}$$

where $\mathbf{h}_i$ is the hidden state that represents the sequence up to i-th time step, θ_{LSTM} denotes the parameters of the LSTM network, and $\mathbf{h}_T$ is utilized as the temporal representation of the sequence of spatial embeddings. The temporal representations are further used to train a sign classifier, f_g, which is the target task.

3.2. Feature Disentanglement with Adversarial Training

Spatial embeddings are supposed to contain a spatial pattern characteristic of the sign action independent of the individual who performs it. However, the spatial embeddings are susceptible to the attributes of the signers, such as their faces and posture. As a result, the generalization of SLR frameworks deteriorates. Inspired by domain adaptation, we add a fully connected signer classifier f_s as in Figure 1 that is simultaneously trained with the spatial embeddings, $\mathbf{e}_i$, to reduce the effect of discriminative signer attributes on the sign representation. The encoder weights, θ_E, are updated to confuse the signer classifier's decision. For this purpose, gradient reversal layer [12], shown in Eq. 4 is included before the signer classifier.

$$\mathcal{R}\left(\mathbf{e}_i, f_s; \alpha\right) = \begin{cases} \mathbf{e}_i, & \text{forward propagation} \\ -\alpha \frac{\partial f_s}{\partial \mathbf{e}_i}, & \text{backpropagation} \end{cases} \tag{4}$$

where α is a hyperparameter to control the amount of the gradient reversal effect.

When the gradient reversal is utilized after the CNN encoder, the encoder weight updates are encouraged to maximize the signer classifier's loss. Meanwhile, the signer classifier, f_s weights are updated to improve the classifier's signer prediction. Thus, adversarial training is intended to remove signer features from the spatial sign representation. However, our experiments show that adversarially training a signer classifier is insufficient to improve an SLR framework's signer-independent generalization. Therefore, the adversarial training is regularized with the discrepancy between the video representations of the same gloss performed by different signers.

3.3. Regularized Adversarial Training

Signer dependency in the SLR is due to the disparity between the latent representations learned for a gloss performed by different signers. However, to facilitate sign recognition, the distinction between representations of different glosses should be attained, while similar representations with slight variances should be maintained for all the input videos of the same gloss. Therefore, we propose to promote signer-independency with regularization based on reducing the discrepancy between distributions of gloss representations performed by different signers. Mainly, KL divergence, OT distance, and Mean Square Error (MSE) are utilized for this purpose.

KL divergence and the OT distance can be used to measure the difference between two distributions. We employ the KL divergence given in Eq. 5 to quantify the divergence between the signer embeddings and uniform distribution.

$$\mathcal{L}_{\text{KL}}\left(p||q\right) = \sum_i p(i) \log \frac{p(i)}{q(i)}, \tag{5}$$

where $p(i)$ is the uniform distribution and $q(i)$ is the signer embedding output of the encoder. When included in the adversarial training, KL divergence between the signer and uniform distribution enforces the frame embedding updates in a way that a signer classifier cannot capture any difference between the frames of different signers. The KL divergence is the number of bits required to transform one distribution to another. When the KL divergence between the signer distribution and the uniform distribution, where each signer is equally likely, is used to regularize the SLR framework, the training will be guided to create ambiguity regarding the signer identity in the learned frame representations. In other words, the optimizer will try to convert the signer distribution into a uniform distribution to reduce the KL divergence. Thus, the KL divergence regularization is considered one of the methods that could facilitate disentangling the signer identity from the sign representation.

In the existence of invariance to signer identity, the learned representations of the same sign are expected to be similar to each other, facilitating its recognition. Therefore, we could similarly penalize the distance between the learned representations of the same signs performed by the different signers. Instead of directly minimizing the Euclidean distances between the representations between the individual samples that might be sensitive to outliers, we aim to reduce the discrepancy between signer distributions of the same signs over a mini-batch. For this purpose, OT distance given in Eq. 6 is preferred. The OT distance is a metric measuring the minimum cost to transport one distribution to another [24]. It is well-known for alleviating issues in GAN training, such as mode collapse [25, 26]. Compared to KL divergence, the OT distance is a valid metric with symmetry. The OT distance might be preferred over the KL divergence under certain circumstances in which the gradients of the KL divergence are ineffective [24]. The OT distance is computed for a mini-batch of signer pairs performing the same sign.

$$\mathcal{L}_{\mathrm{OT}}\left(\boldsymbol{\mu}_{s_i}, \boldsymbol{\mu}_{s_j}\right) = \inf_{\gamma \in \Pi\left(\boldsymbol{\mu}_{s_i}, \boldsymbol{\mu}_{s_j}\right)} E_{(\mathbf{s}_i, \mathbf{s}_j) \sim \gamma} c\left(\mathbf{s}_i, \mathbf{s}_j\right) \tag{6}$$

where $\Pi\left(\boldsymbol{\mu}_{s_i}, \boldsymbol{\mu}_{s_j}\right)$ is the set of joint distributions, γ, with marginal distributions of $\boldsymbol{\mu}_{s_i}$, $c\left(\mathbf{s}_i, \mathbf{s}_j\right)$ is the Euclidean distance, and $\mathbf{s}_i = \frac{1}{T}\sum_{t=1}^{T} f_s\left(\mathbf{e}_t^i\right)$ and $\mathbf{s}_j = \frac{1}{T}\sum_{t=1}^{T} f_s\left(\mathbf{e}_t^j\right)$ are the average pre-softmax logits over the frames corresponding to two videos of the same gloss performed by signers i and j, respectively. When the optimization problem tries to minimize the OT distance between $\mathbf{s}_i$ and $\mathbf{s}_j$, both encoder and LSTM modules are updated to output similar logits for the videos of the same signs regardless of the signer's identity. Thus, the OT distance regularization may reduce the variance in the gloss representations of different signers.

