The NGT200 Dataset
Geometric Multi-View Isolated Sign Recognition

Oline Ranum [1] **David Wessels** [1] **Gomèr Otterspeer** [1] **Erik J Bekkers** [1] **Floris Roelofsen** [1] **Jari I. Andersen** [1]

Editors: *S. Vadgama, E.J. Bekkers, A. Pouplin, S.O. Kaba, H. Lawrence, R. Walters, T. Emerson, H. Kvinge, J.M. Tomczak, S. Jegelka*

Abstract

Sign Language Processing (SLP) provides the foundation for a more inclusive future in language technology; however, the field must first overcome significant challenges. This work addresses multi-view isolated sign recognition (MV-ISR), emphasizing the critical importance of 3D awareness for real-world SLP applications. We introduce a new benchmark for MV-ISR, the NGT200 dataset, and define MV-ISR as a distinct task from single-view ISR. We showcase the benefits of including synthetic data, and propose to condition sign representations on the spatial symmetries inherent to the visual modality of sign language. We enhance MV-ISR performance by 8%-22% using a geometrically grounded model compared to the SL-GCN baseline.

1. Introduction

Sign languages (SL) are dynamic, visual and natural languages articulated using the hands, face, and body. They are expressed through the synthesis of three-dimensional shapes, structures and movements, and leverage temporal and geometric positioning to convey meaning.

Over the past years, the automatic understanding, processing, and analysis of sign languages have gathered an accelerating amount of attention (Koller, 2020a; Rastgoo et al., 2021b). Consequentially, SLP has emerged as a diverse research area, encompassing expertise from varying fields including computer vision, natural language processing (NLP), computer graphics, linguistics, human-computer interaction, and Deaf culture (Bragg et al., 2019).

---

[*]Equal contribution [1]University of Amsterdam, the Netherlands. Correspondence to: Oline Ranum <o.a.ranum@uva.nl>.

SLP applications encompass services with automated SL accommodation, such as SL smart assistants and machine translation (SL-MT). However, despite noticeable advances within the field, SLP methods lag behind other NLP technologies (Yin et al., 2021). Future improvements hinge on overcoming challenges such as the lack of large-scale, high-quality datasets, difficulties in generalizing to new signers and situations, and the need for methods that can handle the structural complexities and visual features of sign languages (Joksimoski et al., 2022; Desai et al., 2024b). Additionally, SL linguistics is a relatively young field, dating back to the 1960s (Stokoe, 1960), with much still to be understood.

Sign Language Recognition (SLR) methods interpret sign language from videos and are essential in many SL applications. However, a gap exists between research and real-world applications, partly because most datasets capture SL from a single, frontal view. The projection of three-dimensional language into video introduces information loss, causing *i.a.* variations in the viewing angle to significantly influence downstream SLR performance.

In daily interactions, such as group conversations or crowded areas, signing is often perceived from multiple angles, making it essential for SLR systems to process signs from varying viewpoints. Additionally, when processing signing from individuals with significant cognitive or functional impairments, imposing stringent regulations on assistive tool usage can be both insensitive and inconvenient. Both scenarios necessitate user-friendly, view-invariant systems to ensure seamless interactions. Therefore, we argue that viewpoints matter in real-world SLR applications.

We take a step towards multi-view SL machine processing by introducing a publicly available multi-view isolated sign dataset and perspectives on multi-view isolated sign recognition (MV-ISR). MV-ISR/SLR is challenging due to the need for models to generalize across signer appearance, articulation style, and viewpoints, compounded by the scarcity of multi-view data. Consequently, MV-SLR algorithms are increasingly compelled to learn more efficiently from limited datasets.

Multi-view SLP introduces several questions. First, how do viewpoint transformations impact SLP accuracy? If the impact is significant, the community should consider how to ethically, consistently, and scalably introduce multi-view representations into SL data, and enhance the view-invariance of SL models. However, incorporating multi-view geometry increases the task complexity. This raises a third question: how can we ground and condition SLR architectures to increase sample efficiency, reduce model complexity, and enhance generalizability in response to the challenges posed by multi-view SLR?

To begin addressing these questions, we introduce the NGT200 dataset detailed in Section 3, designed to explore the multi-view geometry of signs. NGT200 includes 2D landmarks extracted from isolated sign video clips of both human and synthetic signers captured from multiple views. The dataset aligns with the 3D-LEX dataset (Ranum et al., 2024), providing 3D ground truth for each sign. We also release a subsection of the corresponding video data.

We construct *geometric sign graphs* from NGT200 landmarks. Each node i corresponds to a landmark x_i on the human body, with spatial proximity to nodes j corresponding to landmarks x_j. The spatial edges are configured to approximate the human bone structure. In essence, sign graphs are lower-dimensional representations of dynamic geometric shapes and structures, expressed by the articulators of signs.

In Section 4, we use the sign graphs to characterize MV-ISR by its distinction from single-view ISR, and define the task as the challenge of achieving view-invariant predictions of sign language word labels (glosses) from multi-view isolated sign data. In Section 5, we contribute to scalable multi-view data production by demonstrating the efficacy of including synthetic poses in the training data. In Section 6 we propose a method to address the spatial complexity inherent in SL, which is further improved by including multiple views in the dataset. Our approach leverages geometrically grounded models to generate representations that maintain the symmetries intrinsic to the graphs constructed from SL data, including rotational symmetries corresponding to various perspectives. By integrating these inductive biases, the geometric properties are preserved in the representation, which has been shown to improve performance in downstream tasks (Wessels et al., 2024).

Models that focus on learning geometrically grounded representations have led to state-of-the-art outcomes in diverse areas such as protein structure prediction (Jumper et al., 2021; Baek et al., 2021), n-body simulations (Bekkers et al., 2024), and 3D-modeling (Heidari & Iosifidis, 2024). The geometric nature of sign graphs suggests that GDL tools used for shape and geometry analysis (*e.g.* molecular conformations) could significantly impact SLP. In this work,

we demonstrate the potential of using equivariant models to address the complexities caused by variations in articulation style and prosodic factors, such as sign amplitude, thereby enhancing the understanding of local symmetries in neural sign representations.

Geometric SLP presents a novel challenge with NGT200 as a new benchmark for the geometry-grounded machine learning community, characterized by the search for patterns and structures governed by linguistic rules and spatio-temporal dependencies. The contributions of this work can be summarized as follows:

1. We introduce a new dataset and benchmark for the task of MV-ISR: *The NGT200 Dataset*.

2. We provide a proof-of-concept demonstrating that MV-ISR is a distinct task from ISR, necessitating the adoption of novel and more efficient approaches.

3. We demonstrate that avatar-based synthetic pose data can be used to upscale low-resource MV datasets.

4. We propose leveraging a geometrically informed model to tackle the MV-ISR task, demonstrating significant improvements in gloss prediction accuracy.

