DISENTANGLED REPRESENTATION LEARNING IN ISOLATED SIGN LANGUAGE RECOGNITION

by

İpek Erdoğan

B.S., Computer Engineering, Istanbul Technical University, 2020

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2023

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude and appreciation to my advisor Assist. Prof. İnci Meliha Baytaş, for being a tremendous mentor for me. Thank you for encouraging my research and allowing me to grow as a research scientist. It was a real privilege to work under your guidance.

I am deeply indebted to my jury committee, including Prof. Lale Akarun and Prof. Hatice Köse, not only for their time and detailed advice but for their intellectual contributions to this thesis.

I would also like to extend my deepest gratitude to all the Computer Engineering Department faculty members, not only for providing a splendid learning experience but for teaching how a scientist should behave.

To all the talented researchers in my labs, thank you for the fun and support. Learning from each other's experiences was invaluable.

I would like to thank my friends, Umay, Batuhan, and Atakan, who always supported and understood me. There were times I felt demotivated and tired, but with you in my life, it was always much easier.

I sincerely appreciate Ulaş for all the love, patience, unfailing support, and continuous encouragement throughout my years of study.

Finally, but not least, I am extremely grateful to my mother and sister, my biggest chances in this life, who have always believed in me and been there for me.

ABSTRACT

DISENTANGLED REPRESENTATION LEARNING IN ISOLATED SIGN LANGUAGE RECOGNITION

Representation learning is an essential part of all deep learning tasks. Achieving good performance in recognition, generation, and classification heavily depends on learning meaningful and reliable representations. It is important to gather informative representations that are not affected by unnecessary details in all cases. Sign Language Recognition is one of the areas where deep learning models have been successfully used. Sign Language Recognition (SLR) is essential to exchange information between those who know sign language and those who do not.

The input of an SLR model is a video in which an individual performs a sign or multiple signs. Therefore, Convolutional Neural Networks (CNN) are commonly a part of deep learning-based SLR frameworks. However, CNN-based recognition frameworks tend to capture the characteristics of the identity in the foreground, such as face attributes, hand and body shape, and skin color. This challenge is often encountered in problems such as face and gait recognition, image manipulation, and person re-identification problems.

In this thesis, a disentangled representation learning framework is proposed to separate the latent factors in the sign and signer representations and eliminate the irrelevant identity information to improve sign recognition performance. Various disentanglement techniques, including regularized adversarial training, are investigated. Experiments are conducted on two isolated Turkish sign language benchmark datasets. The effect of feature disentanglement and its potential to improve recognition performance are discussed with qualitative and quantitative analysis.

ÖZET

İZOLE İŞARET DİLİ TANIMADA AYRIŞTIRILMIŞ TEMSİL ÖĞRENİMİ

Temsil öğrenimi, derin öğrenme görevlerinin önemli bir parçasıdır. Tanıma, oluşturma ve sınıflandırma modellerinde iyi performans elde etmek, büyük ölçüde anlamlı ve güvenilir temsilleri öğrenmeye bağlıdır. Gereksiz ayrıntılardan etkilenmeyen bilgilendirici temsiller toplamak, her koşulda önemlidir. İşaret Dili Tanıma (İDT), derin öğrenme modellerinin başarıyla kullanıldığı alanlardan biridir.

İşaret dili, işitme ve konuşma engelli insanlarla iletişimde kullanılan birincil iletişim aracıdır. İşaret dilini bilenler ve bilmeyenler arasında bilgi alışverişini sağlamak için kullanılan işaret dili tanıma modelleri vardır. İşaret dili tanıma modellerinin girdileri videolar olduğundan, Evrişimsel Sinir Ağları, tanıma modellerinin temel parçalarından biridir. Bununla birlikte, Evrişimsel Sinir Ağları tabanlı tanıma modelleri, yüz özellikleri, el ve vücut şekli ve ten rengi gibi kimlik ilintili özellikleri yakalama ve kodlama eğilimindedir. Bu, yüz tanıma modelleri, yürüyüş tanıma modelleri, görüntü manipülasyonu modelleri veya kişi yeniden tanımlama modelleri gibi görsel derin öğrenme modellerinde yaygın rastlanan bir sorundur.

Bu tezde, İDT modellerinin işaretçi özelliklerinden etkilenmemesi için işaret ve işaretçi temsillerindeki gizli faktörleri ayırmak ve işaretçi özelliklerinden kaynaklanan ilgisiz bilgilerini ortadan kaldırmak için işaretçi bilgisinden ayıklanmış bir temsil öğrenme yöntemi önerilmiştir. Düzenlileştirilmiş hasmane eğitim de dahil olmak üzere çeşitli ayıklanmış öğrenme teknikleri araştırılmıştır. Deneyler, izole edilmiş iki Türk işaret dili veri kümesi üzerinde yapılmıştır. Özellik ayıklamanın etkisi ve tanıma performansını iyileştirme potansiyeli, niteliksel ve niceliksel analizlerle tartışılmıştır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . viii
LIST OF TABLES . . . x
LIST OF SYMBOLS . . . xii
LIST OF ACRONYMS/ABBREVIATIONS . . . xiii
1. INTRODUCTION . . . 1
1.1. Thesis Focus and Contributions . . . 2
2. RELATED WORK . . . 4
2.1. Disentangled Representation Learning . . . 4
2.2. Isolated Sign Language Recognition . . . 8
2.2.1. 2D-CNN Based ISLR Models . . . 9
2.2.2. 3D-CNN Based ISLR Models . . . 10
2.2.3. GCN Based ISLR Models . . . 11
3. METHOD . . . 14
3.1. Adversarial Training . . . 14
3.2. Regularization Techniques . . . 15
3.3. Proposed Method . . . 17
3.3.1. Backbone Architectures . . . 18
3.3.1.1. 2D CNN + LSTM . . . 18
3.3.1.2. 3D CNN . . . 20
3.3.2. Optimization and Training Scheme . . . 20
4. EXPERIMENTAL RESULTS . . . 23
4.1. Datasets . . . 23
4.2. Implementation . . . 24
4.2.1. Evaluation Metrics . . . 26
4.3. Results . . . 26

4.3.1. Ablation Study and Quantitative Results 27
4.3.1.1. Experiments With the Shallow Encoder 27
4.3.1.2. Experiments with the Resnet3D architecture 29
4.3.1.3. Experiments with the ResNet2D architectures 30
4.3.1.4. Experiments with Cross Datasets 31
4.3.2. Qualitative Analysis of Feature Disentanglement 32
5. CONCLUSION 39
REFERENCES 41

LIST OF FIGURES

Figure 3.1. The proposed model with feature disentanglement. The gradient reversal layer is added to the output of the encoder layer before the signer classifier, f_s. OT loss ($\mathcal{L}_{\text{OT}}$) or KL divergence ($\mathcal{L}_{\text{KL}}$) is calculated using signer representations, yet MSE loss ($\mathcal{L}_{\text{MSE}}$) is calculated using gloss representations. 18

Figure 4.1. The first and the second rows demonstrates samples from BosphorusSign22k [1] and AUTSL [2] datasets, respectively. Images in the last row are the samples after the pre-processing step. 23

Figure 4.2. Cosine similarity between the gloss representations based on a) signer b) label groupings and Euclidean distance between the gloss representations based on c) signer d) label groupings. For the both cases, Gloss representations belongs to AUTSL dataset. Proposed methods here are Experiment 6 and 7 in the Table 4.2. 34

Figure 4.3. t-SNE of the Encoder outputs of a) the base model b) the proposed model on AUTSL dataset. Colors denote 6 signer classes in the Figures 4.3(a), 4.3(b). t-SNE demonstrates that the performance improvement is due to feature disentanglement, as previously hypothesized. Proposed model here is the Experiment 6 in the Table 4.2. 35

Figure 4.4. t-SNE of the LSTM outputs of a) the base model b) the proposed model on AUTSL dataset. Colors denote 30 gloss classes in the Figures 4.4(a), 4.4(b). t-SNE demonstrates that the performance improvement is due to feature disentanglement. Proposed model here is the Experiment 6 in the Table 4.2. 36

Figure 4.5. t-SNE of the Encoder outputs of a) the base model b) the proposed model on AUTSL dataset. Colors denote 6 signer classes in the Figures 4.5(a), 4.5(b). t-SNE demonstrates that the performance improvement is due to feature disentanglement, as previously hypothesized. Proposed model here is the Experiment 4 in the Table 4.2. 37

Figure 4.6. t-SNE of the LSTM outputs of a) the base model b) the proposed model on AUTSL dataset. Colors denote 30 gloss classes in the Figures 4.6(a), 4.6(b). t-SNE demonstrates that the performance improvement is due to feature disentanglement. Proposed model here is the Experiment 4 in the Table 4.2. 38

LIST OF TABLES

Table 4.1. Hyperparameters used in the experiments. 26

Table 4.2. Implemented experiments and their components. 27

Table 4.3. Gloss Classification Accuracy (%) results for all models implemented with the shallow encoder, tested on the AUTSL dataset. The top two rows are the benchmark results for the AUTSL dataset. The best performances among our methods are denoted in bold. . 28

Table 4.4. Gloss Classification Accuracy (%) results for all models implemented with the shallow encoder, tested on the BosphorusSign22k dataset. The top four rows are the benchmark results for the BosphorusSign22k dataset. The best performances among our methods are denoted in bold. 29

Table 4.5. Gloss Classification Accuracy (%) results for all models implemented with the ResNet3D, tested on the AUTSL dataset. The top two rows are the benchmark results for the AUTSL dataset. The best performances among our methods are denoted in bold. . 30

Table 4.6. Gloss Classification Accuracy (%) results for all models implemented with the ResNet3D, tested on the BosphorusSign22k dataset. The top four rows are the benchmark results for the BosphorusSign22k dataset. The best performances among our methods are denoted in bold. 31

Table 4.7. Gloss Classification Accuracy (%) results for all models implemented with the ResNet2D-34, tested on the AUTSL dataset. The top two rows are the benchmark results for the AUTSL dataset. The best performances among our methods are denoted in bold. . 32

Table 4.8. Gloss Classification Accuracy (%) results for all models implemented with the ResNet2D-18, tested on the BosphorusSign22k dataset. The top four rows are the benchmark results for the BosphorusSign22k dataset. The best performances among our methods are denoted in bold. 33

LIST OF SYMBOLS

$c(\mathbf{s}_i, \mathbf{s}_j)$	Cost function
C_g	Number of unique glosses
C_s	Number of unique signers
$\mathbf{e}_i$	Signer embedding
$E(\cdot)$	Convolutional encoder
f_g	Gloss classifier
f_p	Fully-connected layers to project the embeddings into the gloss space
f_s	Signer classifier
$\mathbf{h}_i^+$	Hidden representation of the LSTM unit's last time step for the ith video
$\mathcal{L}_s$	Signer classification loss
$\mathcal{L}_{\text{OT}}$	OT loss
$\mathcal{L}_{\text{KL}}$	KL divergence loss
$\mathcal{L}_{\text{MSE}}$	MSE loss
N	Batch size
$P(i)$	Uniform distribution
$Q(i)$	Signer embedding
T	Length of the input sequence
$\mathbf{x}_i$	ith frame of the input video
y	Ground truth label for gloss classification
y^s	Ground truth label for signer classification
α	Gradient reversal hyperparameter
θ_E	Parameter of convolutional encoder
γ	Joint distribution
$\Pi\left(\boldsymbol{\mu}_{s_i}, \boldsymbol{\mu}_{s_j}\right)$	Set of joint distributions

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
CNN	Convolutional Neural Network
CSLR	Continuous Sign Language Recognition
GAN	Generative Adversarial Network
ISLR	Isolated Sign Language Recognition
KL	Kullback-Leibler
LSTM	Long Short Term Memory
MSE	Mean Squared Error
OT	Optimal Transport
SLR	Sign Language Recognition
VAE	Variational Autoencoder

1. INTRODUCTION

Representation learning plays an essential role in a wide range of applications, particularly the ones concerned with vision problems. Whether the main focus is recognition, generation, or classification, learning distinctive and reliable representations of the input data is the key to providing compelling model performances. A good representation has several qualities. One of those qualities is the invariance against factors that inject unwanted components into the learned representation. For instance, gait recognition aims to learn human gait characteristics and capture a pattern that can discriminate the gaits of different subjects. On the other hand, factors such as clothing and the view angle influence the representations learned using traditional deep learning frameworks degrading the recognition performance [3]. Face recognition models target to recognize human faces correctly, without getting affected by the age or the pose variety in the input images [4, 5]. In image manipulation and person re-identification models, it is essential to separate the patterns (e.g., different face attributes) in the input representation based on relevant context [6, 7].