We also investigate directly reducing the difference between the temporal embeddings of a sign gloss performed by different signers by imposing the MSE loss regularization. The MSE loss below aims to boost the signer independence by minimizing the distances between the gloss representations in the latent space of LSTM obtained from the videos of different signers.

$$\mathcal{L}_{\mathrm{MSE}} = \sum_{i=1}^{N} ||\mathbf{h}_i^+ - \mathbf{h}_i^-||_2^2, \tag{7}$$

where $\mathbf{h}_i^+$ and $\mathbf{h}_i^-$ are the hidden states of the LSTM unit's last time step for the same gloss performed by two different signers. Unlike KL divergence and OT distance, the MSE loss regularization directly penalizes the discrepancy between the gloss embeddings used for recognition.

3.4. Loss Function and Optimization

The proposed regularized adversarial training framework is concerned with optimizing the following loss function.

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_s + \mathcal{L}_{\text{REG}}. \tag{8}$$

where $\mathcal{L}_g$ and $\mathcal{L}_s$ given in Eq. 9 and 10 denote cross-entropy loss functions for gloss and signer classification, respectively.

$$\mathcal{L}_g = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C_g} y_{ij}\log f_{g_j}\left(\mathbf{h}_i\right), \tag{9}$$

$$\mathcal{L}_s = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C_s} y^s_{ij}\log f_{s_j}\left(\tilde{\mathbf{e}}_i\right), \tag{10}$$

where N is the batch size, C_g and C_s denote the number of unique glosses and signers, respectively, $\mathbf{h}_i$ denotes the hidden state of the LSTM network's last time step for the ith video, $\tilde{\mathbf{e}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathcal{R}\left(\mathbf{e}_t, f_s; \alpha\right)$, y is the ground truth label for the gloss classification and y^s is the ground truth label for the signer classification.

In Eq. 8, $\mathcal{L}_{\text{REG}}$ is the regularizer which can be chosen as KL divergence, OT distance and the MSE loss described in Section 3.3. We analyze the benefits of these regularizers in the experiments. The contribution of the terms in Eq. 8 to the signer-independent representation learning can be examined as shown below.

$$\frac{\partial\mathcal{L}}{\partial\theta_E} = \frac{\partial\mathcal{L}_g}{\partial\theta_E} - \alpha\frac{\partial\mathcal{L}_s}{\partial\theta_E} + \frac{\partial\mathcal{L}_{\text{REG}}}{\partial\theta_E} \tag{11}$$

$$\frac{\partial\mathcal{L}}{\partial\theta_{\text{LSTM}}} = \begin{cases} \frac{\partial\mathcal{L}_g}{\partial\theta_{\text{LSTM}}} & , \quad if \quad \mathcal{L}_{\text{REG}} = \mathcal{L}_{\text{KL}} \quad or \quad \mathcal{L}_{\text{OT}} \\ \frac{\partial\mathcal{L}_g}{\partial\theta_{\text{LSTM}}} + \frac{\partial\mathcal{L}_{\text{MSE}}}{\partial\theta_{\text{LSTM}}} & , \quad if \quad \mathcal{L}_{\text{REG}} = \mathcal{L}_{\text{MSE}} \end{cases} \tag{12}$$

$$\frac{\partial\mathcal{L}}{\partial\theta_g} = \frac{\partial\mathcal{L}_g}{\partial\theta_g} \tag{13}$$

$$\frac{\partial\mathcal{L}}{\partial\theta_s} = \frac{\partial\mathcal{L}_s}{\partial\theta_s} \tag{14}$$

where θ_E, θ_{LSTM}, θ_g and θ_s denote the parameters of CNN encoder, LSTM network, fully connected layers used for gloss classification and fully connected layers used for signer classification, respectively. The expressions above indicate that when KL divergence or OT distance is used with gradient reversal, the CNN encoder weights are updated to fool the signer classifier, minimize the regularize, and minimize the cross-entropy loss for sign recognition, $\mathcal{L}_g$. At the same time, the LSTM network's weights are updated to minimize $\mathcal{L}_g$ and the MSE loss if used as the regularizer. If KL divergence and OT distance are computed with the encoder embeddings, LSTM weights will be updated with the gradient of the gloss classification loss only due to the path of the backpropagating error. If the regularization terms take the encoder's output as their only input, then the backpropagation will flow starting from the regularization objective and through the encoder network. Therefore, only the derivative of the regularization objectives with respect to the encoder weights will be computed by the backpropagation algorithm. Therefore, under such cases, LSTM weights will not be updated with the error backpropagated from the regularization terms. Finally, the gloss and signer classifiers' weights, θ_g, θ_s are updated to improve their corresponding classification performance.

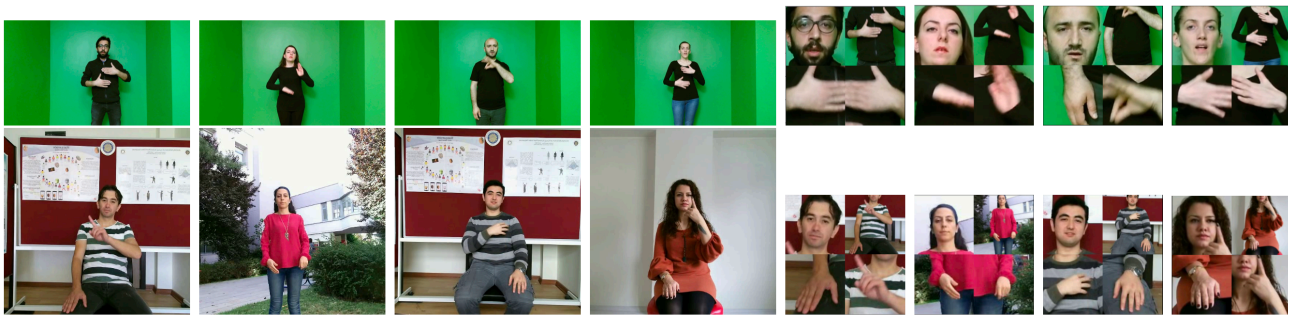


Figure 2: The first and the second rows demonstrate samples from BosphorusSign22k [13] and AUTSL [14] datasets, respectively. The samples after the pre-processing step are presented in the last four columns.