2. Background

2.1. Sign Languages and Visual Linguistics

Sign languages are visual, natural languages with their own unique structure, grammar, and lexicon. They primarily function as the main languages within Deaf communities, where they emerge and continuously evolve (Padden & Humphries, 2005; Leigh et al., 2022 - 2022). Additionally, sign languages are utilized in various forms by hard-of-hearing persons, Children/Siblings of Deaf Adults (CODA/SODA), SL interpreters, second language learners and individuals with cognitive and/or physical disorders that impact (spoken) language learning abilities. Hundreds of sign languages are thought to exist worldwide (Eberhard et al., 2022), though their prevalence, accessibility, and lawful recognition vary from country to country (Meulder, 2015; Murray, 2020).

Sign languages convey meaning through a collection of asynchronous visual information cues, expressed with manual (hands, arms, fingers) and non-manual (*e.g.* facial expressions, gaze direction, torso, head posture) articulators (Cormier et al., 2018). The basic independent meaningful unit is generally a sign language word. When considered in isolation, the structure of a sign word can largely be characterized in terms of its phonological features: handshape, place of articulation, movement and palm orientation (Stokoe, 1960).

In continuous signing scenarios, such as sentence construc-

tion or conversations, the linguistic landscape of signs transforms, as new linguistic phenomena are introduced at the suprasegmental level. While the NGT200 dataset is exclusively comprised of sign language words, its important to acknowledge that continuous sign features render the generalization of methods from isolated to continuous SLP a nonlinear and nontrivial process. Examples of such features include co-articulation, where attributes of a sign are influenced by adjacent signs, and increased variation in articulation speed and amplitude as a means to mark prosody.

2.2. Sign Language Recognition

Sign Language Recognition (SLR) is the task of automatically recognizing and interpreting sign language from videos or other motion capture data. The task is commonly divided into Isolated Sign Recognition (ISR) and Continuous Sign Language Recognition (CSLR). ISR focuses on predicting glosses (*e.g.* Sehyr et al. (2021); Athitsos et al. (2008); Kezar et al. (2023b); Joze & Koller (2019); Li et al. (2020a)), by considering visual features from videos, poses and depth estimates. CSLR is the task of recognizing and interpreting entire sign language sentences from SL-corpora (*e.g.* Forster et al. (2014); von Agris & Kraiss (2010); Schembri et al. (2013)). For an extensive overview of methods and state-of-the-art in SLR see Koller (2020b) or Rastgoo et al. (2021a). For an extensive summary of sign language datasets, see Kopf et al. (2022).

2.3. Multi-View and 3D-aware Sign Language Recognition

MV-SLR has received relatively little attention in the literature, despite its practical importance and potential to promote three-dimensional fidelity in SLP tasks. While sign languages are inherently three-dimensional, most research has focused on two-dimensional projections like single-view videos. However, an emerging body of literature indicates that 3D-awareness matters in SLP.

The study by Watkins et al. (2024) demonstrates that viewing angle significantly influences human SL recognition, suggesting that sign features transform substantially under rotation, a factor relevant to machine recognition. Additionally, other studies have found that neural networks are sensitive to the three-dimensional linguistic structures of sign language (Rodriguez et al., 2023), and conditioning neural models on these structures improves recognition accuracy (Kezar et al., 2023b).

Gao et al. (2023) provides a proof-of-concept for the importance of viewing angle in SLR. They produced a multi-view Chinese Sign Language (CSL) dataset with 14 signers and 50 sign classes. Using a Multi-View Knowledge Transfer (MVKT) model, they showed a recognition accuracy drop of over 50% when trained on frontal views and tested on frontal

and side views, respectively. They also showed that training with multiple views consistently improved accuracy.

There is a growing trend in SL data production to include multiple views. The How2Sign dataset (Duarte et al., 2021) offers over 80 hours of American Sign Language videos, including speech, English transcripts, RGB-D videos, key points, with both frontal and side views for each recording. A three-hour subset was recorded in a Panoptic studio, enabling detailed 3D pose estimations. Additionally, the fable1 dataset was recently released, a small-scale corpus comprising continuous SL fairy tales in German Sign Language recorded from 7 viewpoints (Nunnari et al., 2024). Both datasets comprise SL sentences, while the NGT200 dataset consists of SL words.

3. The NGT200 Dataset

We begin by introducing NGT200, containing pose and video data for 200 common NGT signs, captured from three viewpoints with both human and synthetic signers.

3.1. Vocabulary Construction and Resource Alignment

Due to the limited availability of sign language datasets, it is highly beneficial to position new datasets in relation to preexisting resources to maximize their utility. The vocabulary of NGT200 is aligned with the SignBank NGT Lexicon (Crasborn et al., 2020a) and the 3D-LEX dataset (Ranum et al., 2024). SignBank NGT is an extensive database that provides detailed linguistic information for individual signs, including phonetic characteristics such as handshapes and handedness. Furthermore, the lexicon provides an additional frontal-view example video for each sign. The 3D-LEX dataset contains 3D motion capture data for the NGT200 vocabulary, providing a 3D ground truth for the NGT200 dataset and enabling the sampling of synthetic data from novel views using an avatar. A comparison between the NGT200 dataset and other isolated sign datasets is provided in Appendix D.

3.2. Data Capture

The pose data is obtained from a collection of multi-view videos of signers performing isolated signs. The videos are captured using the signCollect platform, developed by Otterspeer et al. (2024). The signCollect system is a sign recording platform designed to provide a 'touchless' interface for sign capture, enabling system operation through gesture recognition. This platform automates the sign collection workflow to efficiently sample signs from multiple viewing angles. The sign collection setup for the NGT200 dataset is displayed in Figure 1, showcasing how the three views are captured from a left, front and right perspective at respectively -25°, 0 and 25° degrees apart. All three
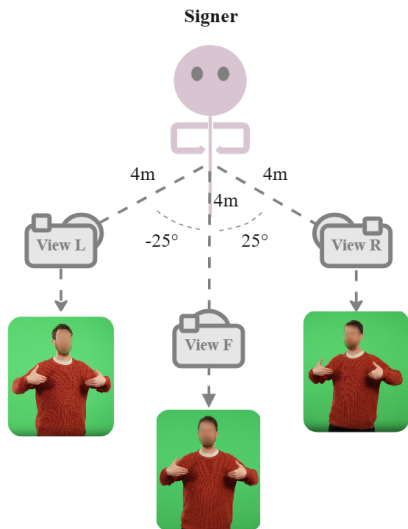
Figure 1. Configuration of video capture setup with the signCollect platform: each camera is positioned 4 meters away from the signer, with a 25° separation between cameras.

Figure 2. Spatio-temporal point clouds extracted with MediaPipe, displaying the front and right view. White landmarks represents a single frame, while the blue landmarks indicate temporal dynamics across multiple frames. Dashed lines connect the landmarks purely for visual enhancement and do not reflect elements in the dataset.

cameras are synchronously triggered to start capturing upon detection of the initialization gesture, ensuring temporal alignment between the different video clips of each view.