Sign Language Recognition (SLR) is also one of the problems where learning a distinctive sign representation is essential rather than the identity and pose of the signers or the details in the background. Sign Language (SL), a combination of hand gestures and facial expressions, is the primary communication tool for deaf and speech-impaired communities [8, 9]. There are different sign languages (e.g., Turkish SL, American SL, British SL, Chinese SL) where different parts of the body (e.g., hands, head, face, arms, torso) contribute to the visual characteristics of their signs [10, 11]. SLR is central in various applications that aim to clear away the communication barrier between the deaf and hearing communities [10]. The input of an SLR framework is a video where an individual, a signer, performs a sign or multiple signs in the foreground. These individuals could be native signers from the deaf community or someone who knows how to sign from the hearing community.

The SLR frameworks aim to learn spatial and temporal sign representations that facilitate discriminating the signs. Therefore, such frameworks are not concerned with learning a representation of the signer's identity or appearance. However, it has been observed that representations learned by Convolutional Neural Network (CNN)-based recognition frameworks tend to be signer dependent and capture the characteristics, such as facial attributes, body shape, and skin color of the signers. In sign language datasets collected in the wild, the invariance becomes even more crucial due to the variety in signer-identity attributes [12] and different backgrounds, which can be misleading for the sign classifier if encoded into the representation. For this reason, it is imperative to address signer dependency in CNN-based SLR frameworks.

One way to achieve signer dependency in CNN-based SLR frameworks is Disentangled Representation Learning (DRL). This technique aims to separate latent factors and eliminate the irrelevant details in the representation learned for the task of interest. Thus, DRL provides a better understanding of the underlying structures in data. The feature disentanglement can be achieved with adversarial training, Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and regularization techniques that focus on reducing the effect of irrelevant details in the learned representations. This thesis proposes a disentangled representation learning framework for Isolated Sign Language Recognition (ISLR) where the input video contains a single sign. In the next section, we detail the thesis focus and contributions.

1.1. Thesis Focus and Contributions

An ISLR framework aims to predict the gloss, i.e., a word associated with a sign, signed in the video. Consequently, the ISLR task can be posed as a gloss classification problem, where learning a distinctive gloss representation is the concern. In this problem, videos of individuals performing face and body gestures are used to elicit the gloss representation. For this reason, ISLR frameworks are prone to be vulnerable to irrelevant patterns stemming from the individuals in the video's foreground. This challenge makes disentangled representation learning worth exploring in the SLR do-

main. Therefore, this thesis investigates the feature disentanglement techniques and proposes a framework tailored for the ISLR task. The contributions of the thesis are summarized below.

- We implemented an SLR model consisting of a CNN encoder to extract spatial representation and a Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) for temporal modeling. The effect of the feature disentanglement is investigated for a shallow and a deeper off-the-shelf CNN encoder. Furthermore, the feature disentanglement is also explored with 3-dimensional CNN, where spatial and temporal modeling is done within one module.
- We first investigate a feature disentanglement by adversarial training with gradient reversal [13]. The effect of disentanglement is analyzed for employing gradient reversal with both spatial and temporal representation learning modules.
- After we observe that solely adversarial training is insufficient, regularized disentangled representation learning is considered. For regularization purposes, we investigate various loss functions, such as Kullback-Leibler (KL) divergence [14], Optimal Transport (OT) distance [15], and Mean Squared Error (MSE). We conduct ablation studies to interpret the contributions of the regularization terms to the feature disentanglement.
- Various feature disentanglement techniques are implemented and experimented with for two Turkish ISLR datasets, BosphorusSign22k [1] and AUTSL [2, 16]. Thus, one of the contributions of the thesis is a benchmarking study of the disentangled representation learning for well-known Turkish ISLR datasets.

2. RELATED WORK

2.1. Disentangled Representation Learning

During the feature extraction phase, machine learning models struggle with separating redundant and important parts of the information they encode in the representations. This struggle may result in architecture performing worse than it can with a more informative representation. The redundant parts of the representations are not just irrelevant but may also be deceptive.

CNN-based recognition frameworks are expected to learn features encoding essential information that facilitate the recognition task. On the other hand, CNNs might also capture redundant information that might degrade the recognition performance in some cases. The problem tackled in this thesis can be considered one of those cases where the CNNs trained for sign language recognition might also capture signer information. As a result, the appearance of the signer might unintentionally influence the sign language recognition performance. Therefore, learning signer-invariant feature representations for signs is imperative.

In pose or action detection problems, models are affected by the identity information of subjects. In gait recognition frameworks, models get affected by subjects' clothes. Disentangled representation learning is a field that aims to solve these problems by either cleaning the main representation from redundant information or separating its relevant and irrelevant parts.

There are various approaches to providing disentangled representation learning. One of these approaches is to use generative adversarial networks. Tran et al. proposed a Generative Adversarial Network (GAN) based model named DR-GAN [4] to extract pose-invariant features for face recognition tasks. The DR-GAN architecture aims to produce generative and discriminative representations, which distinguish it

from prior GAN models. For a face image, it learns an identity representation using its encoder-decoder structured generator. By adding extra information, such as pose information, to the representation encoder gives to the decoder, DR-GAN aims to disentangle pose-related and identity-related features. This method helps DR-GAN to learn a discriminative representation.

Liu et al. also proposed a generative approach, named MTAN [17], inspired by DR-GAN [4], where style-invariant features are extracted using a disentangled feature learning technique. Their model also comprises encoder-discriminator-generator architectures. On the other hand, MTAN proposes two discriminators for two different tasks, which are image classification based on its content and style. While the encoder and the content discriminator are trained in agreement, the encoder and the discriminator compete during training. The feature disentanglement is enabled when a base image with different styles is generated such that the encoder can learn the image content rather than its style. MTAN [17] overperforms the state-of-the-art generative feature disentanglement models.

Another approach for disentangled representation learning is to use encoder-decoder architectures with efficient domain-specific loss functions. Yang and Yao proposed a variational autoencoder-based method [18] for learning disentangled representations of hand poses. By using different factors of variations (hand pose, viewpoint of the camera, and image content), the VAE framework learns a disentangled representation for RGB hand images. Jiang et al. [19], on the other hand, considered disentangling face representations. They claim that since face representations are composed of a nonlinear combination of expression and identity attributes, using a nonlinear disentanglement method rather than a linear one will give better results. From this point of view, they proposed to use a vertex-based deformation presentation for faces, rather than Euclidean coordinates, to get a disentangled 3D face shape representation. Experiments show that they can get more natural and accurate expression transfer results.

Aiming speaker-independent lipreading, Zhang et al. proposed a model [20] that designed in a way to disentangle the content-related and identity-related features, called Disentangled Visual Speech Recognition Network (DVSR-Net). The network consists of three parts: disentanglement module, speaker classification module and visual speech recognition module. Disentanglement module includes an encoder and a decoder. Encoder's aim is to extract content and identity features from videos. The identity features of the videos are exchanged and concatenated with their content features. Decoder reconstructs videos from these concatenated feature vectors.

Disentangled representation learning is commonly studied in gait recognition domain as well. For instance, Zhang et al. [3] investigated an appearance invariant representation learning approach where an encoder-decoder based network is designed to disentangle the appearance and pose features of each video frame. To consider the temporal relation of pose features and combine them into a gait feature for gait identification purposes, a multi-layer LSTM has been used. The components of the overall loss function are gait similarity loss, cross reconstruction loss, and incremental identity loss. Since reconstruction loss may not enable learning a representation that can be disentangled into appearance and pose, the paper proposes a cross reconstruction loss. The cross reconstruction loss can guarantee the two features are representative enough to reconstruct video frames. Also, for gait similarity loss, they enforce the similarity between the two videos' averaged pose features. This overcomes the problem of comparing two videos that don't have strict frame-level alignment.

Person re-identification is another field that suffers from appearance changes in subjects. To generate cloth-invariant shape-based representations, Li et al. [6] proposed CASE-NET, a generative model comprising two encoders for shape and color, one feature discriminator, one image generator, and one image discriminator. To handle different colors of images, the authors generated a gray-scale version of the dataset used in the study. In the model, the shape encoder takes RGB and gray-scale versions of the same image as input, while the color encoder takes the same colored but a different posed image of the same person. The output of the shape encoder is fed

to the feature discriminator to measure the difference between the different colored same posed images. Meanwhile, the image generator takes the feature representation from the color encoder and the output of the shape encoder with a grayscale image. Finally, the output of the image generator is fed to the image discriminator. The person re-identification task is accomplished with RGB and grayscale outputs of the shape encoder. Both qualitative and quantitative results show that the model learned clothing and color invariant representations.

Disentangling feature representations is one of the core approaches in image manipulation. Oldfield et al. [7] proposed a model which consists of encoders per attribute and a decoder to provide attribute-invariant representation learning. The idea is to feed the attribute encoders and try to generate two images. One image in which they try to generate the same input image and a second one in which they try to generate an original image composed of different attribute features from all the minibatch. Different loss functions have been used, such as classification loss for all different attributes, reconstruction loss calculated between the original image and reconstructed image, and adversarial loss to understand if the second reconstructed image is realistic. There is also a novel loss function researchers propose in this paper, as a disentanglement loss. This loss function ensures that feature representations of different attributes are distinctive across the attribute classes.

Similar to the motivation of this thesis, Ferreira et al. [12] addressed the signer dependency problem but considered hand-sign datasets. The authors followed a similar design where an encoder learns the sign representations, and two classifiers are trained to recognize signs and signers to avoid the signer identities leaking into the learned representations. During adversarial training, the encoder is updated so that the sign classification loss is reduced and the signer classifier is fooled. In the experiments with three different datasets, the authors reported performance improvements compared to CNN-based classifiers without a signer classifier and the models trained with triplet loss.

Domain adaptation approaches can be inspiring while solving the disentangled representation learning problem. In domain adaptation, the objective is to sharpen the target domain's effects while decreasing the source domain's effects. Disentangled feature learning aims to learn representations that highlight the features related to the target domain and alleviate the influence of the features related to the source domain. In this regard, Ganin and Lempitsky [13] introduced a framework where the model in one domain can be adapted to another domain. The framework incorporates a domain classifier for distinguishing the source and target domain features. The authors introduced a new concept named gradient reversal [13]. The gradient reversal layer behaves as an identity function during inference. On the other hand, in the training phase, the gradient reversal layer, when added in front of the domain classifier, reverses the feedback from the domain classifier. In other words, the gradient of the domain classifier's loss with respect to the representation learning layer parameters is multiplied by $-\alpha$. Thus, the learned representations are less affected by the domain shift since the domain classifier eventually cannot successfully discriminate source and target domain features [21].

2.2. Isolated Sign Language Recognition

Sign Language Recognition aims to recognize words (glosses) or sentences represented with sign language. There are two approaches to achieving this goal: Isolated Sign Language Recognition(ISLR) and Continuous Sign Language Recognition(CSLR). ISLR aims to identify one gloss from a video, whereas CSLR seeks to identify a sequence of glosses, a sentence, by a video. In this thesis, the target is providing signer independence in isolated sign language recognition models, so in this chapter, we will focus on different approaches of ISLR. We can group ISLR models into 2D-Convolutional Neural Network (CNN) [22], 3D-Convolutional Neural Network (CNN) and Graph Convolutional Network (GCN) [23] based models.

2.2.1. 2D-CNN Based ISLR Models

One of the approaches in ISLR is to extract spatial information with 2D convolution layers and pass this spatial information into a recurrent or transformer network to interpret the meaning of the temporal relationship between the video frames.