4. Experiments

The proposed framework is evaluated based on quantitative and qualitative measures. Quantitative results are presented for shallow and deeper CNN encoders, ResNet-18, and ResNet-34 architectures. All the models are trained for 100 to 200 epochs. The best-performing checkpoints are used for the final evaluation. Each epoch took around 20 and 45 minutes for shallow and deep networks used in the experiments, respectively. All the experiments are conducted with two Turkish sign language data sets, BosphorusSign22k [13] and AUTSL [14]. Gloss Classification Accuracy (GCA) is calculated below to evaluate the isolated SLR performance.

$$GCA = \frac{\text{number of correct predictions}}{\text{total number of samples}} \tag{15}$$

GCA corresponds to classification accuracy. It is a commonly reported metric to evaluate the recognition performance of isolated signs. GCA is considered the primary metric for performance evaluation since there is no class imbalance issue in the training datasets used in the experiments. All the metrics are macro averages over the number of classes. Training, validation, and test splits provided by the benchmark datasets are used in training and evaluation. The softmax probabilities of predictions are sorted in descending order to report Top-1, 3, and 5 accuracies. The true positive is obtained if a correct prediction is observed in the top one, among the top three, and the top five predictions with the highest probabilities for top-1, top-3, and top-5 accuracies, respectively.

4.1. Datasets

Both BosphorusSign22k [13] and AUTSL [14] have undergone a similar pre-processing approach as Gökçe et al. [27] where signer's face, torso, and hands are cropped to form new frames. Thus, facial expressions and hand gestures become more prominent on the input frames. Sample and pre-processed frames from both datasets are presented in Figure 2.

BosphorusSign22k dataset [13] comprises 22.542 videos of 744 signs. The dataset is split into a training set of 18.018 and a test set of 4.524 videos. Videos in the training set are performed by 5 unique native signers, while the test set has only one native signer whose videos are not used in training. In BosphorusSign22k, signs are performed in a controlled environment with a green background. All the signers are in a standing-up position wearing a dark-colored outfit.

AUTSL dataset [14, 28] provided by the ChaLearn Challenge [1] is used in this study. This dataset has

[1] https://chalearnlap.cvc.uab.es/challenge/43/description/

36.302 videos of 226 sign glosses performed by 43 unique signers divided into 31, 6, and 6 signers for training, validation, and test sets, respectively. In the AUTSL dataset, not all the signers are necessarily native signers. Therefore, some videos are performed by individuals who mimic signing the gloss in the video. Therefore, intra-class variations are expected to be higher than BosphorusSign22k. Moreover, the AUTSL dataset is collected rather in the wild. For this reason, the backgrounds of the videos differ. Another notable difference between the two datasets is that while some AUTSL signers stand up, some sit down while signing. Therefore, the position of hands relative to the torso might vary in the dataset. In AUTSL, significant similarities between the signing action of some glosses can be observed [14]. For this reason, top-3 and top-5 GCA are also reported along with top-1 GCA.

4.2. Implementation Details

PyTorch library is used to implement the neural network architectures and the loss functions used in this study. There are three types of 2D CNNs used in this thesis: a shallow encoder, ResNet-18, and ResNet-34. The shallow encoder consists of eight convolutional layers and two or three fully connected layers for projection. After each convolution layer, leaky ReLU with a leakiness of 0.2 is used. The eight convolutional layers start with convolution with a kernel of size 3, stride of 1 and padding of 1. The next convolution has a kernel size of 4, stride of 2 and padding of 1. These two layers are repeated four times with the same order to form the eight layers. The first of the eight layers raises the input channels of 3 to 8 and the next convolutional layers continue to increase the number of feature maps to 16, 24, 32, 40, 48, 56, and 64. The shallow encoder takes the output of the convolutional layers and maps to 512 dimensional representation with the fully connected layers of size 4096, 1024, and 512. ResNet-18 and ResNet-34 consist of 17 and 33 convolutional layers along with a fully connected layer that outputs a 512 dimensional representation, respectively.

All the LSTM models in the experiments follow the same one-layer LSTM architecture with a hidden layer dimension of 1024 and one fully connected output layer for gloss prediction. The output dimensionalities of the gloss classifiers are 744 and 226 for the BosphorusSign22k [13] and AUTSL [14] datasets, respectively. Signer classifiers, two and three-layer fully connected layers with Leaky ReLU activation, are added to the output of the CNN encoder and the LSTM network, respectively. The output layers of the signer classifiers have dimensionalities of 6 and 43 for the BosphorusSign22k [13] and AUTSL [14], respectively. Reversal layers, which behave as identity functions during forward pass and reverse the gradient's sign during backpropagation, are located before the signer classifiers. Hyperparameter α of the reversal layer seen in Eq. 4 changes between $(0, 1)$ during training by following Ganin and Lempitsky [12]. Adam and AdamW optimizers are used with a batch size of 16 and 32, respectively. For the experiments conducted with AUTSL, we used a cyclical learning rate policy with a base learning rate of 10^{-6} and a maximum learning rate of 10^{-4}. For the experiments conducted with BosphorusSign22k, we used a static learning rate of 5×10^{-5}.

4.3. Ablation Study and Recognition Performance

This section presents the recognition performance comparison for the proposed framework with various regularizers and baseline methods. Results are divided into the shallow encoder and ResNet-18 sections. When the shallow encoder described in Section 4.2 is used to learn the spatial embeddings, we could observe a more significant contribution of the proposed framework than the deeper encoder. However, the shallow encoder might be inadequate to attain state-of-the-art performances. Therefore, we also report the ablation study results for a deeper CNN encoder, ResNet-18.

Table 1: BosphorusSign22k dataset GCA (%) results for the shallow encoder. Baseline performances are given in the first four rows. The best performances are highlighted in bold.