3.3. Spatio-Temporal Point Clouds

We use Holistic MediaPipe version 0.10.11 (Lugaresi et al., 2019) to extract landmarks from the videos, as illustrated in Figure 2. This comprehensive framework extracts pose, face, and hand key points, enabling the analysis of full-body gestures, poses, and actions. We extract 11 landmarks from the face, 14 landmarks from the body, and 21 landmarks from each hand per frame. These landmarks are combined such that each sign is represented by a matrix of size $T_c \times N_{lm} \times 3$, where T_c is the number of frames in a clip, $N_{lm} = 75$ is the total number of landmarks extracted per frame, and 3 is the x,y and z dimensions, respectively. While the x and y dimensions provide precise positions, the z dimension represents estimated depth and is prone to inaccuracies.

3.4. Pose Extraction from the Synthetic Signer

The 3D ground truth from 3D-LEX is used to create synthetic data to expand the NGT200 dataset. We retarget 3D-LEX animation files onto an avatar generated with *Ready Player Me Studio* and rendered using the open-source framework Babylon.js. Each animation clip is recorded from the screen in the browser from three perspectives matching the signCollect system's camera angles. We then extract detailed landmarks from the synthetic signer using the Holistic MediaPipe framework, as described in Section 3.3. Further
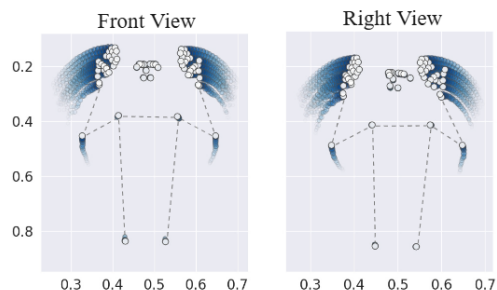
details on the process for producing the synthetic videos are provided in Appendix B.

3.5. Pose Quality Validation

Landmark detection can fail due to occlusions or rapid movements that blur body parts. We assess pose estimation quality by calculating the average ratio of successful to unsuccessful landmark extractions per sign. When extraction fails, MediaPipe returns a zero value for the affected keypoint. The NGT200 dataset has an average success rate of 97.6% for real videos and 97.8% for avatar videos. Detailed ratios are shown in Figure 7 in the Appendix C.

3.6. Dataset Characteristics

Three native NGT signers and one synthetic signer contributed to the NGT200 dataset. Two signers consented to release videos and poses, while the third consented only to videos. Each signer was assigned a unique identifier, as detailed in Table 1.

Table 1. Details on the availability of NGT200 modalities.

SIGNER ID	1	2	3	A	TOTAL
# VIDEOS	600	×	600	600	1,800
# POSES	600	600	600	600	2,400

Understanding the shapes and lingustic structures within data distribution can enhance inductive biases for ISR methods (Kezar et al., 2023b; Ranum et al., 2024). We provide some linguistic information to inform on the data distribution in the NGT200 dataset. Table 2 details the handedness distribution from SignBank: Class *1* includes one-handed signs, Class *2a* includes asymmetrical two-handed signs (non-dominant hand as location), and Class *2s* includes

4

symmetrical two-handed signs (both hands moving with the same handshape) (Crasborn et al., 2020b). Notably, signers may not consistently use their dominant hand for strong handshapes, suggesting flip symmetries as a potential data augmentation technique. However, caution is needed as some signs encode directionality, and we have not yet assessed sign directionalities in the NGT200 dataset.

Table 2. Handedness of signs in the NGT200 vocabulary. There are in total 122 one-handed signs and 78 two-handed signs.

HANDEDNESS	*1*	*2s*	*2a*
COUNT	122	63	15

The distribution of handshapes in the NGT200 vocabulary is illustrated in Figure 3. Each sign in the NGT200 vocabulary is annotated with a specific handshape for the dominant hand. Additionally, there are 78 two-handed signs, where labels are provided for the non-dominant hand as well.
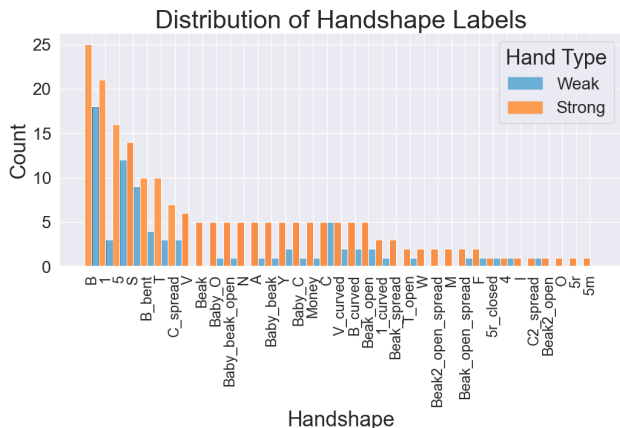

Figure 3. The frequency of each handshape type within the NGT200 vocabulary, categorized by strong (dominant hand) and weak (non-dominant hand).

4. The MV-ISR Task

We now shift our focus towards the second contribution of this work: we provide a proof-of-concept demonstrating that the MV-ISR task is distinct from that of frontal-view ISR. We conducted a series of experiments using the NGT200 dataset where we learn from and test on different sets of views. Specifically, we addressed the following questions:

Q1: *Does viewing angle matter in pose-based ISR?*
Q2: *How does the inclusion of additional views during training impact performance?*

To address these questions, we use a state-of-the-art Sign Language Graph Convolution Network (SL-GCN) (Jiang et al., 2021). We construct pose-graphs from the point clouds and train the SL-GCN to predict glosses.

4.1. Method & Experiments

Graph Construction We adopt the graph reduction scheme introduced by Jiang et al. (2021) to mitigate noise from the numerous nodes and edges in a human skeleton and to reduce distances within the graph. We downsample to 27 nodes: 10 per hand and 7 for the overall pose. We configure the spatial edges to approximate the human bone structure, as illustrated in Figure **??**. A spatial pose graph is constructed for each frame, and the ordered graph sequence represents one sign. In this work, we do not explicitly include temporal edges, which connect the same node between consecutive frames. Instead, we manage the temporal dynamics through temporal convolutions.

The SL-GCN SL-GCN is a state-of-the-art model for pose-based ISR, featuring a sophisticated design engineered towards the SLR task. It builds on a spatio-temporal GCN with a spatial partition strategy to model dynamic skeletons. Furthermore, the network is enhanced with decoupled spatial convolution layers, a spatial, temporal, and channel-wise attention module, a temporal convolutional layer, and a DropGraph module. In total, 10 spatio-temporal GCN blocks are used, followed by global average spatio-temporal pooling before classification with a fully connected layer. We use the Openhands (Selvaraj et al., 2021) implementation, featuring an SL-GCN encoder with a fully-connected classification decoder.

Training Details We construct three train-validation-test split-blocks, configured according to the k-fold cross-validation scheme illustrated in Figure 4. Each block has a distinct test set and train-validation splits. Each test set includes a novel human signer for a given sign, but the signer appears in the training set for other signs. Synthetic data and the SignBank video is excluded from test-set. The k-value is adjusted based on available data, detailed in Table 3. Single-view models use a three-fold cross-validation scheme, while models incorporating two or three views use a six-fold cross-validation scheme. The final score is the average across all eighteen folds.