An example research by Bantupalli and Xie [24] proposes a model for American Sign Language recognition which is composed of a CNN, specifically the model called Inception [25], and Long Short Term Memory (LSTM) [26] as RNN. Researchers evaluated recognition performance using the output of the global pooling layer and the output of the softmax layer. Even though model accuracy reached 91% with the output of the softmax layer, researchers stated that variety in signer appearance and clothes affected the model's performance negatively. The inclusion of the faces dropped the model's accuracy, and the model tended to encode irrelevant information in features. Hence, researchers had to trim the frames to only include the hand gestures. Unseen skin tones in the test set also decreased the performance.

As these challenges attract researchers' attention, Sincan and Keleş [2] presented a large-scale Turkish Sign Language Dataset consisting of 226 signs performed by 43 different signers and 38,336 isolated sign video samples. The dataset is diverse regarding background, illumination, appearance, and posture of signers. Researchers trained different deep learning models and provided evaluation results of these models on the benchmark training and testing datasets, split in a user-independent manner. They used CNNs to extract the features and gave these features to unidirectional and bidirectional LSTMs to extract temporal information. In the evaluation part, to understand their models' performance on a well-known dataset, researchers tested their model on the Montalbano dataset. The model reached 96.11% accuracy, which is a competitive result with the state-of-the-art methods. Then, on the new dataset they presented, The Ankara University Turkish Sign Language Dataset (AUTSL), with random training-testing splits, their model performed up to 95.95% accuracy, but with user-independent training-testing splits, their model achieved 62.02% accuracy. This

gap between random and user-independent splits results shows the signer dependency challenge in sign language recognition tasks.

One real-life large-scale sign language dataset was proposed by Koller, and Joze [27], which is called MS-ASL. MS-ASL contains over 25000 videos shot for 1000 signs with 222 signers in realistic settings. With variety between its signers, MS-ASL allows separate training-testing sets in a signer-independent manner. Inspired by Donahue et al. [28], they determined a model composed of a VGG16 [29] followed by an LSTM [26] as one of their evaluation models. Experimental results suggest that the MS-ASL dataset is difficult for the 2D-CNN model, and LSTM couldn't propagate the recurrent information enough.

Another large-scale sign language dataset was presented from Li et al. [30], a Word-Level American Sign Language (WLASL) video dataset. WLASL contains over 2000 glosses performed by more than 100 signers. Researchers provided various baseline models performed on RGB images or joints. One of the baseline models they proposed is a combination of a pre-trained VGG16 [29] and a stacked GRU [31], which ended up with poor performance, especially with number of glosses such as 1000 and 2000.

2.2.2. 3D-CNN Based ISLR Models

Another common approach of ISLR is extracting spatial and temporal information together using 3D Convolutional Neural Networks. Huang et al. [32] proposed a 3D-CNN model to extract spatio-temporal features from the raw video. Until this research, existing SLR methods used hand-crafted features to represent sign language movements. Researchers used a sequence of 3D convolution layers as a feature extractor, followed by a multi-layer perceptron classifier. To convert the model into a multi-channel model and boost the performance, they feed the model with five different channels: color R, color G, color B, depth, and body joints. The proposed model outperformed the conventional GMM-HMM model. Researchers enhanced their study by adding an attention mechanism into their 3D-CNN model [33]. Inspiring from hu-

man vision, they have benefitted from a spatial attention mechanism that makes a viewpoint selection to focus on the most relevant parts of the frame and ignore the background and irrelevant parts. They implemented a pooling mechanism to combine clips to highlight the important ones as a temporal attention mechanism. Results seem satisfying on the "ChaLearn Looking at People Challenge 2014" (ChaLearn14) dataset [34] compared to the 3D-CNN models without attention mechanisms.

With their large-scale dataset proposal, Koller and Joze [27] and Li et al. [30] also proposed 3D-CNN baseline models for their datasets. Koller and Joze [27] adopted the architecture of the I3D networks [35] which includes 3D convolutional layers, 3D max-pooling layers, and inflated Inception-V1 submodules. They fine-tuned the pre-trained model, which was trained on ImageNet [36] and Kinetics [35], with MS-ASL and ended up overperforming the state-of-the-art methods with a large margin. Li et al. [30] also employed the network architecture of I3D [35] by fine-tuning the network on their proposed dataset, WLASL. The only part they modified in the network was the last classification layer because of the difference in the class number. Fine-tuned I3D overperformed the other baseline models they have determined for training on WLASL.

Another large-scale sign language dataset, BosphorusSign22k [1], is proposed by Akarun et al. One of the baseline experiments on the BosphorusSign22k was fine-tuning the MC3 network [37] in different levels. Researchers claim that since the network was pre-trained on Kinetics-400 dataset [38], it has never seen any sign gestures, so fine-tuning only the last fully connected layer didn't end up with a good performance. Indeed, fine-tuning some additional layers with the final fully connected layer performed better.

2.2.3. GCN Based ISLR Models

Using body joints as informative inputs for the recognition models is a common approach. Amorim et al. [39] proposed to use Spatial Temporal Graph Convolutional

Networks [40] architecture for sign language recognition. ST-GCN aims to use a hierarchical representation of the local regions rather than the whole skeleton. Even though it didn't overperform some of the traditional methods mentioned in the paper, considering Spatial Temporal Graph Convolution Neural Network as an approach for sign language recognition is a promising step. Researchers suggest weighting the body parts and including the depth information to make the observation three-dimensional as future works.

Besides other methods they covered in their research, Li et al. [30] proposed a pose-based approach for ISLR called Temporal Graph Convolution Network. TGCN takes concatenated 2D keypoint coordinates for each frame, pass them through 3 residual graph convolutional blocks, and uses a pooling layer along the temporal dimension followed by a softmax layer to classify the gloss. Despite the challenges and high gloss variety in their new WLASL dataset, TGCN shows a comparable performance.

Sincan et al. proposed a new ISLR dataset AUTSL [2] and AUTSL was the provided dataset for ChaLearn LAP Large Scale Signer Independent Isolated Sign Language Recognition Challenge. In this challenge [16], the group that came first in this competition used a multi-stream Graph Convolutional Network model [41]. They proposed to use whole-body keypoints as skeleton modality and whole-body features extracted from a pre-trained pose estimation model as another modality. The recognition results were combined with other modalities of RGB and optical flows to improve the accuracy.

Comparing the SOTA models they have trained on RGB versus key points, Vazquez et al. [42] proposed benchmark research on AUTSL [2]. For the key points, researchers have selected MS-G3D [43] as the model. Similar to the TGCN architecture Li et al. proposed in [30], MS-G3D is composed of stacked Spatial Temporal Graph Convolutional Networks, followed by a global average pooling layer, and finally, a softmax layer for classification. MS-G3D proposes a unified spatial-temporal graph convolution part, G3D, that combines the GCN module (for spatial features) and TCN

module (for temporal features). Researchers proposed that by providing a better understanding of various levels of semantic information in the graph, MS-G3D can help to catch the related information between both hands and the other body parts in sign language videos.

3. METHOD

In the domains where inputs are image or video, deep learning models tend to get affected by different aspects of the images which are not the main target. For example, a face recognition model's performance can be affected by the pose differences beside its inputs [4, 5], or an action recognition model's performance can be affected by the clothes or appearance of the person in the video [3]. Different feature disentanglement methods are being used to provide robustness in the recognition models' performance and increase generative models' interpretability [4–7].

As signers' faces, upper bodies, and how they perform the sign vary, spatial embeddings tend to capture these characteristics. However, the SLR task is expected to focus on the spatial and temporal information about the glosses, independent of the signers. Signer identity information should be removed or disentangled from the embeddings used in gloss classification, to improve the SLR performance. Feature disentanglement methods can help at this point.

In this chapter, we introduce adversarial training and different regularization techniques such as KL divergence [14], OT distance [15], MSE loss; and finally how we used them in our proposed method given in Figure 3.1, to provide feature disentanglement and signer independency in isolated SLR models.

3.1. Adversarial Training

One of the approaches we used in this study to remove the signer-dependent information from embeddings is using a gradient reversal layer, before a signer classifier. This approach aims to reduce the discriminative signer characteristic information in the encoder's representation. As an auxiliary, signer classifier is used to facilitate feature disentanglement. The forward propagation and backpropagation expressions of this

approach can be written as

$$\mathcal{R}\left(\mathbf{e}_i, f_s; \alpha\right) = \begin{cases} \mathbf{e}_i, & \text{forward propagation} \\ -\alpha \frac{\partial f_s}{\partial \mathbf{e}_i}, & \text{backpropagation,} \end{cases} \tag{3.1}$$

where α is a hyperparameter to control the amount of the adverse effect of gradient reversal, f_s is signer classifier and $\mathbf{e}_i$ as signer embedding.

The gradient reversal approach [13] ensures that the encoder weights are updated to decrease the signer classification performance. Thus, the learned spatial representation of the video frame should not contain the signer characteristics and should confuse the signer classifier. At the same time, the signer classifier, denoted by f_s, is trained to improve the signer classification performance to ensure that the signer classifier can provide relevant feedback in backpropagation. This adversarial behaviour facilitates disentangling gloss-related features (those passed to the LSTM and used for gloss classification, or 3D CNN representations) from distinctive signer features.

However, our experiments suggest that more than using the signer classifier with a gradient reversal layer approach is required for disentangling signer and sign representations. Therefore, we propose using different regularization approaches to reduce the discrepancy between the distributions of the gloss and signer representations. We hypothesize that, even if we try to separate signer-related and sign-related information in the encoder by dividing the output into two branches, since the SLR has temporal and spatial aspects, there may still be an information leakage between the parts. This leakage may result in the poor disentanglement of redundant information in the representations.

3.2. Regularization Techniques

There are different metrics in the literature to measure the distance between two distributions. This study considers Kullback-Leibler (KL) divergence, Optimal Transport (OT) distance, and Mean Squared Error (MSE) to facilitate the elimination

of signer characteristics from the sign representation.

Kullback-Leibler (KL) divergence [14] measures the difference between two probability distributions, which can be written as

$$KL(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}, \tag{3.2}$$

where P(i) is the uniform distribution and Q(i) is the signer embedding output of the encoder. It is calculated as the difference between the expected log-likelihood of the data under the ground truth distribution and the expected log-likelihood of the data under the predicted distribution. Eventually, it measures how closely the predicted distribution matches the ground truth distribution. When we minimize the distance between the uniform distribution and signer classifier output's output, we aim to enforce the encoder to output sign representations where signer information is eliminated.

In our problem, our ground truth distribution is a uniform distribution over the signer identities which means each identity is equally likely. The KL divergence between the signer distribution predicted by the signer classifier and a uniform distribution is calculated to encourage the encoder to learn a signer representation such that the signer classifier cannot discriminate between signers. In this manner, KL divergence, given in Equation (3.2), can reinforce the feature disentanglement through the signer classifier's predictions.

Optimal Transport (OT) distance, is also a measure of the distance between two probability distributions [15]. It quantifies the amount of "mass" that needs to be moved to transform one distribution to another. In other words, it measures the minimum cost to transform one distribution into the other. It is often used as OT loss function in machine learning and statistics to compare the similarity of two distributions and evaluate the performance of algorithms that work with probabilistic data [15]. OT loss is also known for stabilizing the GAN training with large mini-batches [44, 45]. Wasserstein distance is popular in GAN training since it improves

stability and eliminates the mode collapse problem [46].

In the context of disentanglement, OT loss is used to improve the disentanglement by reducing the variations in the representation of videos that belongs to different signers. A lower OT distance indicates a closer match between the two representations (which belong to two different signers), allowing the encoder to learn to encode signer independently. The OT Loss can be expressed as

$$\mathcal{L}_{\mathrm{OT}}\left(\boldsymbol{\mu}_{s_i}, \boldsymbol{\mu}_{s_j}\right) = \inf_{\gamma \in \Pi\left(\boldsymbol{\mu}_{s_i}, \boldsymbol{\mu}_{s_j}\right)} E_{(\mathbf{s}_i, \mathbf{s}_j) \sim \gamma} c\left(\mathbf{s}_i, \mathbf{s}_j\right), \tag{3.3}$$

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 cost function (which can be Euclidean distance, Wasserstein distance, cosine distance etc.) 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.