Model	BosphorusSign22k		
	Top-1	**Top-3**	**Top-5**
Temporal Accumulative Features [29]	81.37	-	97.47
3D ResNets (MC3) [13]	78.85	-	94.76
IDT (HOG + HOF + MBH) [13]	88.53	-	-
Score-level Multi Cue Fusion (3D ResNets) [27]	94.94	-	99.76
Encoder + LSTM (Our Baseline)	75.00	89.10	94.09
Encoder + LSTM + MSE	**83.00**	91.68	94.34
Encoder + Reversal + LSTM	77.00	91.15	94.16
Encoder + Reversal + LSTM + MSE	81.00	**94.51**	**96.81**
Encoder + KL Divergence + Reversal + LSTM	78.00	92.00	95.00
Encoder + OT Distance + LSTM	80.00	92.00	95.00
Encoder + OT Distance + Reversal + LSTM	78.00	93.00	96.00

Table 2: AUTSL dataset GCA (%) results for the shallow encoder. Baseline performances are given in the first two rows. The best performances are highlighted in bold.

Model	AUTSL (Val)			AUTSL (Test)		
	Top-1	**Top-3**	**Top-5**	**Top-1**	**Top-3**	**Top-5**
AUTSL Baseline[14] (CNN+FPM+BLSTM+Attention)	-	-	-	49.22	68.89	75.78
SAM-SLR (GCN based)[10]	-	-	-	98.42	-	-
Encoder + LSTM (Our Baseline)	59.00	77.29	83.20	56.00	73.30	80.86
Encoder + LSTM + MSE	63.00	78.42	84.65	60.00	77.60	83.45
Encoder + Reversal + LSTM	65.00	81.91	86.93	62.00	80.54	86.45
Encoder + Reversal + LSTM + MSE	67.56	84.54	89.04	57.93	76.5	82.09
Encoder + KL Divergence + Reversal + LSTM	**72.00**	**89.00**	**94.00**	62.00	80.00	88.00
Encoder + OT Distance + LSTM	65.00	80.00	84.00	62.00	79.00	87.00
Encoder + OT Distance + Reversal + LSTM	71.00	86.00	90.00	**64.00**	**81.00**	**87.00**

We report various proposed approaches in addition to baselines. Encoder + LSTM denotes the base model where adversarial training or regularizations are not applied. Encoder + LSTM + MSE is the model trained with only MSE loss regularizer penalizing the differences in temporal representations of the same gloss performed by different signers. Models with Encoder + Reversal + LSTM have the gradient reversal layer added to the end of the spatial Encoder adversarially trained with a signer classifier. The models with KL Divergence and OT Distance have the corresponding regularizers added to their loss functions.

4.3.1. Shallow Encoder

Shallow encoder results of BosphorusSign22k and AUTSL datasets are given in Tables 1 and 2, respectively. The tables report baseline [10, 13, 14, 27, 29] and proposed framework recognition performances as in GCA. For the BosphorusSign22k dataset, MSE loss regularization, denoted by Encoder + LSTM + MSE loss, performs best compared to the Encoder + LSTM model and some of the baselines from the literature. When we add the

Table 3: BosphorusSign22k GCA (%) results for ResNet-18. The top four rows are the benchmark results for the BosphorusSign22k dataset. The best performances among our methods are denoted in bold.

Model	BosphorusSign22k		
	Top-1	**Top-3**	**Top-5**
Temporal Accumulative Features [29]	81.37	-	97.47
3D ResNets (MC3) [13]	78.85	-	94.76
IDT (HOG + HOF + MBH) [13]	88.53	-	-
Score-level Multi Cue Fusion (3D ResNets) [27]	94.94	-	99.76
ResNet-18 + LSTM (Our Baseline)	90.00	95.00	96.00
ResNet-18 + LSTM + MSE	92.00	97.00	**99.00**
ResNet-18' + LSTM + MSE	**94.00**	**98.00**	**99.00**
ResNet-18 + Reversal + LSTM	88.00	97.00	98.00
ResNet-18 + OT Distance + LSTM	88.00	96.00	97.00

Table 4: AUTSL GCA (%) for ResNet-34. The best performances among our methods are denoted in bold.

Model	AUTSL (Val)			AUTSL (Test)		
	Top-1	**Top-3**	**Top-5**	**Top-1**	**Top-3**	**Top-5**
AUTSL Baseline [14] (CNN+FPM+BLSTM+Attention)	-	-	-	49.22	68.89	75.78
SAM-SLR (GCN based) [10]	-	-	-	98.42	-	-
ResNet-34 + LSTM (Our Baseline)	89.00	95.00	97.00	90.00	96.00	98.00
ResNet-34 + OT Distance + LSTM	**89.00**	**96.00**	**97.00**	89.00	**96.00**	**98.00**
ResNet-34 + KL Divergence + Reversal + LSTM	87.00	95.00	97.00	86.00	95.00	97.00
ResNet-34 + LSTM + MSE	89.00	95.00	97.00	88.00	96.00	97.00
ResNet-34 + KL Divergence	80.00	92.00	95.00	76.00	91.00	94.00

reversal layer with the MSE regularizer, top-3 and top-5 accuracies improve further. We can also see that the shallow Encoder cannot reach the state-of-the-art set by the 3D CNN models [27]. According to the ablation study, although the adversarial training with gradient reversal and signer classifier might not improve the top-1 accuracy in every case, it can potentially boost top-3 and top-5 accuracies.

For AUTSL experiments, Table 2 shows that adversarial training with gradient reversal, KL divergence, and OT distance regularizers significantly impact recognition performance. KL divergence regularizer performs the best in the validation set, while the OT distance regularizer yields the best performance in the test set. We also provide the state-of-the-art performance for AUTSL denoted by SAM-SLR [10], which is based on Graph Convolutional Networks (GCNs) on Table 2. The significant difference in the recognition performances of AUTSL and BosphorusSign22k stems from the fundamental challenges AUTSL constitutes as described in Section 4.1. The impact of adversarial training with the signer classifier and regularizers on the signer independence varies depending on the number of signers and glosses and the data collection environment.

4.3.2. Deep Encoder

ResNet results of BosphorusSign22k and AUTSL datasets are given in Tables 3 and 4, respectively. ResNet-18 is used in BosphorusSign22k experiments. In AUTSL experiments, we observe better results with ResNet-34 than with ResNet-18. The tables report the performance of the baseline and the best-performing proposed methods

Table 5: Evaluation of the best-performing models with more metrics. All the metrics are macro averages over the number of classes.