Experiments We first examine whether models trained exclusively on frontal views maintain accuracy when evaluated on side views. Next, we retrain the models with the additional SignBank front-view data (ID: S_b) to assess how this impacts performance across all views. To assess the impact of multi-view data on performance, we train models on individual views and then progressively integrate additional views.

4.2. Results

Table 4 presents the results of testing models trained with only front-views on front- and side-views. Recognition accuracy sees a relative drop of more than 50% when front-view models are evaluated on side views. Including the extra S_b-view leads to a relative increase of 122% on the front-view, but only marginally improves side-view predictions. These findings suggest that more frontal-view data alone is not the solution for the MV-SLR task, as it doesn't help the model generalize across different viewpoints. In conclusion, the results indicate that viewing angle matter in MV-SLR.

Table 5 presents the results from incorporating additional views into the training dataset. Adding more views consistently enhances recognition accuracy across all perspectives, suggesting that the network learns distinct and complimentary features from each view.

5. Scaling Up Sign Language Datasets with Synthetic Data for ISR

In the preceding section, we utilized pose data from both human and synthetic signers. However, the impact of leveraging synthetic data to support SLP tasks remains uncertain, as the development of synthetic SL representations is still in its nascent stages. We ask the following question:

Q3: *Can synthetic data be effectively used to supplement MV-SL datasets in the context of boosting pose-based MV-ISR performance?*

To address this question, we conduct experiments by iteratively augmenting the training dataset with synthetic and

Train Views	Test Views	Top1 Acc	Top3 Acc
L^{123A}	L^{123}	$\mathbf{.05}_{(\pm.03)}$	$\mathbf{.10}_{(\pm.04)}$
	F^{123}	$.03_{(\pm.01)}$	$.07_{(\pm.03)}$
	R^{123}	$.01_{(\pm.01)}$	$.04_{(\pm.01)}$
R^{123A}	L^{123}	$.02_{(\pm.01)}$	$.05_{(\pm.02)}$
	F^{123}	$.03_{(\pm.01)}$	$.09_{(\pm.03)}$
	R^{123}	$\mathbf{.06}_{(\pm.02)}$	$\mathbf{.12}_{(\pm.04)}$
F^{123A}	L^{123}	$.03_{(\pm.02)}$	$.07_{(\pm.02)}$
	F^{123}	$\mathbf{.09}_{(\pm.02)}$	$\mathbf{.20}_{(\pm.03)}$
	R^{123}	$.03_{(\pm.01)}$	$.10_{(\pm.02)}$
F^{123AS_b}	L^{123}	$.06_{(\pm.02)}$	$.14_{(\pm.04)}$
	F^{123}	$\mathbf{.20}_{(\pm.03)}$	$\mathbf{.36}_{(\pm.05)}$
	R^{123}	$.05_{(\pm.02)}$	$.13_{(\pm.03)}$

Table 4. Classification accuracies and model standard deviations achieved with SL-GCN when training strictly on frontal views and testing across different views (L: left; R: right; F: front). The highlighted results indicate the top-performing view of each model.

Train Views	Test Views	Top1 Acc	Top3 Acc
L	L	$.01_{(\pm.00)}$	$.04_{(\pm.01)}$
L+F	L	$.25_{(\pm.05)}$	$.51_{(\pm.06)}$
L+R	L	$.27_{(\pm.05)}$	$.47_{(\pm.07)}$
L+F+R	L	$\mathbf{.46}_{(\pm.04)}$	$\mathbf{.69}_{(\pm.03)}$
F	F	$.09_{(\pm.02)}$	$.21_{(\pm.03)}$
L+F	F	$.35_{(\pm.05)}$	$.59_{(\pm.06)}$
F+R	F	$.42_{(\pm.05)}$	$.67_{(\pm.05)}$
L+F+R	F	$\mathbf{.49}_{(\pm.03)}$	$\mathbf{.74}_{(\pm.03)}$
R	R	$.04_{(\pm.01)}$	$.09_{(\pm.04)}$
L+R	R	$.28_{(\pm.05)}$	$.51_{(\pm.06)}$
F+R	R	$.39_{(\pm.05)}$	$.62_{(\pm.06)}$
L+F+R	R	$\mathbf{.47}_{(\pm.04)}$	$\mathbf{.72}_{(\pm.03)}$

Table 5. Classification accuracies and model standard deviations achieved with the SL-GCN using different combinations of views. The highlighted results indicate the top-performing model per test-view.

human signers to assess the impact of including synthetic pose data.

5.1. Method & Experiments

In the experiments in this section, we reuse the graph construction method and the SL-GCN described in Section 4.1.

Training Details   To provide an additional perspective on the MV-ISR task, we redefine the test set in this section to

Table 3. Details on the allocation of train-validation-test examples per number of included training views. If the SB front view is included, the number of training examples increases with one.

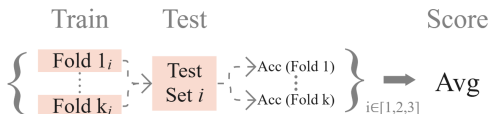| # Views	# Train	# Val	# Test
1	$2 (+1)_{Sb}$	1	1
2	$5 (+1)_{Sb}$	1	1
3	$7 (+1)_{Sb}$	2	1

Figure 4. Performance evaluation scheme using k-fold cross-validation across three distinct test sets. Accuracy scores (Acc) are computed for each fold within a test set, and the average accuracy (Avg) is calculated across all folds.

use only signer 3 for testing. Training is then conducted using exclusively poses from signers 1, 2, A, and Sb. This train-validation-test split allows us to evaluate the model's performance in the context of a novel signer prediction task, which is considered significantly more challenging than predicting signs from signers that has been seen during training.

Experiments   We iteratively augment the training data subsets to include additional signer identities and views. The experiments aimed to evaluate if:

   i The addition of a synthetic signer improves overall gloss recognition accuracy

   ii There is a performance difference between adding synthetic poses and human poses

5.2. Results

Table 6 presents the results of experiments on including synthetic data to boost the recognition performance of an MV-ISR model. We use the notation $\{v_j\}^I$, where $v_j \in \{\text{L: left, F: front, R: right}\}$, and the superscripts $I \in \{1, 2, 3, A, Sb\}$ denote the signer identities used for the selected collection of views.