To reduce the differences between the gloss representations depending on signer characteristics, in some of our experiments, we calculated the Mean Square Error (MSE) between the representations of the same gloss performed by different signers. The MSE Loss can be written as

$$\text{MSE Loss} = \sum_{i=1}^{N} ||\mathbf{h}_i^+ - \mathbf{h}_i^-||_2^2, \tag{3.4}$$

where $\mathbf{h}_i^+$ and $\mathbf{h}_i^-$ are the hidden representations of the LSTM unit's last time step for the same gloss performed by two different signers. This approach differs from KL and OT by directly penalizing the differences between the gloss embeddings rather than manipulating the encoder weights over the signer embedding of the encoder.

3.3. Proposed Method

In this subsection, we present proposed architectures and training frameworks.

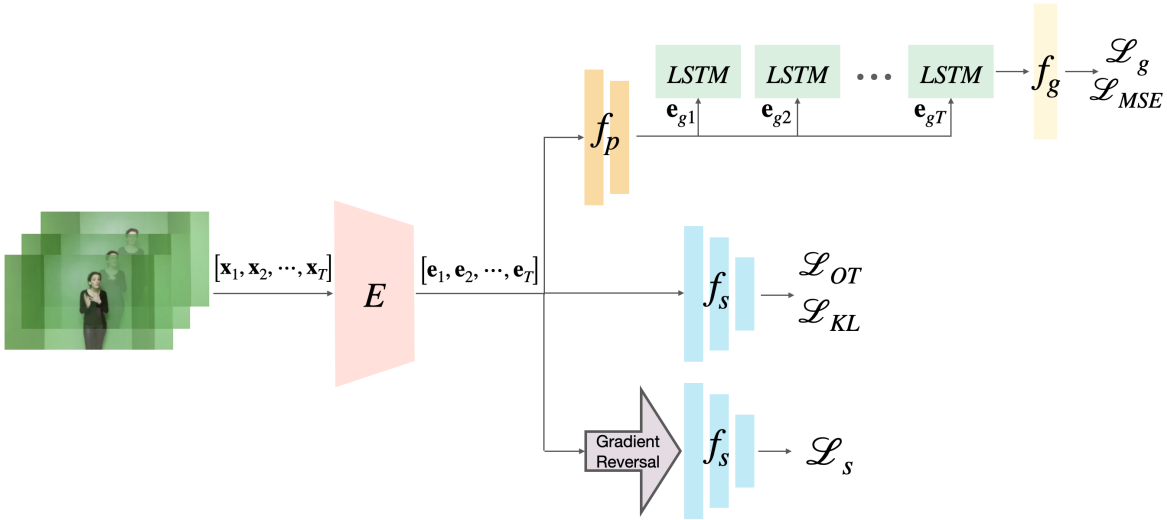


Figure 3.1. The proposed model with feature disentanglement. The gradient reversal layer is added to the output of the encoder layer before the signer classifier, f_s. OT loss ($\mathcal{L}_{\text{OT}}$) or KL divergence ($\mathcal{L}_{\text{KL}}$) is calculated using signer representations, yet MSE loss ($\mathcal{L}_{\text{MSE}}$) is calculated using gloss representations.

3.3.1. Backbone Architectures

Isolated Sign Language Recognition is a multi-class classification task that heavily depends on the spatial and temporal features extracted from the sign language videos. In this thesis, we used two approaches while processing spatial and temporal information:

1. Extract spatial features with an encoder (2D CNN) and pass these features to an RNN.
2. Process spatial and temporal information in a single architecture as 3D CNN.

We train all models end-to-end instead of using pre-trained architectures.

3.3.1.1. 2D CNN + LSTM. The spatial representations of input video frames are encoded using an encoder consisting of convolutional layers and the embeddings extracted from the encoder are transformed into frame representations using fully-connected pro-

jection layers. They can be expressed as

$$\mathbf{e}_{ei} = E\left(\mathbf{x}_i; \theta_E\right), \quad i = 1, \cdots, T, \tag{3.5}$$

$$\mathbf{e}_{gi} = f_p\left(\mathbf{e}_{ei}\right), \quad i = 1, \cdots, T, \tag{3.6}$$

respectively, where $\mathbf{x}_i$ is the ith frame of the input video, $E(\cdot)$ denotes a convolutional encoder parameterized by θ_E, f_p denotes fully-connected layers to project the embeddings into the gloss space, and T is the length of the input sequence.

The SLR task requires video classification, and the third dimension of the video inputs is the temporal dimension. Recurrent neural networks and transformer architectures [47] might be considered for temporal modeling. We trained Long Short Term Memory (LSTM) network [26] with the frame embeddings, $\mathbf{e}_{gi}$. The hidden state of the last time step, $\mathbf{h}_T$, considered as the temporal representation of the gloss, is used to train a gloss classifier, f_g. There are three different encoders used in this study: A shallow encoder with 8 convolutional layers and 2 or 3 fully connected layers for projection (depending on the experiment type), Resnet-18 [48] and Resnet-34 [48].

Resnet-18 and Resnet-34 are residual networks proposed by He et al. [48]. Adding short-cut (identity) connections between convolutional layers makes residual networks easier to optimize and has remarkable generalization performance on recognition tasks. Resnet-18 consists of 17 convolutional layers and a fully connected layer. Resnet-34 consists of 33 convolutional layers and a fully connected layer. We initially used a shallow encoder to investigate the affect of feature disentanglement without the improvements stemming from residual connections and batch normalization layers in the off-the-shelf architectures. However, we do not suggest the shallow encoder since its performance cannot be compared with state-of-the-art recognition performances of the sign language benchmark datasets. For this reason, we continue our analysis with Resnet-18 and Resnet-34.

3.3.1.2. 3D CNN. Depending on the literature review we did, we suppose that signer identity affects spatial features more than it affects temporal features. In order to manipulate spatial latent space more efficiently, we formed spatial and temporal parts of the architecture modular rather than forming them in a combined architecture such as a 3D CNN. Nonetheless, after the experiments we conducted with the shallow encoder + LSTM architecture, we realized our results were behind the state-of-the-art results in Turkish SLR tasks. Thus we extended our experiments by using our proposed methods with 3D CNN. The 3D CNN architecture used in this thesis, named Res3D, is a residual network with 3D convolution layers. It is proposed by Tran et al. in [49] and improved further in [37].

3.3.2. Optimization and Training Scheme

The loss function used to train the proposed model comprises cross-entropy losses for gloss classification, $\mathcal{L}_g$, and signer classification, $\mathcal{L}_s$, can be written as

$$\mathcal{L}_g = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C_g} y_{ij}\log f_{g_j}\left(\mathbf{h}_i\right), \tag{3.7}$$

$$\mathcal{L}_s = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C_s} y^s_{ij}\log f_{s_j}\left(\tilde{\mathbf{e}}_i\right), \tag{3.8}$$

respectively, where N denotes the batch size, C_g is the number of unique glosses, C_s is the number of unique signers, $\mathbf{h}_i$ denotes the hidden state of the LSTM network at the 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.

The optimization problem of the proposed approach is posed as an adversarial training framework. The total loss function, when a regularization method is used, can be expressed as

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_s + \mathcal{L}_{\text{REG}}. \tag{3.9}$$

When OT Loss is used as the regularization method in the experiments, the $\mathcal{L}_{\text{REG}}$

term is replaced by the OT loss ($\mathcal{L}_{\text{OT}}$) calculated between the signer embeddings of two different signers performs the same sign. When KL divergence is used in the experiments, the $\mathcal{L}_{\text{REG}}$ term is replaced by the KL divergence ($\mathcal{L}_{\text{KL}}$) calculated between the signer predictions and the uniform distribution. When Mean Square Error (MSE) loss is computed between the representations of the same gloss performed by different signers, the $\mathcal{L}_{\text{REG}}$ term is replaced by $\mathcal{L}_{\text{MSE}}$.

During the optimization step of the experiments with the regularization methods such as KL divergence and OT loss; gradients of the total loss, $\mathcal{L}$, with respect to the encoder layer E, LSTM layer and gloss classifier f_g are computed as

$$\frac{\partial \mathcal{L}}{\partial E} = \frac{\partial \mathcal{L}_g}{\partial E} + \frac{\partial \mathcal{L}_{\text{REG}}}{\partial E}, \tag{3.10}$$

$$\frac{\partial \mathcal{L}}{\partial LSTM} = \frac{\partial \mathcal{L}_g}{\partial LSTM}, \tag{3.11}$$

$$\frac{\partial \mathcal{L}}{\partial f_g} = \frac{\partial \mathcal{L}_g}{\partial f_g}, \tag{3.12}$$

respectively, where LSTM parameters are updated to minimize the gloss loss function. Similarly, gloss classifier also minimizes the gloss loss function.

In the experiments that MSE loss used, gradients of the total loss with respect to the LSTM given in Equation (3.11) turns into

$$\frac{\partial \mathcal{L}}{\partial LSTM} = \frac{\partial \mathcal{L}_g}{\partial LSTM} + \frac{\partial \mathcal{L}_{\text{MSE}}}{\partial LSTM}. \tag{3.13}$$

In the experiments with the gradient reversal layer, gradients of the total loss with respect to the encoder given in Equation (3.10) change. Also, signer classification loss is added to the total loss. The gradients of the total loss with respect to the encoder and signer classifier are given as

$$\frac{\partial \mathcal{L}}{\partial E} = \frac{\partial \mathcal{L}_g}{\partial E} - \alpha \frac{\partial \mathcal{L}_s}{\partial E}, \tag{3.14}$$

$$\frac{\partial \mathcal{L}}{\partial f_s} = \frac{\partial \mathcal{L}_s}{\partial f_s}, \tag{3.15}$$

respectively. As seen in Equation (3.15), the gradient of signer classifier's loss function with respect to the spatial encoder has an adverse effect.

Eventually, in the experiments both of the gradient reversal layer and regularization techniques are used, the gradients of the total loss with respect to the encoder given in Equation (3.10) is updated as

$$\frac{\partial \mathcal{L}}{\partial E} = \frac{\partial \mathcal{L}_g}{\partial E} - \alpha \frac{\partial \mathcal{L}_s}{\partial E} + \frac{\partial \mathcal{L}_{\text{REG}}}{\partial E}. \tag{3.16}$$

In the following section, we present our experimental analysis and discuss our results. The effects of different scenarios discussed in this section are investigated in the following ablation studies.

4. EXPERIMENTAL RESULTS

4.1. Datasets

We used two sign language datasets in our experiments: BosphorusSign22k [1] and AUTSL [2]. We followed a similar pre-processing approach as Gokce *et al.* [50] for both datasets, by using openpose coordinates. Sample and pre-processed frames from the both datasets are presented in Figure 4.1.



Figure 4.1. The first and the second rows demonstrates samples from BosphorusSign22k [1] and AUTSL [2] datasets, respectively. Images in the last row are the samples after the pre-processing step.

BosphorusSign22k dataset [1] contains 22.542 videos of 744 unique Turkish sign glosses. 428 of these glosses are in the Health domain, 163 are in the Finance domain, and 174 are commonly used sign glosses. Videos are performed by six native signers, four of whom are a woman and two are men. BosphorusSign22k has been divided into training and testing datasets in a signer-independent manner, which made it very beneficial for us to use it in this thesis: The test set has one signer, and the remaining five are in the training set. The training dataset consists of 18.018 videos, and the testing dataset consists of 4.524 videos.

AUTSL dataset [2, 16] contains 36.302 videos of 226 unique Turkish sign glosses, selected from the daily spoken vocabulary. Sign videos are performed by 43 different signers. We used the dataset provided for ChaLearn Challenge [51]. The dataset has been divided into training, validation, and testing sets in a signer-independent manner; thus, there is no overlapping in signers between the sets. The training set contains 28,142 videos performed by 31 signers, the validation set contains 4,418 videos performed by six signers, and the test set contains 3,742 videos performed by six signers. In the training set, there are ∼124 samples per sign, these value is 19 and 17 for the validation set and the test set, respectively.

Both datasets contain various hand shapes and movements, including instances where only one hand is used and others where both hands are involved. Also, some signs are similar in hand shape, orientation, position, and movement; there are only minor differences. This similarity is so high for some signs belonging to different classes that contain the same hand gesture but differ only by the number of repetitions of the same gesture.