Dataset	Model	Validation Dataset				Test Dataset			
		GCA (%)	Precision (%)	Recall (%)	F1-Score (%)	GCA (%)	Precision (%)	Recall (%)	F1-Score (%)
BosphorusSign22k	Base (Shallow Encoder + LSTM)	-	-	-	-	0.76	0.78	0.75	0.73
	Shallow Encoder + Reversal + LSTM + MSE	-	-	-	-	0.82	0.83	0.82	0.80
	Base (Resnet18 + LSTM)	-	-	-	-	0.85	0.87	0.86	0.83
	Resnet18 + LSTM + MSE	-	-	-	-	0.90	0.89	0.89	0.87
AUTSL	Base (Shallow Encoder + LSTM)	0.58	0.63	0.58	0.58	0.54	0.58	0.54	0.53
	Shallow Encoder + Reversal + LSTM + KL Divergence	0.71	0.73	0.71	0.71	0.57	0.62	0.57	0.57
	Base (Resnet34 + LSTM)	0.89	0.90	0.89	0.89	0.90	0.90	0.90	0.89
	Resnet34 + LSTM + MSE	0.89	0.89	0.88	0.88	0.88	0.89	0.88	0.88

as in GCA. When deep encoders are used, we obtain similar behaviors to shallow encoder experiments. MSE and OT distance regularizers contribute more to the recognition performance than the other proposed methods for BosphorusSign22k and AUTSL datasets, respectively. We can achieve a state-of-the-art performance with 2D ResNet-18 and MSE regularizer for the BosphorusSign22k. On the other hand, performance improvement over the ResNet-34 + LSTM baseline is subtle for the AUTSL dataset. We hypothesize that support of adversarial training and regularizers might diminish with deeper spatial encoders since gradients during deep network training are prone to vanish. Recognition performance with additional metrics, Precision, Recall, and F1-Score, are also reported for selected models in Table 5.

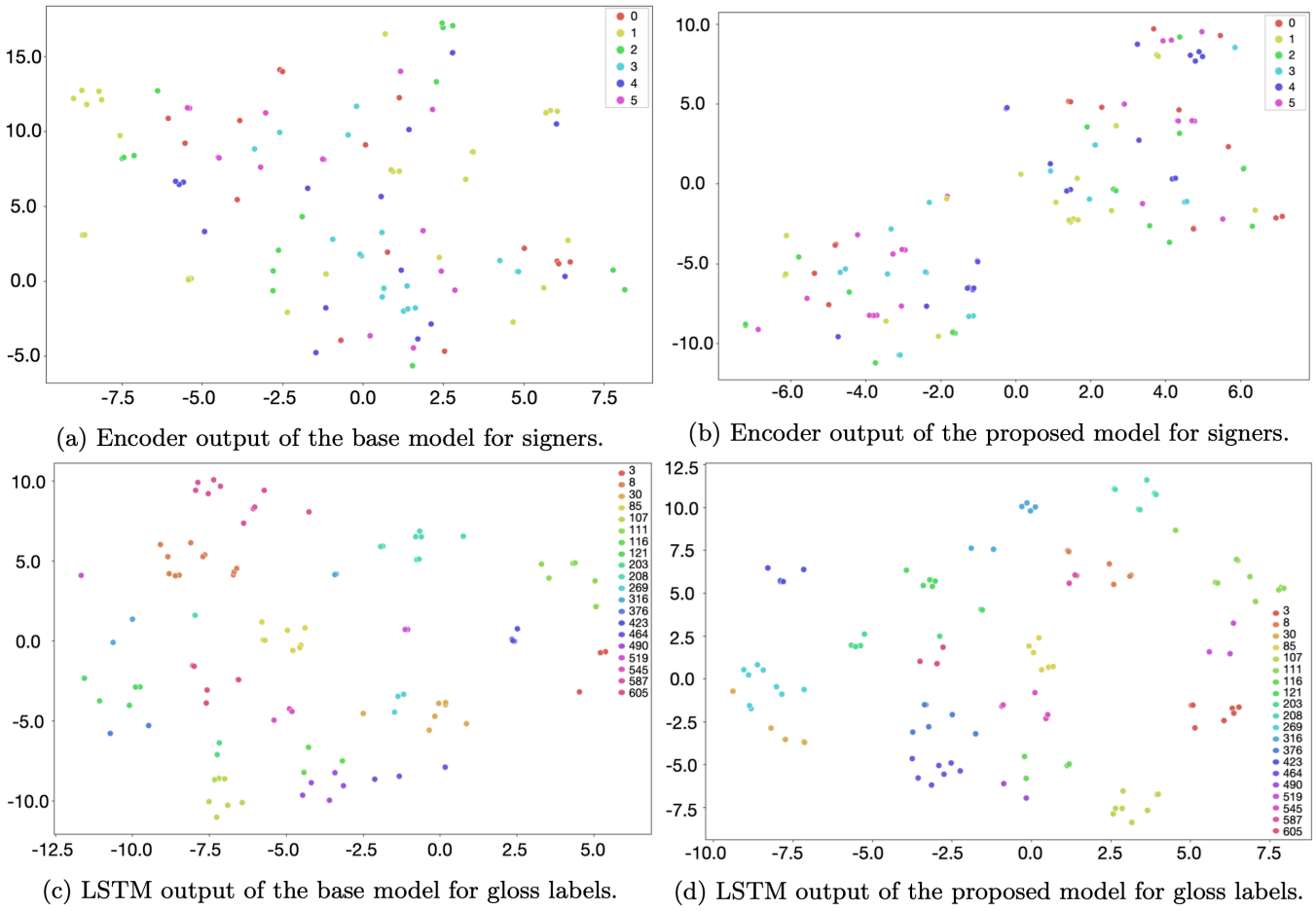
4.3.3. Qualitative Analysis

To investigate the effect of disentanglement on the learned representations, we illustrate the t-SNE plots of the spatial and temporal embeddings in Figures 3 and 4 for AUTSL and BosphorusSign22k datasets, respectively. The expected behavior is that groupings in the latent space should be due to the signs rather than the signers. Therefore, clusters around signers should be dissolved after the feature disentanglement. The plots in Figures 3 and 4 show that the proposed feature disentanglement can improve the gloss groupings compared to the Encoder + LSTM baseline. This conclusion indicates that the feature disentanglement can impose signer independence on the spatial and temporal embeddings.