Train Set	Test Set	Top1 Acc	Top3 Acc
$\{LFR\}^{12}$	L^3	$.03_{(\pm.01)}$	$.08_{(\pm.02)}$
	F^3	$.14_{(\pm.04)}$	$.27_{(\pm.04)}$
	R^3	$.14_{(\pm.02)}$	$.27_{(\pm.04)}$
$\{LFR\}^{12}+\{F\}^A$	R^3	$.09_{(\pm.02)}$	$.17_{(\pm.04)}$
	F^3	$.27_{(\pm.03)}$	$.43_{(\pm.04)}$
	L^3	$.22_{(\pm.02)}$	$.39_{(\pm.03)}$
$\{LFR\}^{12} + \{F\}^{Sb}$	L^3	$.10_{(\pm.02)}$	$.20_{(\pm.04)}$
	F^3	$.28_{(\pm.04)}$	$.45_{(\pm.04)}$
	R^3	$.26_{(\pm.03)}$	$.41_{(\pm.04)}$
$\{LFR\}^{12A}$	L^3	$.19_{(\pm.02)}$	$.34_{(\pm.03)}$
	F^3	$.43_{(\pm.02)}$	$.61_{(\pm.02)}$
	R^3	$.38_{(\pm.04)}$	$.57_{(\pm.03)}$
$\{LFR\}^{12A} + \{F\}^{Sb}$	L^3	$.32_{(\pm.04)}$	$.49_{(\pm.04)}$
	F^3	$.48_{(\pm.02)}$	$.68_{(\pm.02)}$
	R^3	$.43_{(\pm.02)}$	$.60_{(\pm.02)}$

Table 6. Classification accuracies for multi-view SL-GCN experiments with and without synthetic data, tested across different views. Accuracies are averaged over 10 runs with standard deviations. Experiments are trained on all 3 views from signers 1 and 2. When indicated, 1 frontal view from SignBank and 1-3 views of the synthetic avatar. All models are tested on signer ID 3.

The results show that adding a single frontal view from either the avatar or SignBank data improves performance across all three views. The difference between using the human signer from SignBank and the synthetic data is marginal, with human signer data providing a slightly higher boost. Additionally, incorporating more views from the synthetic signer significantly increases recognition accuracy from the baseline, with the best results achieved by leveraging all available data.

Furthermore, the experiments using the $\{LFR\}^{12A}$ train set in Table 6 are equivalent in size and views to the **L+F+R** experiments in Table 5. There is a drop of 27%, 6% and 9% in accuracy for the left, front, and right views, respectively, demonstrating that predicting the signs of a novel signer is indeed a more challenging task.

These findings provide empirical evidence that synthetic data can substantially inform recognition models when training with a pose modality. This is an important observation for the SLR community, suggesting a viable approach to scale up multi-view pose-based datasets to make practical applications of recognition models more feasible. We conclude that adding synthetic data boosts sign recognition accuracy in the pose modality and that including synthetic data in the NGT200 dataset is a beneficial strategy for enhancing overall model performance, especially across diverse viewpoints.

6. The Case for Geometric MV-SLR

As observed in the Section 4.2, the SL-GCN achieves a top recognition accuracy of 51% in our experiments on the NGT200 dataset. To explore more efficient learning in the context of MV-ISR, we propose leveraging geometrically grounded models. To take a first step in this direction, we ask the question:

Q4: *Is a geometrically grounded model viable for ISR?*

To address this question, we modify a SE(2)-equivariant neural network proposed by Bekkers et al. (2024). We explore the possibility of leveraging equivariance towards the group of roto-translations in the 2D-plane to enhance learning across inter-view intra-signer variations. Exploration of models equivariant towards the group of perspective transformations, addressing intra-view variations, is left for future work.

6.1. Method & Experiments

The PONITA Architecture   PONITA is a general purpose SE(N)-Equivariant model proposed by Bekkers et al. (2024), which achieves state-of-the-art results in tasks including interatomic potential energy prediction, trajectory forecasting

in N-body systems, and molecule generation. They formalize the notion of weight sharing in convolutional networks as the sharing of message functions over point-pairs that should be treated equally. They derive practical pair-wise attributes that uniquely identify such equivalence classes of point pairs, and subsequently use them to build efficient equivariant architectures.

To adapt PONITA to the SLR task, we make two modifications to the model architecture. The PONITA architecture including our modifications (temporal-PONITA) is summarized in Figure 5.

i After each spatial PONITA layer we add a temporal convolution module consisting of two convolution kernels, GeLU activation functions and a residual connection.

ii After the final temporal convolution block in the last network layer, we add a spatio-temporal average-based pooling operation across each pose-graph.
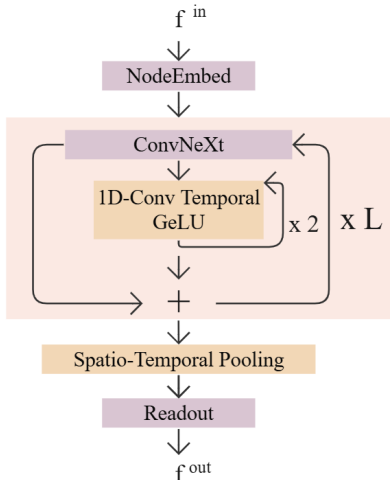

Figure 5. The modified PONITA architecture with temporal learning mechanisms (temporal-PONITA) first embeds the input features with a linear layer, which are then passed through L temporal-PONITA layers. Each layer includes one ConvNeXt block and one temporal block comprising two convolutional layers with GeLU activations.

Experiments To establish a benchmark for geometric models in the MV-ISR task, we reproduce a subsection of the experiments conducted in Section 4, replacing the SL-GCN with our temporal-PONITA. Detailed hyperparameters and training conditions are outlined in the appendix.

TRAIN VIEWS	TEST VIEWS	TOP1 ACC	ABSOLUTE GAIN
L+F	L	$.43_{\pm.03}$	+.18
L+R	L	$.48_{\pm.03}$	+.21
L+F+R	L	$\mathbf{.54}_{\pm.03}$	+.08
L+F	F	$.55_{\pm.03}$	+.20
F+R	F	$.57_{\pm.02}$	+.15
L+F+R	F	$\mathbf{.59}_{\pm.03}$	+.10
L+R	R	$.50_{\pm.03}$	+.22
F+R	R	$.49_{\pm.02}$	+.10
L+F+R	R	$\mathbf{.55}_{\pm.02}$	+.08

Table 7. Classification accuracies and model standard deviations achieved with the temporal-PONITA using different combinations of views. The *Absolute Gain* column indicates the absolute improvement of the results presented here compared to the Top-1 accuracies in Table 5.

6.2. Results

Table 7 showcases the recognition accuracy of temporal-PONITA when trained and tested on NGT200. The experiments in this table correspond to those in Table 5, and the *Absolute Gain* refers to the improvement in Top-1 accuracy compared to Table 5. Temporal-PONITA achieves higher performance across all combinations of views compared to SL-GCN. Additionally, temporal-PONITA demonstrates higher efficiency in terms of speed and exhibits more stable training runs with lower variance in predictions. An example comparing the training of temporal-PONITA and the SL-GCN is available in the Appendix Section E. These results confirm that geometrically grounded models are viable for training MV-ISR models, offering considerable benefits.

7. Discussion and Conclusion

NGT200 is a pose-based dataset, which is less common than standardized video datasets. The pose modality offers advantages, such as a lower-dimensional representation of signs that generalizes better to unseen signers and backgrounds. It can embed a skeletal inductive bias into SLP models by constructing spatial edges that mirror natural human body connections (Saunders et al., 2021). Ethically, it enhances signer anonymity compared to video. However, the pose modality has limitations in SLR due to the landmark estimation process, leading to information loss, especially when considering interacting body parts (Moryossef et al., 2021).