While there is a green background used in BosphorusSign22k videos, in AUTSL dataset background choice is dynamic (such as moving trees, or moving people behind the signer). Also in terms of lighting variability and different postures of signers, we can say AUTSL is more challenging than BosphorusSign22k. The AUTSL dataset contains 20 different backgrounds. Specifically test set contains 8 different backgrounds, 3 of which are not included in the training or validation sets.

4.2. Implementation

PyTorch library is used to implement all of the architectures. PyTorch implementations of KL divergence and OT loss are used in the experiments.

- 2D CNN Architecture: There are three types of 2D CNNs used in this thesis: a shallow encoder, ResNet-18 [48] and ResNet-34 [48]. The shallow encoder consists

of 8 convolutional layers and 2 or 3 fully connected layers for projection. ResNet-18 consists of 17 convolutional layers and a fully connected layer. ResNet-34 consists of 33 convolutional layers and a fully connected layer.

- LSTM Architecture: All the LSTM models in the experiments follow the same one-layer LSTM architecture with the hidden layer dimension of 1024. We give the last hidden state to the Gloss Classifier, given in Equation (3.7).
- 3D CNN Architecture: Res3D is used as the 3D CNN architecture used in this thesis, which is a residual network consisting of 17 3D-convolution layers and one pooling layer following. It is proposed by Tran et al. in [49] and improved further in [37].
- Gloss and Signer Classifiers: One fully-connected layer is added to the end of the LSTM network for the gloss classification. Two and three fully-connected layers with Leaky ReLU activation functions are added to the end of the encoder or LSTM as signer classifier respectively, in the experiments where signer classification is a task.
- Reversal Layer: In the experiment where gradient reversal approach implemented, we add reversal layers before the signer classifiers. This reversal layer manipulates the backpropagation operation by multiplying the gradients with a negative hyperparameter, α.
- Training Details: All the experiments have been implemented with the same PyTorch framework. In experiments, the hyperparameter of the reversal layer, α, gradually changes between $(0, 1)$, as suggested in [13]. Adam and AdamW optimizers are used with a batch sizes of 16 and 32. 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}. Embedding size of the encoder was 512 and hidden size of LSTM was 1024, yet according to the changes in the architecture, these values varied.

Table 4.1. Hyperparameters used in the experiments.

Dataset	**Hyperparameters**			
	Batch Size	Optimizer	Learning Rate	α in GRL
BosphorusSign22k [1]	(16, 32)	Adam, AdamW	0.00005	[0, 1]
AUTSL [2]	(16, 32)	Adam, AdamW	Cyclical LR [0.000001, 0.0001]	[0, 1]

4.2.1. Evaluation Metrics

We used Gloss Classification Accuracy (GCA) to evaluate the model's isolated SLR performance. GCA is calculated as

$$GCA = \frac{\text{number of correct predictions}}{\text{total number of samples}}. \tag{4.1}$$

There are significant similarities between the signs of different glosses in the AUTSL dataset, which can be very confusing for the recognition models [2]. For this reason, in addition to top-1 classification accuracy, we also considered top-3 and top-5 performances of the models as suggested in the AUTSL baseline study [2].

4.3. Results

We designed various experiments to investigate the effects of the reversal layer and different regularization terms. The designed experiments are presented in the Table 4.2. The performance of the proposed approaches is investigated both quantitatively and qualitatively. There are three groups of quantitative results: Results of experiments where shallow encoder is used, tested on BosphorusSign22k are given in Table 4.4 and tested on AUTSL are given in Table 4.3. Results of experiments where Resnet2D-18 is used and tested on BosphorusSign22k are given in Table 4.8 and Resnet2D-34 is used and tested on AUTSL are given in Table 4.7. Finally results of experiments where Resnet3D is used and tested on BosphotusSign22k are given in Table 4.6 and tested on AUTSL are given in Table 4.5. In the experiments, the models are trained for at least 100 epochs and the results of best performing epochs are reported.

Table 4.2. Implemented experiments and their components.

Experiment	**Architecture Components**				
	Reversal Layer (after Encoder)	Reversal Layer (after LSTM)	OT Loss	KL Divergence	MSE Loss
Experiment 1	✓				
Experiment 2	✓	✓			
Experiment 3			✓		
Experiment 4				✓	
Experiment 5					✓
Experiment 6	✓		✓		
Experiment 7	✓			✓	
Experiment 8	✓				✓
Experiment 9		✓			✓

4.3.1. Ablation Study and Quantitative Results

4.3.1.1. Experiments With the Shallow Encoder. Table 4.4 and 4.3 present the GCA performances of various frameworks on BosphorusSign22k dataset and AUTSL dataset respectively. In the tables, Encoder + LSTM denotes the baseline framework without any signer classifier and regularization. In this section "Encoder" refers to the shallow encoder we introduced in 3. We first examined the contribution of adversarial training with a signer classifier. For this purpose, Experiments 1 and Experiment 2 in Table 4.2 are trained to demonstrate the effect of gradient reversal on disentangling the signer and gloss features. We observed that adversarial training with a signer classifier after encoder, Experiment 1, improves the overall prediction performance for both datasets. In Experiment 2, we added the gradient reversal layer after the encoder and LSTM. We tried to boost the performance by disentangling all the identity-related information through both spatial and temporal dimensions. However, the performance improvement was not as high as the increase in Experiment 1. We interpret it as the difficulty of disentangling identity-related and action-related spaces in a temporal dimension.

Table 4.3. Gloss Classification Accuracy (%) results for all models implemented with the shallow encoder, tested on the AUTSL dataset. The top two rows are the benchmark results for the AUTSL dataset. 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 [2] (CNN+FPM+BLSTM+Attention)	-	-	-	49.22	68.89	75.78
SAM-SLR (GCN based) [41]	-	-	-	98.42	-	-
Encoder + LSTM (Our Baseline)	59.00	77.29	83.20	56.00	73.30	80.86
Encoder + LSTM + MSE Loss	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 Loss	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 Loss + LSTM	65.00	80.00	84.00	62.00	79.00	87.00
Encoder + OT Loss + Reversal + LSTM	71.00	86.00	90.00	**64.00**	**81.00**	**87.00**

Next, we considered adding a regularization term such as MSE Loss, denoted by Experiment 5 in Table 4.2. The MSE Loss term penalizes the differences in temporal representations of the same gloss due to signer variations. Thus, the MSE Loss regularization aims to eliminate the signer features leaked into the temporal representation of the glosses. We observe that the MSE regularization alone might indeed alleviate signer dependency. We also considered a scenario where there is no additional signer classifier, but only OT loss or KL divergence is added to the baseline. Thus, Experiments 3 and 4 in Table 4.2 are trained to explore the feature disentanglement ability of these regularizers without the gradient reversal since adding a KL divergence, or OT loss imposes an adversarial effect on learning the encoder weights. Both regularizers improved the baseline performances in both datasets, as we expected.

Table 4.4. Gloss Classification Accuracy (%) results for all models implemented with the shallow encoder, tested on the BosphorusSign22k dataset. 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 [52]	81.37	-	97.47
3D ResNets (MC3) [53]	78.85	-	94.76
IDT (HOG + HOF + MBH) [53]	88.53	-	-
Score-level Multi Cue Fusion (3D ResNets) [50]	94.94	-	99.76
Encoder + LSTM (Our Baseline)	75.00	89.10	94.09
Encoder + LSTM + MSE Loss	**83.00**	91.68	94.34
Encoder + Reversal + LSTM	77.00	91.15	94.16
Encoder + Reversal + LSTM + MSE Loss	81.00	**94.51**	**96.81**
Encoder + KL Divergence + Reversal + LSTM	78.00	92.00	95.00
Encoder + OT Loss + LSTM	80.00	92.00	95.00
Encoder + OT Loss + Reversal + LSTM	78.00	93.00	96.00

Finally, we combined the approaches above and presented the results for the proposed architectures with Experiment 6, Experiment 7, and Experiment 8 in Table 4.2. The proposed architectures improve the prediction performance over the baselines for both datasets. On the other hand, we did not observe clear domination of one of the approaches over the other ones. Best performing regularizer changes for different datasets and evaluation subsets. This result indicates that the number of signers and glosses in the training data and various challenges in the background incur different requirements for learning signer-independent representations.

4.3.1.2. Experiments with the Resnet3D architecture. Based on the results from our experiments with the shallow encoder, we have clearly observed the effects of the pro-

posed signer-independence methods. Although, for both datasets, higher state-of-the-art results are obtained using 3DCNNs or GCNs with full body joint information. Consequently, we designed Experiment 1, Experiment 3, Experiment 4 and Experiment 5 in Table 4.2 by adding a gradient reversal layer and regularizers individually at the end of the Res3D. However, none of the methods worked on a 3DCNN architecture as they did on an Encoder+LSTM architecture, as we hypothesized in the Chapter 3. None of the experiments improved the baseline performance, and we suggest the reason for it as the difficulty in manipulating a spatial-related pattern in an integrated spatiotemporal latent space.

Table 4.5. Gloss Classification Accuracy (%) results for all models implemented with the ResNet3D, tested on the AUTSL dataset. The top two rows are the benchmark results for the AUTSL dataset. 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 [2] (CNN+FPM+BLSTM+Attention)	-	-	-	49.22	68.89	75.78
SAM-SLR (GCN based) [41]	-	-	-	98.42	-	-
Resnet 3D-18 (Our Baseline)	**87.00**	**95.00**	**97.00**	**85.00**	**94.00**	**97.00**
Resnet 3D-18 + Reversal	80.00	92.00	95.00	77.00	91.00	94.00
Resnet 3D-18 + OT Loss	74.00	90.00	94.00	77.00	91.00	94.00
Resnet 3D-18 + MSE Loss	73.00	88.00	92.00	70.00	87.00	92.00
Resnet 3D-18 + KL Divergence	80.00	92.00	95.00	76.00	91.00	94.00

4.3.1.3. Experiments with the ResNet2D architectures. Finally, to reach state-of-the-art results in SLR but keep the disentanglement methods we proposed valid, we decided to use deeper 2D encoders, such as ResNet18 and ResNet34. We used ResNet18 in the experiments we conducted on the BosphorusSign22k dataset and ResNet34 in the experiments we conducted on the AUTSL dataset based on their highest baseline performance. We conducted Experiments 1, 3, 4, 5, and 7 in Table 4.2.

Table 4.6. Gloss Classification Accuracy (%) results for all models implemented with the ResNet3D, tested on the BosphorusSign22k dataset. 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 [52]	81.37	-	97.47
3D ResNets (MC3) [53]	78.85	-	94.76
IDT (HOG + HOF + MBH) [53]	88.53	-	-
Score-level Multi Cue Fusion (3D ResNets) [50]	94.94	-	99.76
Resnet 3D-18 (Our Baseline)	**92.00**	**98.00**	**99.00**
Resnet 3D-18 + Reversal	91.00	**98.00**	**99.00**
Resnet 3D-18 + OT Loss	81.00	95.00	98.00
Resnet 3D-18 + MSE Loss	89.00	**98.00**	**99.00**

We can still observe the improvements in the results of the experiments ran on BosphorusSign22k, yet there is no improvement in the results of the experiments ran on AUTSL. Since we used ResNet18 in the experiments with BosphorusSign22k and ResNet34 in the experiments with AUTSL, we suggest that the reason behind the difference in the results may be the depth difference between the encoders. In ResNet34, regularizers may lose their effects since the back-propagation chain is longer than the ResNet18 (and the shallow encoder). Considering that in the research gradient reversal approach proposed [13], the deepest network used was an AlexNet with five convolutional layers, we can expect the gradient reversal layer to lose its effect on gradients in deep architectures such as ResNet18 and ResNet34.

4.3.1.4. Experiments with Cross Datasets. We designed some cross-dataset experiments to see the signer independence effect of the proposed methods even further. Between the two datasets we used in this research, AUTSL, and BosphorusSign22k, there are 56 common glosses. We have created an AUTSL subset and a Bosphorus-

Sign22k subset with these common glosses. We have tested the model we trained on BosphorusSign22k with AUTSL subset and the model we trained on AUTSL with BosphorusSign22k subset. As we expected, the results were not good. This is because the signer difference is not the only difference between the datasets. AUTSL is a challenging dataset in terms of variety in backgrounds, lightenings, and signer body positions. We hypothesize that our models need to be able to disentangle not only signer information but also signer position and background information from gloss representations. This way, models would be more invulnerable to the background and signer body position changes.