In addition to the t-SNE plots, we quantify the distances and similarities between the embeddings with and without the proposed feature disentanglement. For this purpose, Euclidean distance and cosine similarity between the temporal embeddings based on signer and gloss groups are reported in Figure 5. Here, we expect to observe that the similarity between the embeddings of different signs performed by the same signer should decrease and that the same sign performed by different signers should increase after the feature disentanglement. The plots in Figure 5 demonstrate that regularized adversarial training can facilitate signer independence in the latent space of CNN-based SLR frameworks.

5. Conclusion

CNN-based SLR frameworks are prone to overfitting signer-specific features. The SLR aims to learn distinctive representations for different signs rather than the signer's identity. This study proposes an adversarial training-based feature disentanglement approach for signer-independent representation learning. Notably, a CNN encoder and LSTM network are trained with a signer classifier and various regularizers, such as KL divergence, OT distance, and MSE loss. The proposed training scheme is evaluated on two isolated Turkish sign language datasets, BosphorusSign22k and AUTSL. Quantitative and qualitative results show that the proposed approach facilitates signer independence. Ablation study shows that MSE loss regularizers can be effective in



(a) Encoder output of the base model for signers.

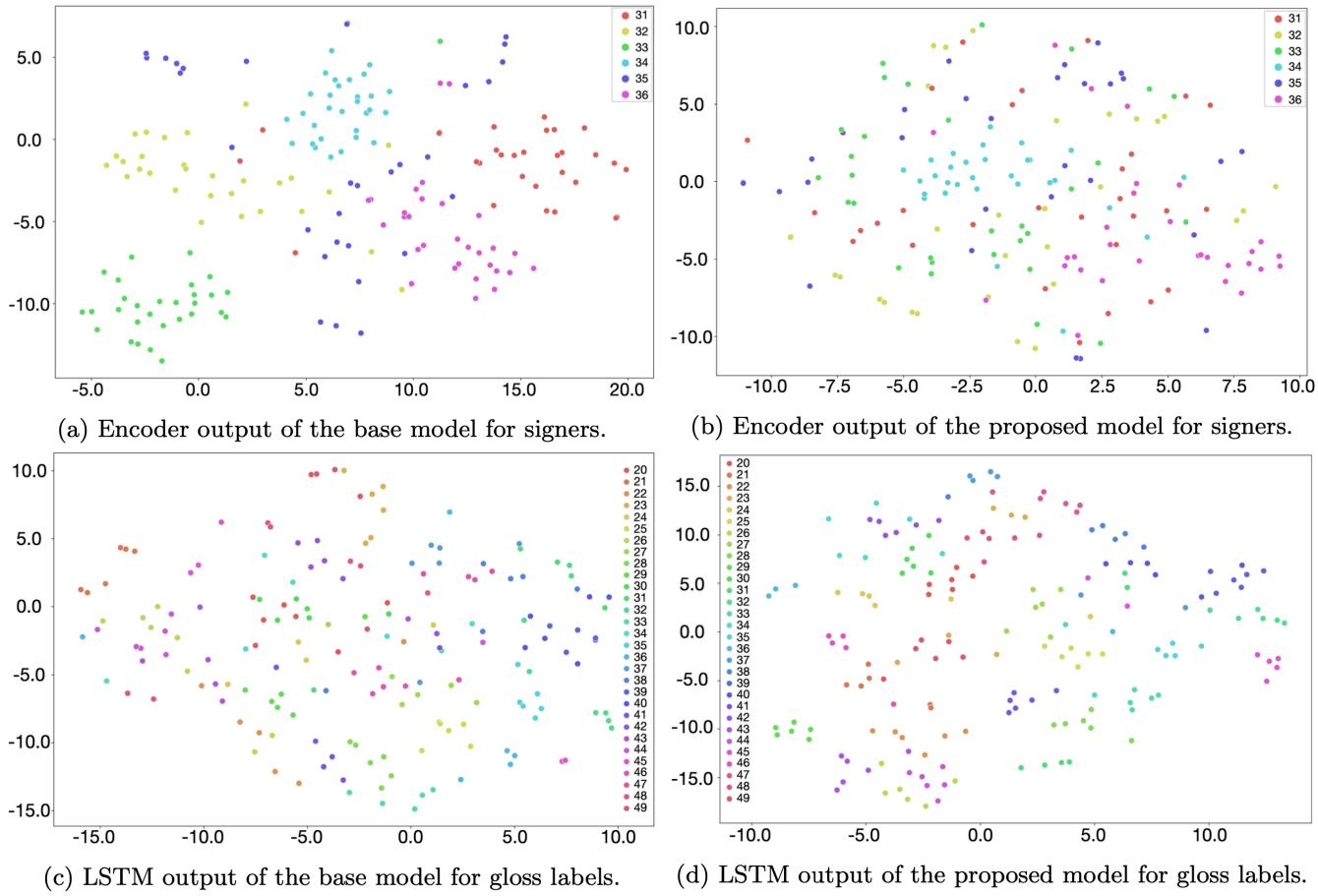
(b) Encoder output of the proposed model for signers.

(c) LSTM output of the base model for gloss labels.

(d) LSTM output of the proposed model for gloss labels.

Figure 3: t-SNE of the Encoder and LSTM outputs on the BosphorusSign22k dataset. Colors denote six signer classes and 20 gloss classes in Figures 3a, 3b, and Figures 3c, 3d, respectively. The base model is Encoder + LSTM, and the proposed model is chosen as the best-performing regularized model for the dataset.