One of the main limitations of the NGT200 dataset is its size and scope. While NGT200 serves as a valuable research dataset, it is not suitable for training real-world ISR systems. In future work, we aim to expand NGT200 to include new signers, larger vocabularies, and continuous signing.

The NGT200 dataset has some technical limitations: a few videos are missing due to issues occurring during collection. Additionally, the synthetic data was recorded in the browser, which occasionally can result in a lower-than-normal frame rate. Details are provided in Appendix C.

We evaluated the use of synthetic data by including a single signing avatar in the training dataset, focusing on the pose modality. Our promising results indicate the potential of synthetic data in SLP, but its effectiveness in the RGB modality, which poses specific challenges related to signer appearance, remains unclear. Furthermore, we have not evaluated whether synthetic data might hinder the model's ability to learn authentic SL features at scale. Synthetic data could introduce false motion patterns and unrealistic articulation styles, potentially affecting real-world recognition performance. With the availability of larger multi-view datasets, the impact of synthetic data should be reassessed, especially for methods involving larger data distributions and continuous signing. Furthermore, capturing 3D ground truths can be costly. Future research should consider using 3D ground truths to generate variations in synthetic data, such as different sign amplitudes and articulation styles. Our study highlights the early stage of synthetic SL representations and raises the question of how to optimally leverage synthetic data to support SLP tasks.

We took a first step in assessing the potential of GDL tools for supporting SLP tasks by demonstrating the application of an SE(2)-equivariant model, which achieved significant improvements over the baseline. However, many geometrically grounded methods may be better suited for SLP tasks. Future work should consider, *e.g.*, models equivariant to 2D perspective transformations for MV-ISR/SLR.

In this work, we highlighted the importance of viewing angles in MV-ISR. Our contributions include: i) the NGT200 dataset; ii) demonstrating that recognition models trained on frontal views lose accuracy on side views, and showing that MV recognition accuracy improves when learning from multiple views; iv) enhancing the NGT200 dataset with synthetic multi-view data, which demonstrates the potential for scaling up multi-view datasets; and v) showcasing the benefits of considering geometrically grounded models for MV-ISR tasks. We hope this dataset will benefit the research community by providing a foundation for exploring a novel and intriguing task in geometric deep learning, inspiring new and stronger approaches to SLP.

8. Privacy and Ethical Considerations

The increasing demand for data to drive computational methods and machine learning algorithm introduces significant privacy risks and ethical concerns across the computational sciences. These issues are particularly pronounced in data collection involving minority groups, such as sign language communities. Bragg et al. (2020) emphasizes that gathering data from small populations inherently reduces anonymity, a situation that may be particularly undesirable in communities associated with medical conditions.

Another critical issue is the collection of data without obtaining informed consent from contributors. In the case of the NGT200 dataset, all participants provided informed consent and received compensation. The video modality of one signer is not released to preserve the anonymity of this signer. The other two signers did give informed consent for publication of video data. Names of signers are not disclosed. Instead, each signer was assigned a unique signer ID as described above.

9. Positionality Statement and Contributions

Research into the automatic understanding and processing of sign languages requires collaboration across multiple disciplines, bringing diverse positionalities, knowledge, and expertise into the team. Consequently, we include a brief note on our research team members and their respective contributions to this project.

Ranum is a hearing sibling of a signing adult with a language learning disability, and Norwegian SL is her second language; Otterspeer is deaf and an expert NGT signer; Roelofsen is a hearing parent of a deaf child, and proficient in NGT; Andersen is hearing with basic proficiency in NGT; Wessels and Bekkers are non-signing.

Ranum, Roelofsen, Wessels, and Bekkers have backgrounds in Artificial Intelligence, with Roelofsen additionally having a background in linguistics. The methods considered were mostly implemented by Ranum, who also primarily authored the current manuscript. Roelofsen supervised the development of this project, providing feedback on the manuscript and contributing to discussions. Wessels and Bekkers contributed to this project with their insights and expertise on the topic of geometric principles in deep learning; Otterspeer and Andersen have a background in programming, signing avatars and system engineering for sign capture, and developed the pipeline for producing the synthetic data. Additionally, Otterspeer conducted the collection of the video data. All authors edited and commented on previous versions of the manuscript. All authors read and approved the final manuscript.

10. Data and Code

NGT200 is available through OSF: osf.io/5zuyd/. The code for reproducing our experiments is available on OSF or (WIP) at GitHub: github.com/OlineRanum/GMVISR.

References

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., and Thangali, A. The american sign language lexicon video dataset. pp. 1–8. 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. doi: 10.1109/CVPRW.2008.4563181.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi: 10.1126/science.abj8754. URL https://www.science.org/doi/abs/10.1126/science.abj8754.

Bekkers, E. J., Vadgama, S., Hesselink, R., der Linden, P. A. V., and Romero, D. W. Fast, expressive se(n) equivariant networks through weight-sharing in position-orientation space. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dPHLbUqGbr.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Ringel Morris, M. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pp. 16–31, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366762. doi: 10.1145/3308561.3353774. URL https://doi.org/10.1145/3308561.3353774.

Bragg, D., Koller, O., Caselli, N., and Thies, W. Exploring collection of sign language datasets: Privacy, participation, and model performance. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371032. doi: 10.1145/3373625.3417024. URL https://doi.org/10.1145/3373625.3417024.

Cormier, K., Brentari, D., and Fenlon, J. *Sign language phonology*. 07 2018.

Crasborn, O., Bank, R., Zwitserlood, I., van der Kooij, E., Ormel, E., Ros, J., Schüller, A., de Meijer, A., van Zuilen, M., Nauta, Y. E., van Winsum, F., and Vonk, M. Ngt dataset in global signbank. 2020a.

Crasborn, O., Zwitserlood, I., van der kooij, E., and Ormel, E. Global signbank manual, version 2. 11 2020b. doi: 10.13140/RG.2.2.16205.67045/1.

Das, S., Biswas, S. K., and Purkayastha, B. A deep sign language recognition system for indian sign language. *Neural Comput. Appl.*, 35(2):1469–1481, sep 2022. ISSN 0941-0643. doi: 10.1007/s00521-022-07840-y. URL https://doi.org/10.1007/s00521-022-07840-y.

Deng, Z., Leng, Y., Chen, J., Yu, X., Zhang, Y., and Gao, Q. Tms-net: A multi-feature multistream multi-level information sharing network for skeleton-based sign language recognition. *Neurocomputing*, 572:127194, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127194. URL https://www.sciencedirect.com/science/article/pii/S0925231223013176.

Desai, A., Berger, L., Minakov, F. O., Milan, V., Singh, C., Pumphrey, K., Ladner, R. E., Daumé, H., Lu, A. X., Caselli, N., and Bragg, D. Asl citizen: a communitysourced dataset for advancing isolated sign language recognition. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024a. Curran Associates Inc.

Desai, A., De Meulder, M., Hochgesang, J. A., Kocab, A., and Lu, A. X. Systemic biases in sign language AI research: A deaf-led call to reevaluate research agendas. In Efthimiou, E., Fotinea, S.-E., Hanke, T., Hochgesang, J. A., Mesch, J., and Schulder, M. (eds.), *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pp. 366–377, Torino, Italy, May 2024b. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL). ISBN 978-2-493814-30-2. URL https://www.sign-lang.uni-hamburg.de/lrec/pub/24045.pdf.

Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., and Giro-i Nieto, X. How2sign: A large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2735–2744, June 2021.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. *Ethnologue: Languages of the World*. SIL International, Dallas, 25 edition, 2022. URL http://www.ethnologue.com.

Forster, J., Schmidt, C., Koller, O., Bellgardt, M., and Ney, H. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1911–1916, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/585_Paper.pdf.

Gao, L., Zhu, L., Xue, S., Wan, L., Li, P., and Feng, W. Multi-view fusion for sign language recognition through knowledge transfer learning. In *Proceedings of the 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI '22, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700316. doi: 10.1145/3574131.3574434. URL https://doi.org/10.1145/3574131.3574434.

Heidari, N. and Iosifidis, A. Geometric deep learning for computer-aided design: A survey, 2024.

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., and Fu, Y. Skeleton aware multi-modal sign language recognition. *CoRR*, abs/2103.08833, 2021. URL https://arxiv.org/abs/2103.08833.

Joksimoski, B., Zdravevski, E., Lameski, P., Pires, I. M., Melero, F. J., Martinez, T. P., Garcia, N. M., Mihajlov, M., Chorbev, I., and Trajkovik, V. Technological solutions for sign language recognition: A scoping review of research trends, challenges, and opportunities. *IEEE Access*, 10:40979–40998, 2022. doi: 10.1109/ACCESS.2022.3161440.

Joze, H. R. V. and Koller, O. MS-ASL: A large-scale data set and benchmark for understanding american sign language. *CoRR*, abs/1812.01053, 2018. URL http://arxiv.org/abs/1812.01053.

Joze, H. R. V. and Koller, O. MS-ASL: A large-scale data set and benchmark for understanding american sign language. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, pp. 100. BMVA Press, 2019. URL https://bmvc2019.org/wp-content/uploads/papers/0254-paper.pdf.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596: 1–11, 07 2021. doi: 10.1038/s41586-021-03819-2.

Kezar, L., Carlin, R., Srinivasan, T., Sehyr, Z., Caselli, N., and Thomason, J. Exploring strategies for modeling sign language phonology, 2023a. URL https://arxiv.org/abs/2310.00195.

Kezar, L., Thomason, J., Caselli, N., Sehyr, Z., and Pontecorvo, E. The sem-lex benchmark: Modeling asl signs and their phonemes. In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '23. ACM, oct 2023b. doi: 10.1145/3597638.3608408. URL http://dx.doi.org/10.1145/3597638.3608408.

Koller, O. Quantitative survey of the state of the art in sign language recognition. *CoRR*, abs/2008.09918, 2020a. URL https://arxiv.org/abs/2008.09918.

Koller, O. Quantitative survey of the state of the art in sign language recognition. volume abs/2008.09918, 2020b. URL https://api.semanticscholar.org/CorpusID:221265990.

Kopf, M., Schulder, M., and Hanke, T. The sign language dataset compendium: Creating an overview of digital linguistic resources. pp. 102–109, June 2022. URL https://aclanthology.org/2022.signlang-1.16.

Leigh, I. W., Andrews, J. F., Harris, R. L., and Gonzalez Avila, T. *Deaf culture : exploring deaf communities in the United States*. Plural Publishing, San Diego, CA, second edition. edition, 2022 - 2022. ISBN 9781635501803.

Li, D., Rodriguez, C., Yu, X., and Li, H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. pp. 1459–1469. The IEEE Winter Conference on Applications of Computer Vision, 2020a.

Li, D., Rodriguez, C., Yu, X., and Li, H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 1459–1469, 2020b.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C., Yong, M. G., Lee, J., Chang, W., Hua, W., Georg, M., and Grundmann, M. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. URL http://arxiv.org/abs/1906.08172.

Meulder, M. D. The legal recognition of sign languages. *Sign Language Studies*, 15(4):498–506, 2015. ISSN 03021475, 15336263. URL http://www.jstor.org/stable/26191000.

Moryossef, A., Tsochantaridis, I., Dinn, J., Camgöz, N. C., Bowden, R., Jiang, T., Rios, A., Müller, M., and Ebling, S. Evaluating the immediate applicability of pose estimation for sign language recognition, 2021.

Murray, J. J. 'the recognition of sign languages in the achievement of deaf people's human rights' side event. https://wfdeaf.org/cosp2020-sideevent/WFD, November 2020. Accessed: 2024-6-14.

Naz, N., Sajid, H., Ali, S., Hasan, O., and Ehsan, M. K. Mipa-resgcn: a multi-input part attention enhanced residual graph convolutional framework for sign language recognition. *Computers and Electrical Engineering*, 112:109009, 2023. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2023.109009. URL https://www.sciencedirect.com/science/article/pii/S0045790623004330.

Nunnari, F., Avramidis, E., España-Bonet, C., González, M., Hennes, A., and Gebhard, P. Dgs-fabeln-1: A multi-angle parallel corpus of fairy tales between german sign language and german text. In *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4847–4857, Torino, Italy, May 2024. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL). URL https://aclanthology.org/2024.lrec-main.434.

Otterspeer, G., Klomp, U., and Roelofsen, F. Signcollect - a 'touchless' pipeline for constructing large-scale sign language repositories. In *LREC-COLING 2024 - 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, Turin, Italy, May 2024.

Padden, C. and Humphries, T. *Inside Deaf Culture*. Harvard University Press, 2005. ISBN 9780674015067. URL http://www.jstor.org/stable/j.ctvjz83v3.

Ranum, O., Otterspeer, G., Andersen, J. I., Belleman, R. G., and Roelofsen, F. 3D-LEX v1.0 – 3D lexicons for American Sign Language and Sign Language of the Netherlands. In Efthimiou, E., Fotinea, S.-E., Hanke, T., Hochgesang, J. A., Mesch, J., and Schulder, M. (eds.), *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pp. 252–263, Torino, Italy, May 2024. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL). ISBN 978-2-493814-30-2. URL https://www.sign-lang.uni-hamburg.de/lrec/pub/24030.pdf.

Rastgoo, R., Kiani, K., and Escalera, S. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021a. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.113794. URL https://www.sciencedirect.com/science/article/pii/S095741742030614X.

Rastgoo, R., Kiani, K., Escalera, S., and Sabokrou, M. Sign language production: A review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3451–3461, June 2021b.

Rezende, T. M., Almeida, S. G. M., and Guimarães, F. G. Development and validation of a brazilian sign language database for human gesture recognition. *Neural Computing and Applications*, 33:10449 – 10467, 2021. URL https://api.semanticscholar.org/CorpusID:233779719.

Rodriguez, J. M., Larson, M., and ten Bosch, L. Exploring the importance of sign language phonology for a deep neural network. *ESANN 2023 proceesdings*, 2023. URL https://api.semanticscholar.org/CorpusID:262058998.