Table 4.7. Gloss Classification Accuracy (%) results for all models implemented with the ResNet2D-34, tested on the AUTSL dataset. The top two rows are the benchmark results for the AUTSL dataset. 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 [2] (CNN+FPM+BLSTM+Attention)	-	-	-	49.22	68.89	75.78
SAM-SLR (GCN based) [41]	-	-	-	98.42	-	-
Resnet 2D-34 + LSTM (Our Baseline)	89.00	95.00	97.00	90.00	96.00	98.00
Resnet 2D-34 + OT + LSTM	**89.00**	**96.00**	**97.00**	89.00	**96.00**	**98.00**
Resnet 2D-34 + KL + Reversal + LSTM	87.00	95.00	97.00	86.00	95.00	97.00
Resnet 2D-34 + LSTM + MSE	89.00	95.00	97.00	88.00	96.00	97.00
Resnet 2D-34 + KL	80.00	92.00	95.00	76.00	91.00	94.00

4.3.2. Qualitative Analysis of Feature Disentanglement

To investigate the effect of disentanglement on the LSTM outputs used for gloss classification, we plot the spatial embeddings and the hidden states at the last step of baseline and proposed models using t-SNE in Figure 4.3 and 4.5 for AUTSL.

Table 4.8. Gloss Classification Accuracy (%) results for all models implemented with the ResNet2D-18, tested on the BosphorusSign22k dataset. 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 [52]	81.37	-	97.47
3D ResNets (MC3) [53]	78.85	-	94.76
IDT (HOG + HOF + MBH) [53]	88.53	-	-
Score-level Multi Cue Fusion (3D ResNets) [50]	94.94	-	99.76
Resnet 2D-18 + LSTM (Our Baseline)	90.00	95.00	96.00
Resnet 2D-18 + LSTM + MSE Loss	92.00	97.00	**99.00**
Resnet 2D-18' + LSTM + MSE Loss	**94.00**	**98.00**	**99.00**
Resnet 2D-18 + Reversal + LSTM	88.00	97.00	98.00
Resnet 2D-18 + OT Loss + LSTM	88.00	96.00	97.00

With the t-SNE plots, we aim to demonstrate that the performance improvement is due to feature disentanglement, as previously hypothesized. We may observe from the plots that the signer variability within the gloss groupings obtained by the proposed approach, is higher than the baseline. Furthermore, the baseline LSTM outputs, demonstrate groupings based on signers rather than the glosses.

Moreover, we investigated the euclidean distance and cosine similarity in between the LSTM outputs belong to baseline and one of the proposed methods, based on signer and gloss groupings in Figure 4.2. We can clearly see that cosine similarity in between the gloss representations belong to the same signer decreased whereas cosine similarity in between the gloss representations belong to the same label increased in both of the proposed models. Accordingly, euclidean distance in between the gloss representations belong to the same signer increased whereas euclidean distance in between the gloss representations belong to the same label decreased in both of the proposed models.

Thus, we conclude that regularized adversarial training can indeed eliminate the effects of signer characteristics from the gloss embeddings in an RGB-based SLR framework.

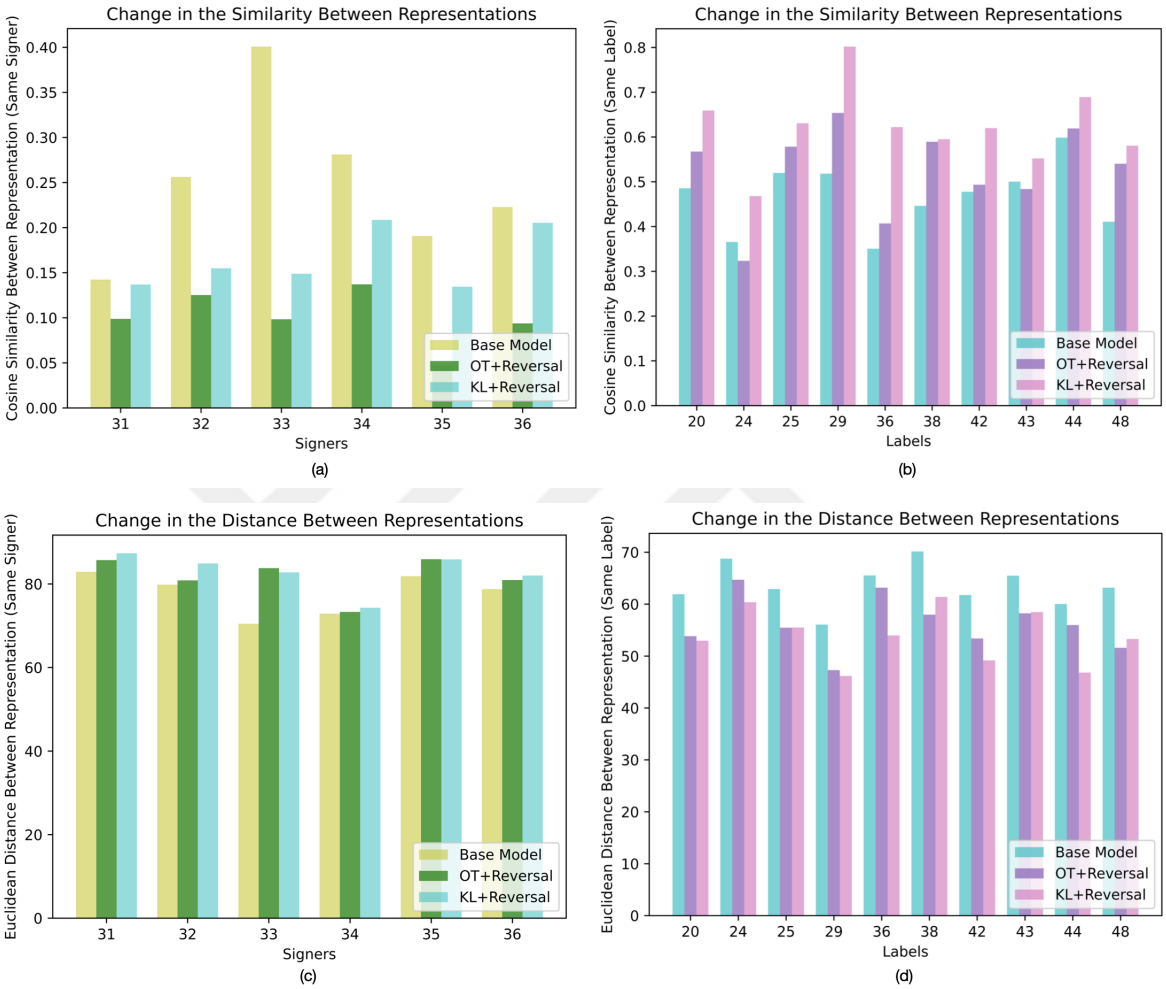


Figure 4.2. Cosine similarity between the gloss representations based on a) signer b) label groupings and Euclidean distance between the gloss representations based on c) signer d) label groupings. For the both cases, Gloss representations belongs to AUTSL dataset. Proposed methods here are Experiment 6 and 7 in the Table 4.2.

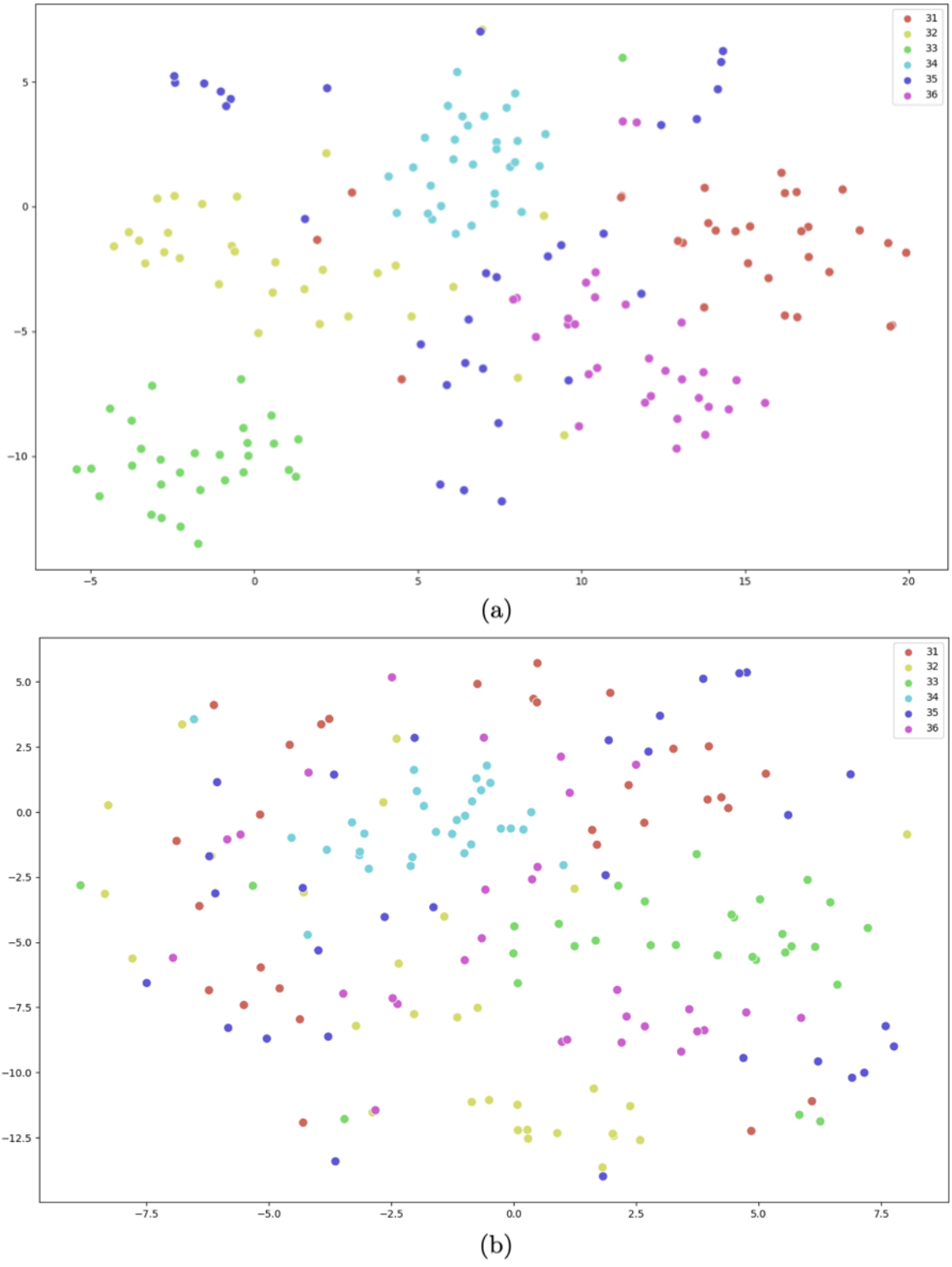


Figure 4.3. t-SNE of the Encoder outputs of a) the base model b) the proposed model on AUTSL dataset. Colors denote 6 signer classes in the Figures 4.3(a), 4.3(b). t-SNE demonstrates that the performance improvement is due to feature disentanglement, as previously hypothesized. Proposed model here is the Experiment 6 in the Table 4.2.