BosphorusSign22k, while OT distance regularization might boost top-3 and top-5 recognition performance in AUTSL. The qualitative analysis demonstrates that the proposed feature disentanglement mitigates the variance in sign embeddings due to signer differences. On the other hand, when deeper CNN encoders are used to extract spatial embeddings, the impact of the feature disentanglement diminishes compared to the state-of-the-art performances. Therefore, the recognition performance of the proposed framework when deep spatial encoders are used could be considered a limitation. In addition, the proposed framework requires RGB input which requires high model complexity and memory during training. For this reason, the batch size might have to be chosen small due to memory limitations causing a slower convergence. In future work, we will consider generative models for feature disentanglement. GANs and VAEs can often be encountered for disentangled representation learning for images in literature. 3D GAN and VAE architectures can be analyzed to disentangle spatiotemporal and signer-related features to achieve signer independence in the SLR applications. On the other hand, skeleton-based SLR frameworks are becoming ubiquitous, where skeleton joints with GCNs are used to learn a sign representation. Compared to RGB-based SLR frameworks, using skeleton joints might



(a) Encoder output of the base model for signers.

(b) Encoder output of the proposed model for signers.

(c) LSTM output of the base model for gloss labels.

(d) LSTM output of the proposed model for gloss labels.

Figure 4: t-SNE of the Encoder and LSTM outputs on the AUTSL dataset. Colors denote six signer classes and 30 gloss classes in Figures 4a, 4b, and Figures 4c, 4d, respectively. t-SNE demonstrates that the performance improvement is due to feature disentanglement, as previously hypothesized. The base model is Encoder + LSTM, and the proposed model is chosen as the best-performing regularized model for the dataset.

alleviate the influence of the signer's appearance on the learned sign representation. However, in future work, we will investigate the possible signer dependency in skeleton-based SLR models. Finally, signer-independent SLR should also be studied for continuous sign language videos, where more than one sign is performed in one video. The continuous SLR requiring recognition and segmentation of multiple glosses in a video poses various challenges, including the recurring challenge of singer dependency.

Acknowledgment

This study was supported by Bogazici University Research Fund under grant number 17004. The numerical calculations reported in this study were fully/partially performed at TUBITAK ULAKBIM, and High Performance and Grid Computing Center (TRUBA resources). **İ.E did the experiments, wrote the experiments section, İ.M.B gave the idea, wrote and edited the paper. Both İ.E. and İ.M.B interpreted the results.**

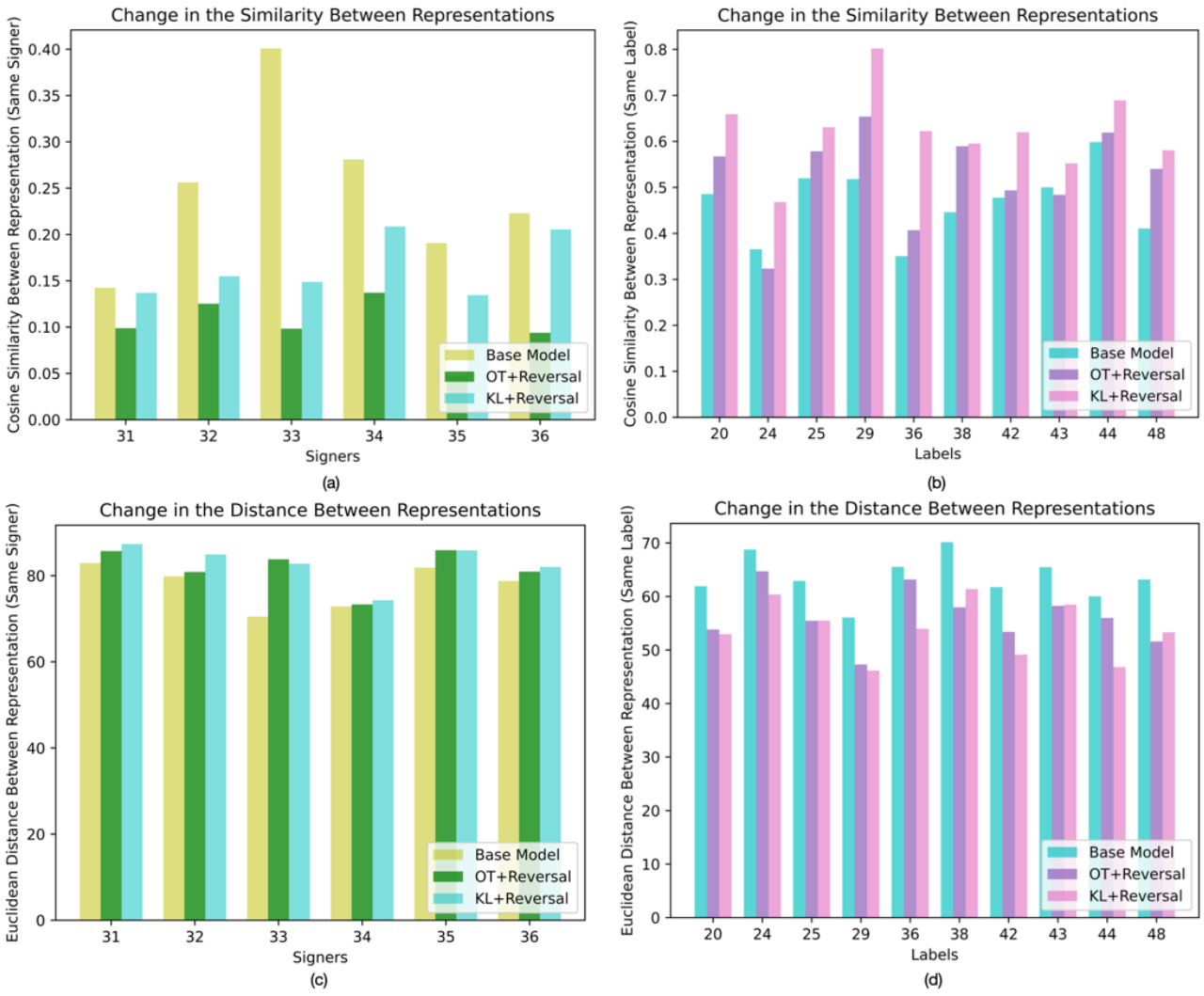


Figure 5: Average cosine similarity between the AUTSL gloss representations based on a) signer and b) label groupings. Euclidean distance between the AUTSL gloss representations based on c) signer d) label groupings.