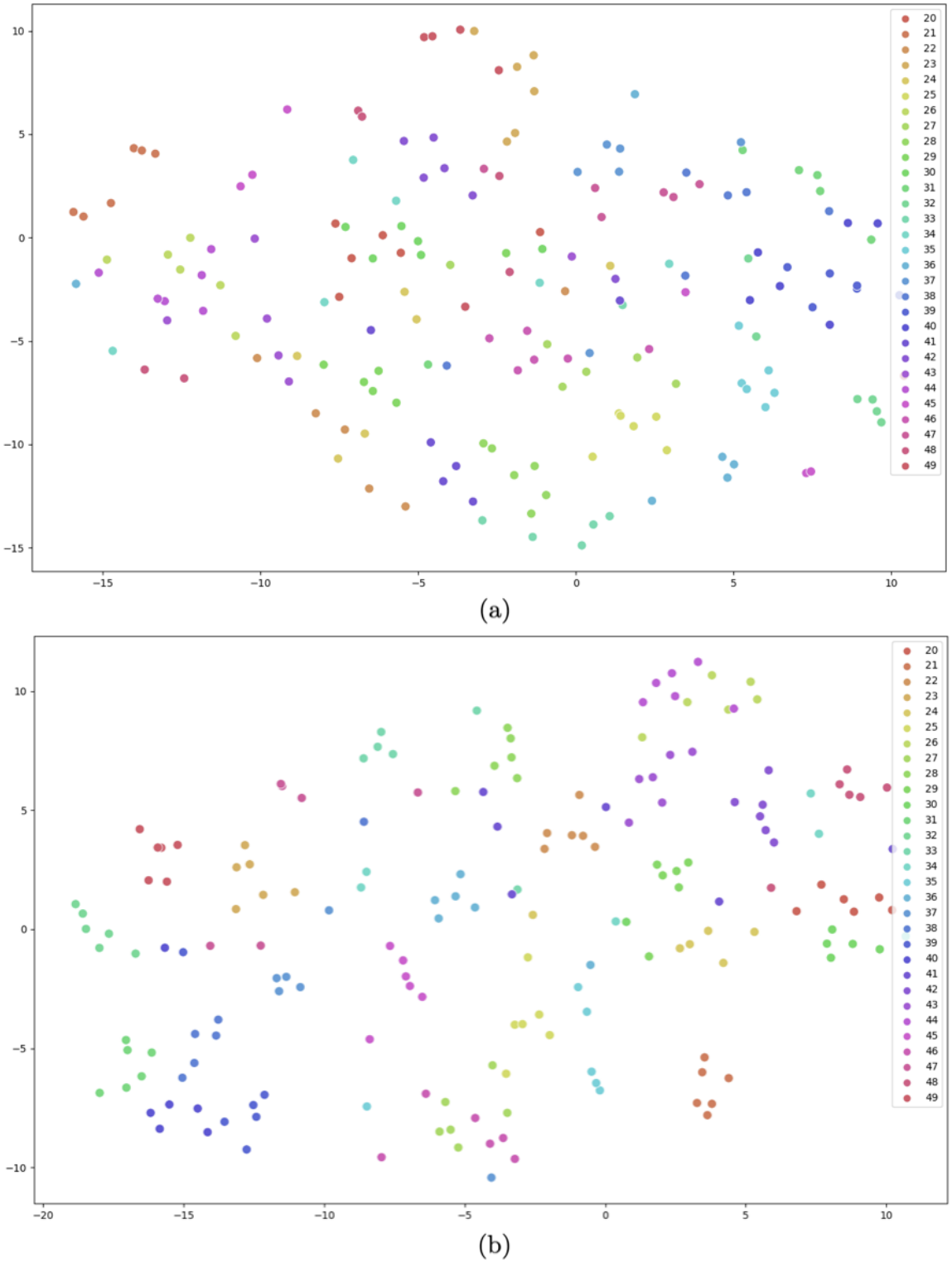


Figure 4.4. t-SNE of the LSTM outputs of a) the base model b) the proposed model on AUTSL dataset. Colors denote 30 gloss classes in the Figures 4.4(a), 4.4(b). t-SNE demonstrates that the performance improvement is due to feature disentanglement. Proposed model here is the Experiment 6 in the Table 4.2.

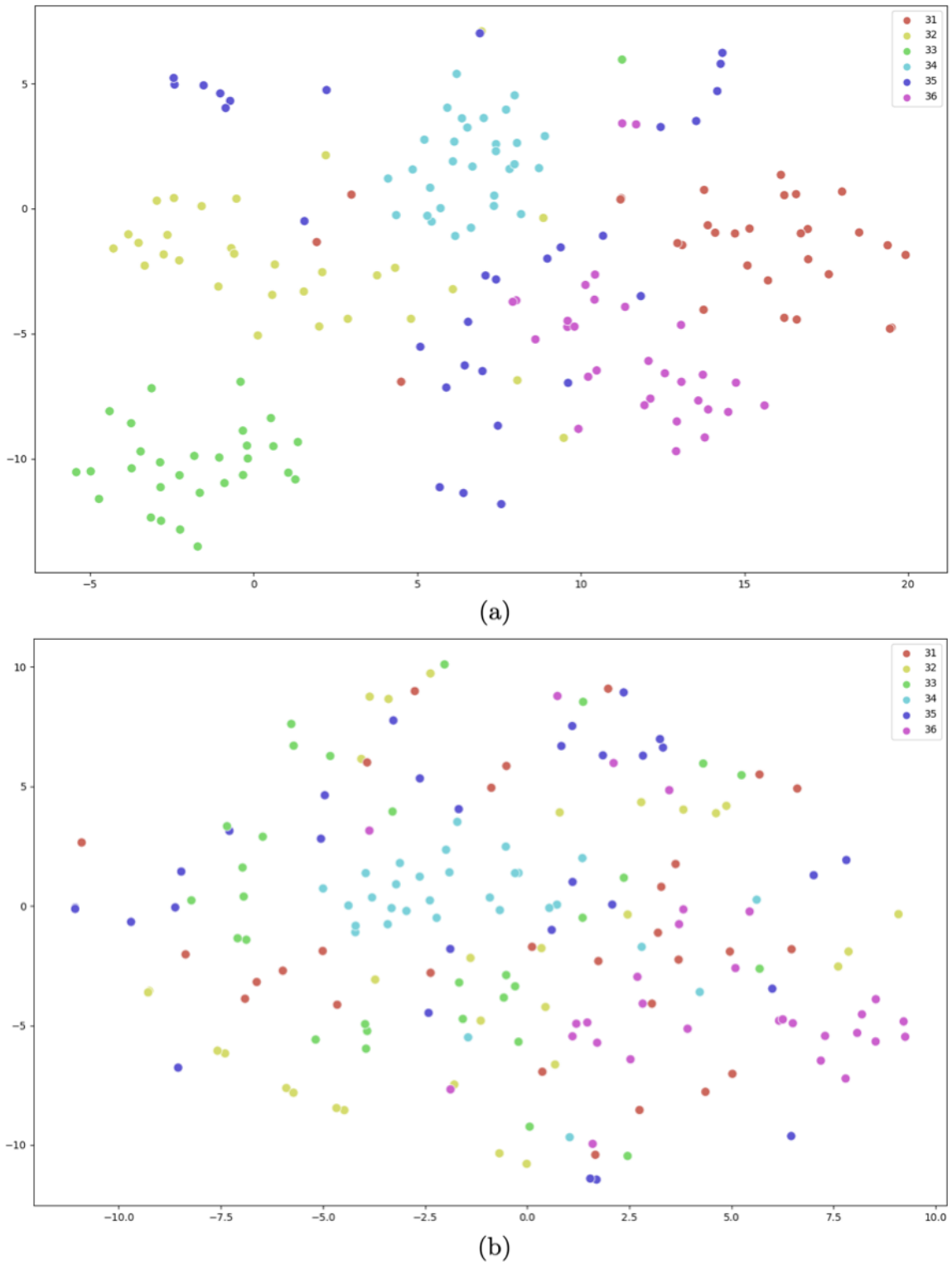


Figure 4.5. t-SNE of the Encoder outputs of a) the base model b) the proposed model on AUTSL dataset. Colors denote 6 signer classes in the Figures 4.5(a), 4.5(b). t-SNE demonstrates that the performance improvement is due to feature disentanglement, as previously hypothesized. Proposed model here is the Experiment 4 in the Table 4.2.

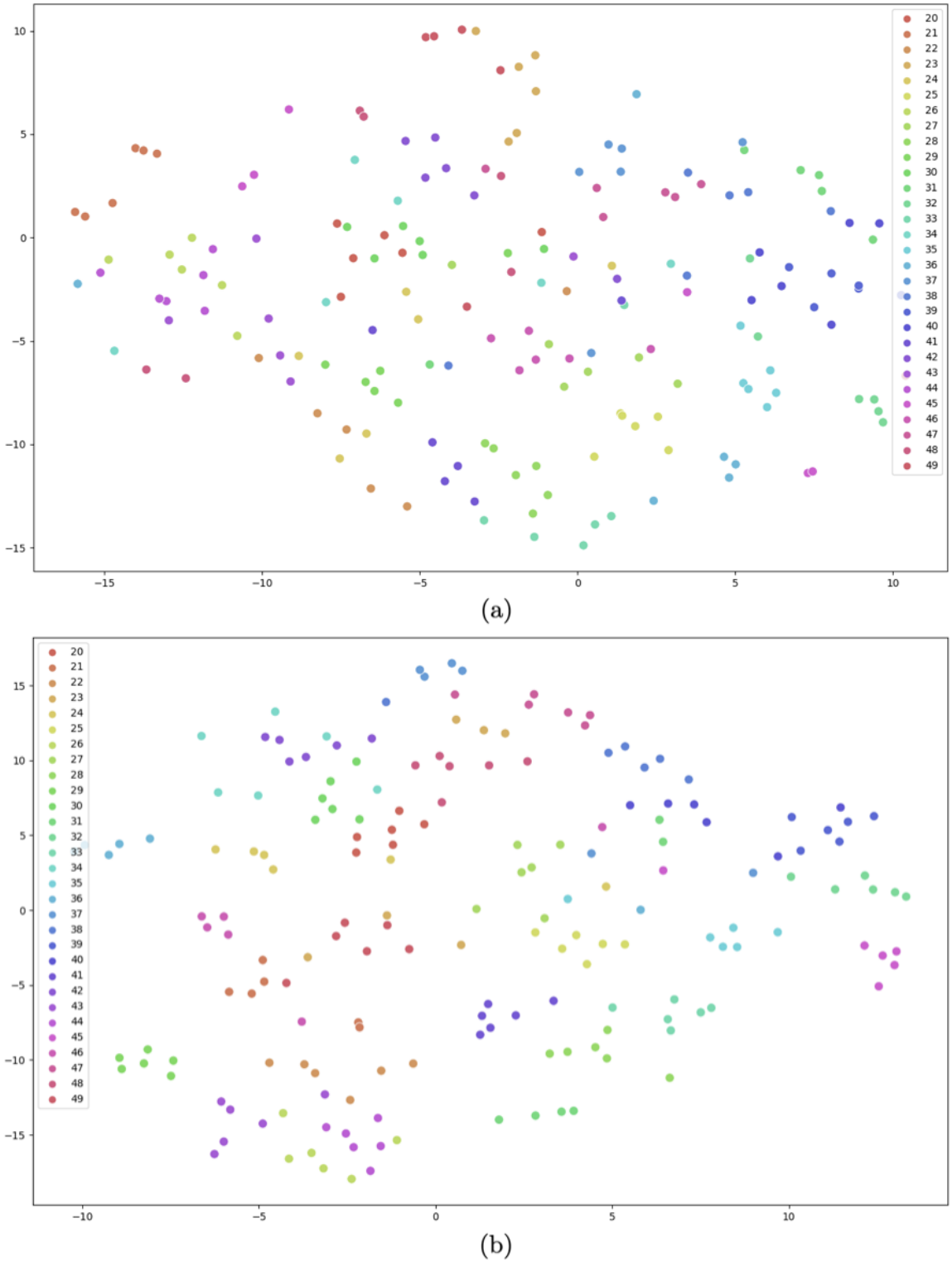


Figure 4.6. t-SNE of the LSTM outputs of a) the base model b) the proposed model on AUTSL dataset. Colors denote 30 gloss classes in the Figures 4.6(a), 4.6(b). t-SNE demonstrates that the performance improvement is due to feature disentanglement. Proposed model here is the Experiment 4 in the Table 4.2.

5. CONCLUSION

RGB-based spatial and temporal SLR frameworks are prone to memorizing signer-specific features in sign videos. The SLR aims to learn distinctive representations for different glosses rather than signer characteristics. This thesis tackles the signer dependency in convolution-based SLR frameworks by posing this task as disentangled representation learning problem. We adversarially train an encoder with a signer classifier to alleviate this challenge. KL divergence and OT distance regularizations also enhance the disentanglement of gloss representation from the signer-related features. We evaluated the proposed approach on two Turkish isolated sign language datasets. An ablation study demonstrating the effect of each component on the gloss prediction performance of 2D and 3D models is presented. Quantitative and qualitative analyses show that regularized adversarial training improves predictive performance by eliminating variations in the gloss representations due to signer characteristics. Finally, we observe that adversarial training with a signer classifier might not be sufficient. The SLR framework might also benefit from regularization, focusing on reducing the distributional differences between representations of the same gloss performed by different signers. Proposed methods with the shallow encoder provided 13% and 8% improvement in gloss classification accuracies of AUTSL validation and test datasets, respectively. In the experiments tested BosphorusSign22k test dataset, with the shallow encoder, the increase in gloss classification accuracies was 8% and with the ResNet18, the increase was 4%.

As future works, we would like to discuss potential research directions of disentangled representation learning for sign language recognition.

1) Generative models such as GANs and VAEs are popular methods to use in disentangled representation learning [4–7]. Their generative behavior is functional when integrated into an adversarial training framework. Generating images from manipulated or disentangled representations is a functioning measure and feedback mechanism.

There are many examples of GANs and VAEs used with disentangled representation learning target in the literature, yet they are mostly 2D disentangled representation learning problems. Benefiting from a generative model when the input is a video is tricky since there is no exact alignment between the videos. It's not possible to compare two videos frame by frame and get informative discriminator feedback from this comparison. There are 3DGANs, and 3DVAEs provided to overcome this problem. Even though they require much more sources to train, 3D generative models are worth exploring in the SLR task, with the aim of disentangled representation learning.

2) Body joints are informative inputs for the action recognition tasks since they don't include side information, such as redundant visuals or backgrounds. Using skeleton information with GCN-based approaches is becoming popular and gives promising results in SLR [30], [39], [42]. But there still might be signer-related differences between the body joints processed for the same sign. Implementing disentangling representation learning methods into GCN models for SLR tasks can be promising future work. Inherently, this approach might be less effective by signer identities since no RGB data is processed; but it is worth investigating.

3) There are two types of SLR tasks: Isolated Sign Language Recognition and Continuous Sign Language Recognition (CSLR). We conducted our experiments for the ISLR task, yet CSLR is worth working on regarding the signer dependency problem. In CSLR, multiple glosses must be represented simultaneously during the recognition process. Also, the segmentation of these glosses is not clear in the input and representation. There are different methods to implement a CSLR model [54–56]. These models include either a CNN+LSTM architecture that first extracts the spatial representations and then processes them in the temporal dimension or a 3D CNN model that directly extracts the spatiotemporal representation. Since both ways are CNN-based methods, the signer dependency problem will occur in these models too.

REFERENCES

1. Özdemir, O., A. A. Kındıroğlu, N. C. Camgöz and L. Akarun, "BosphorusSign22k Sign Language Recognition Dataset", *arXiv preprint arXiv:2004.01283*, 2020.

2. Sincan, O. M. and H. Y. Keles, "AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods", *IEEE Access*, Vol. 8, pp. 181340–181355, 2020.

3. Zhang, Z., L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan and N. Wang, "Gait Recognition via Disentangled Representation Learning", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4710–4719, CA, USA, 2019.

4. Tran, L., X. Yin and X. Liu, "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1283–1292, Honolulu, HI, USA, 2017.

5. Hou, X., Y. Li and S. Wang, "Disentangled Representation for Age-Invariant Face Recognition: A Mutual Information Minimization Perspective", *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3672–3681, Montreal, QC, Canada, 2021.

6. Li, Y.-J., Z. Luo, X. Weng and K. M. Kitani, "Learning Shape Representations for Clothing Variations in Person Re-identification", *arXiv preprint arXiv:2003.07340*, 2020.

7. Oldfield, J., Y. Panagakis and M. A. Nicolaou, "Adversarial Learning of Disentangled and Generalizable Representations of Visual Attributes", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 33, pp. 3498–3509, 2019.

8. Rastgoo, R., K. Kiani and S. Escalera, "Sign Language Recognition: A Deep

Survey", *Expert Systems with Applications*, Vol. 164, p. 113794, 2021.

9. Gökgöz, K., "Syllables in TİD", *Dilbilim Araştırmaları Dergisi*, Vol. 29, No. 1, pp. 29–49, 2018.

10. Cheok, M. J., Z. B. Omar and M. H. Jaward, "A Review of Hand Gesture and Sign Language Recognition Techniques", *International Journal of Machine Learning and Cybernetics*, Vol. 10, pp. 131–153, 2019.

11. Benitez-Quiroz, C. F., K. Gökgöz, R. B. Wilbur and A. M. Martinez, "Discriminant Features and Temporal Structure of Nonmanuals in American Sign Language", *PLOS ONE*, Vol. 9, No. 2, 02 2014.

12. Ferreira, P. M., D. Pernes, A. Rebelo and J. S. Cardoso, "Signer-Independent Sign Language Recognition with Adversarial Neural Networks", *International Journal of Machine Learning and Computing*, Vol. 11, pp. 121–129, 2021.

13. Ganin, Y. and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation", *International Conference on Machine Learning*, pp. 1180–1189, PMLR, Lille, France, 2015.

14. Kullback, S. and R. A. Leibler, "On Information and Sufficiency", *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79 – 86, 1951.

15. Zhang, H. and J. Wang, "Defense Against Adversarial Attacks Using Feature Scattering-based Adversarial Training", *Advances in Neural Information Processing Systems*, Vol. 32, 2019.

16. Mercanoglu, O., J. Jacques, S. Escalera and H. Keles, "ChaLearn LAP Large Scale Signer Independent Isolated Sign Language Recognition Challenge: Design, Results and Future Research", *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3467–3476, Nashville, TN, USA, 06 2021.

17. Liu, Y., Z. Wang, H. Jin and I. Wassell, "Multi-task Adversarial Network for Disentangled Feature Learning", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3743–3751, Salt Lake City, UT, USA, 2018.

18. Yang, L. and A. Yao, "Disentangling Latent Hands for Image Synthesis and Pose Estimation", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9877–9886, Long Beach, CA, USA, 2019.

19. Jiang, Z.-H., Q. Wu, K. Chen and J. Zhang, "Disentangled Representation Learning for 3D Face Shape", *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11949–11958, Long Beach, CA, USA, 2019.

20. Zhang, Q., S. Wang and G. Chen, "Speaker-Independent Lipreading By Disentangled Representation Learning", *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2493–2497, Anchorage, Alaska, USA, 2021.

21. Orbay, A. and L. Akarun, "Neural Sign Language Translation by Learning Tokenization", *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 222–228, Buenos Aires, Argentina, 2020.

22. LeCun, Y. and Y. Bengio, *Convolutional Networks for Images, Speech, and Time Series*, p. 255–258, MIT Press, Cambridge, MA, USA, 1998.

23. Kipf, T. N. and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks", *International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

24. Bantupalli, K. and Y. Xie, "American Sign Language Recognition Using Deep Learning and Computer Vision", *2018 IEEE International Conference on Big Data (Big Data)*, pp. 4896–4899, Seattle, WA, USA, 2018.

25. Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper With Convolutions", *2015 IEEE Con-*

ference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, Boston, Massachusetts, USA, 2015.

26. Hochreiter, S. and J. Schmidhuber, "Long Short-term Memory", *Neural computation*, Vol. 9, pp. 1735–80, 12 1997.

27. Joze, H. R. V. and O. Koller, "MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language", *ArXiv*, Vol. abs/1812.01053, 2019.

28. Donahue, J., L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description", Vol. 39, No. 4, p. 677–691, 04 2017.

29. Simonyan, K. and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *CoRR*, Vol. abs/1409.1556, 2015.

30. Li, D., C. R. Opazo, X. Yu and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison", *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1448–1458, Snowmass Village, CO, USA, 2020.

31. Cho, K., B. van Merrienboer, Çaglar Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation", *Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

32. Huang, J., W. Zhou, H. Li and W. Li, "Sign Language Recognition Using 3D Convolutional Neural Networks", *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Turin, Italy, 2015.

33. Huang, J., W. Zhou, H. Li and W. Li, "Attention-Based 3D-CNNs for Large-Vocabulary Sign Language Recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 29, No. 9, pp. 2822–2832, 2019.

34. Escalera, S., X. Baró, J. Gonzàlez, M. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. Escalante, J. Shotton and I. Guyon", "Chalearn Looking at People Challenge 2014: Dataset and Results", *Computer Vision - ECCV 2014 Workshops, Proceedings*, pp. 459–473, Springer, Zurich, Switzerland, 2015.

35. Carreira, J. and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, Honolulu, HI, USA, 2017.

36. Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A Large-scale Hierarchical Image Database", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, USA, 2009.

37. Tran, D., H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, Salt Lake City, UT, USA, 2018.

38. Kay, W., J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, A. Natsev, M. Suleyman and A. Zisserman, "The Kinetics Human Action Video Dataset", *ArXiv*, Vol. abs/1705.06950, 2017.

39. de Amorim, C. C., D. Macêdo and C. Zanchettin, "Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition", *International Conference on Artificial Neural Networks*, pp. 646–657, Munich, Germany, 2019.

40. Yan, S., Y. Xiong and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, 2018.

41. Jiang, S., B. Sun, L. Wang, Y. Bai, K. Li and Y. R. Fu, "Skeleton Aware Multi-modal Sign Language Recognition", *2021 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition Workshops (CVPRW), pp. 3408–3418, Nashville, TN, USA, 2021.

42. Vázquez-Enríquez, M., J. L. Alba-Castro, L. Docío-Fernández and E. Rodríguez-Banga, "Isolated Sign Language Recognition with Multi-Scale Spatial-Temporal Graph Convolutional Networks", *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3457–3466, Nashville, TN, USA, 2021.

43. Liu, Z., H. Zhang, Z. Chen, Z. Wang and W. Ouyang, "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition", *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 140–149, Seattle, WA, USA, 2020.

44. Genevay, A., G. Peyré and M. Cuturi, "GAN and VAE from an optimal transport point of view", *arXiv preprint arXiv:1706.01807*, 2017.

45. Salimans, T., H. Zhang, A. Radford and D. Metaxas, "Improving GANs Using Optimal Transport", *arXiv preprint arXiv:1803.05573*, 2018.

46. Arjovsky, M., S. Chintala and L. Bottou, "Wasserstein GAN", *ArXiv*, Vol. abs/1701.07875, 2017.

47. Vaswani, A., N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All you Need", *ArXiv*, Vol. abs/1706.03762, 2017.

48. He, K., X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2015.

49. Tran, D., J. Ray, Z. Shou, S.-F. Chang and M. Paluri, "ConvNet Architecture Search for Spatiotemporal Feature Learning", *ArXiv*, Vol. abs/1708.05038, 2017.

50. Gökce, C., O. Özdemir, A. A. Kındıroğlu and L. Akarun, "Score-Level Multi Cue Fusion for Sign Language Recognition", *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, p. 294–309, Springer-Verlag, Berlin, Heidelberg, 2020.

51. "2021 Looking at People Large Scale Signer Independent Isolated SLR CVPR Challenge", `https://chalearnlap.cvc.uab.es/challenge/43/description/`, 2021, accessed 5-Feb-2021.

52. Kındıroğlu, A. A., O. Özdemir and L. Akarun, "Temporal Accumulative Features for Sign Language Recognition", *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 1288–1297, Seoul, Korea (South), 2019.

53. Özdemir, O., A. A. Kındıroğlu, N. Cihan Camgöz and L. Akarun, "Bosphorus-Sign22k Sign Language Recognition Dataset", *Proceedings of the LREC2020 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*, pp. 181–188, European Language Resources Association (ELRA), Marseille, France, May 2020.

54. Camgoz, N. C., S. Hadfield, O. Koller and R. Bowden, "SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition", *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

55. Yang, Z., Z. Shi, X. Shen and Y.-W. Tai, "SF-Net: Structured Feature Network for Continuous Sign Language Recognition", *ArXiv*, Vol. abs/1908.01341, 2019.

56. Pu, J., W. Zhou and H. Li, "Iterative Alignment Network for Continuous Sign Language Recognition", *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.