References

[1] Bragg D, Koller O, Bellard M, Berke L, Boudreault P et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. In: International ACM SIGACCESS Conference on Computers and Accessibility; Pittsburgh, PA, USA; 2019. pp. 16-31.

[2] Süzgün M, Özdemir H, Camgöz NC, Kındıroğlu AA, Başaran D et al. Hospisign: an interactive sign language platform for hearing impaired. Journal of Naval Sciences and Engineering 2015; 11 (3): 75-92.

[3] Kim J, Hwang EJ, Cho S, Lee DH, Park Jong C. Sign language production with avatar layering: A critical use case over rare words. In: Conference on Language Resources and Evaluation; Marseille, France; 2022. pp. 1519-1528.

[4] Tornay S, Razavi M, Doss MM. Towards multilingual sign language recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing; Barcelona, Spain; 2020. pp. 6309-6313.

[5] Orbay A, Akarun L. Neural sign language translation by learning tokenization. In: IEEE International Conference on Automatic Face and Gesture Recognition; Buenos Aires, Argentina; 2020. pp. 222-228.

[6] Camgöz NC, Koller O, Hadfield S, Richard B. Multi-channel transformers for multi-articulatory sign language translation. In: European Conference on Computer Vision; Virtual; 2020. pp. 301-319.

[7] Saunders B, Camgöz NC, Bowden R. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. International Journal of Computer Vision 2021; 129 (7): 2113-2135.

[8] Vázquez-Enríquez M, Alba-Castro JL, Docío-Fernández L, Rodríguez-Banga E. Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; Nashville, TN, USA; 2021. pp. 3457-3466.

[9] Özdemir O, Baytaş İM, Akarun L. Multi-cue temporal modeling for skeleton-based sign language recognition. Frontiers in Neuroscience 2023; 17: 1148191. https://doi.org/10.3389/fnins.2023.1148191

[10] Jiang S, Sun B, Wang L, Bai Y, Li K, Fu YR. Skeleton aware multi-modal sign language recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; Nashville, TN, USA; 2021. pp. 3408-3418.

[11] Si C, Jing Y, Wang W, Wang L, Tan T. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: European Conference on Computer Vision; Munich, Germany; 2018. pp. 103-118.

[12] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning; Lille, France; 2015. pp. 1180-1189.

[13] Özdemir O, Kındıroğlu AA, Camgöz NC, Akarun L. BosphorusSign22k sign language recognition dataset. In: The 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives; Marseille, France; 2020. pp. 181-188.

[14] Sincan OM, Keleş HY. AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods. IEEE Access 2020; 8: 181340-181355. https://doi.org/10.1109/ACCESS.2020.3028072

[15] Tran L, Yin X, Liu X. Disentangled representation learning GAN for pose-invariant face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Honolulu, HI, USA; 2017. pp. 1283-1292.

[16] Liu Y, Wang Z, Jin H, Wassell I. Multi-task adversarial network for disentangled feature learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Salt Lake City, UT, USA; 2018. pp. 3743-3751.

[17] Oldfield J, Panagakis Y, Nicolaou MA. Adversarial learning of disentangled and generalizable representations of visual attributes. IEEE Transactions on Neural Networks and Learning Systems 2019; 33: 3498-3509. https://doi.org/10.1109/TNNLS.2021.3053205

[18] Zhang Z, Tran L, Yin X, Atoum Y, Liu X et al. Gait recognition via disentangled representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Long Beach, CA, USA; 2019. pp. 4710-4719.

[19] Li Y, Luo Z, Weng X, Kitani K. Learning shape representations for clothing variations in person re-identification. In: IEEE Winter Conference on Applications of Computer Vision; Waikoloa, HI, USA; 2021. pp. 2431-2440.

[20] Yang L, Yao A. Disentangling latent hands for image synthesis and pose estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Long Beach, CA, USA; 2019. pp. 9877-9886.

[21] Ferreira PM, Pernes D, Rebelo A, Cardoso JS. Signer-independent sign language recognition with adversarial neural networks. International Journal of Machine Learning and Computing 2021; 11: 121-129.

[22] Zhang Q, Wang S, Chen G. Speaker-independent lipreading by disentangled representation learning. In: IEEE International Conference on Image Processing; Anchorage, AK, USA; 2021. pp. 2493-2497.

[23] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computing 1997; 9: 1735-1780.

[24] Zhang H, Wang J. Defense against adversarial attacks using feature scattering-based adversarial training. In: Advances in Neural Information Processing Systems; Vancouver, Canada; 2019.

[25] Salimans T, Zhang H, Radford A, Metaxas D. Improving GANs using optimal transport. arXiv [cs.LG] preprint 2017; 1803.05573. https://doi.org/10.48550/arXiv.1803.05573

[26] Genevay A, Peyré G, Cuturi M. GAN and VAE from an optimal transport point of view. arXiv [stat.ML] preprint 2017; 1706.01807. https://doi.org/10.48550/arXiv.1706.01807

[27] Gökçe Ç, Özdemir O, Kındıroğlu AA, Akarun L. Score-level multi cue fusion for sign language recognition. In: European Conference on Computer Vision; Glasgow, UK; 2020. pp. 294-309.

[28] Mercanoğlu O, Jacques J, Escalera S, Keleş H. ChaLearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; Nashville, TN, USA; 2021. pp. 3467-3476.

[29] Kındıroğlu AA, Özdemir O, Akarun L. Temporal accumulative features for sign language recognition. In: IEEE/CVF International Conference on Computer Vision Workshop; Seoul, South Korea; 2019. pp. 1288-1297.