Saunders, B., Camgöz, N. C., and Bowden, R. Skeletal graph self-attention: Embedding a skeleton inductive bias into sign language production. *CoRR*, abs/2112.05277, 2021. URL https://arxiv.org/abs/2112.05277.

Schembri, A. C., Fenlon, J. B., Rentelis, R., Reynolds, S., and Cormier, K. Building the british sign language corpus. volume 7, pp. 136–154. University of Hawaii Press, 2013. URL https://api.semanticscholar.org/CorpusID:55544346.

Sehyr, Z. S., Caselli, N., Cohen-Goldberg, A. M., and Emmorey, K. The ASL-LEX 2.0 Project: A Database of Lexical and Phonological Properties for 2,723 Signs in American Sign Language. 26(2):263–277, 02 2021. ISSN 1081-4159. doi: 10.1093/deafed/enaa038. URL https://doi.org/10.1093/deafed/enaa038.

Selvaraj, P., NC, G., Kumar, P., and Khapra, M. Openhands: Making sign language recognition accessible with pose-based pretrained models across languages, 2021.

Sincan, O. M. and Keles, H. Y. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020. doi: 10.1109/ACCESS.2020.3028072.

Sridhar, A., Ganesan, R., Kumar, P., and Khapra, M. Include: A large scale dataset for indian sign language recognition. pp. 1366–1375, 10 2020. doi: 10.1145/3394171.3413528.

Stokoe, W. *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf*. Studies in Linguistics. Occasional Papers. University of Buffalo, 1960. URL https://books.google.nl/books?id=XGCbPgAACAAJ.

von Agris, U. and Kraiss, K.-F. SIGNUM database: Video corpus for signer-independent continuous sign language recognition. In Dreuw, P., Efthimiou, E., Hanke, T., Johnston, T., Martínez Ruiz, G., and Schembri, A. (eds.), *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pp. 243–246, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL https://www.sign-lang.uni-hamburg.de/lrec/pub/10006.pdf.

Watkins, F., Abdlkarim, D., and Winter, B. e. a. Viewing angle matters in british sign language processing. *Scientific Reports 14, 1043*, 2024. doi: https://doi.org/10.1038/s41598-024-51330-1. URL https://doi.org/10.1038/s41598-024-51330-1.

Wessels, D. R., Knigge, D. M., Papa, S., Valperga, R., Vadgama, S., Gavves, E., and Bekkers, E. J. Grounding continuous representations in geometry: Equivariant neural fields, 2024. URL https://arxiv.org/abs/2406.05753.

Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., and Alikhani, M. Including signed languages in natural language processing. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7347–7360, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.570. URL https://aclanthology.org/2021.acl-long.570.

Zhao, W., Hu, H., Zhou, W., Mao, Y., Wang, M., and Li, H. Masa: Motion-aware masked autoencoder with semantic alignment for sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024. doi: 10.1109/TCSVT.2024.3409728.

A. Training Details Supplements

Table 8 details the training configurations and hyperparameters used in our experiments. We utilized the default parameters for SL-GCN as provided in the OpenHands framework and adhered to the OpenHands standard for the temporal convolution kernel in both models. For the additional parameters of the temporal-PONITA architecture, we conducted a hyperparameter sweep using Optuna, an open-source framework that automates hyperparameter tuning for machine-learning models.

HYPERPARAMETERS	SL-GCN	TEMPORAL-PONITA
EPOCH STRATEGY	EARLY STOPPING	EARLY STOPPING
WARMUP	-	100
BATCH SIZE	32	32
LEARNING RATE	1E-3	5E-3
HIDDEN DIM	64, 128, 256	64
LAYERS	10	6
TEMPORAL CONVOLUTION KERNEL SIZE	9	9
TEMPORAL CONVOLUTION WEIGHT DECAY	-	1E-3
NUMBER OF ORIENTATIONS	-	1
BASIS DIM	-	128
DEGREE OF POLYNOMIAL EMBEDDING	-	1
WIDENING FACTOR	-	4
LAYER SCALE	-	0

Table 8. Hyperparameters used in training of each model. "-" Indicates that the parameter does not apply to this model.

B. Details on the Synthetic Pose Production

The 3D-LEX dataset employs three distinct motion capture systems to accurately capture handshapes, facial expressions, and full-body postures directly from human signers, ensuring high fidelity to signs. Handshapes are recorded using StretchSense gloves, and full-body poses with a Vicon motion capture rig. The dataset includes animation files combining handshapes and full-body poses. For synthetic avatar videos, we extract 200 animation files from 3D-LEX that overlap with the NGT200 vocabulary, retarget this data onto the Ready Player Me avatar, and use it to display the recorded signs.

To produce synthetic multi-view videos using 3D ground truth, we developed a web application using BabylonJS, an open-source web rendering engine. This application retargets motion capture data from 3D-LEX and supports batch processing of motion capture files. Within the app, the uploaded files are played sequentially, and a screen recorder captures videos of the signs from distinct viewpoints. These videos are then downloaded, and MediaPipe is used to extract poses from the synthetic videos. The pipeline, including a screenshot of the Ready Player Me avatar, is shown in Figure6. Our (working) repository is available at https://github.com/J-Andersen-UvA/BabylonSignLab.git, which includes a live demo.
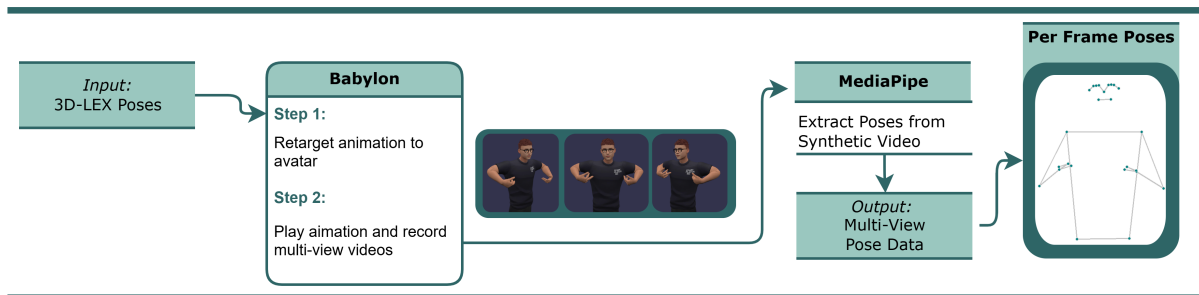

Figure 6. Pipeline for producing the synthetic animation data with Babylons.js and MediaPipe. The chart showcases the multi-view recordings captured of the Ready Player Me Avatar.

C. Pose Quality Validation Supplements

Estimating accuracy in key pose extraction is challenging. To indicate the quality of the Mediapipe pose estimation process, we present the ratio of successful to unsuccessful landmark extractions from the video in Figure 7. Estimating accuracy in key pose extraction is challenging. To assess the quality of the Mediapipe pose estimation process, we present the ratio of successful to unsuccessful landmark extractions from the video in Figure 7. In this figure, successful keypoints are shown in blue, while failed keypoint extractions are shown in orange. Each bar was calculated by summing the total number of zeros occurring in each spatial graph across all consecutive time-frames within a sign, and comparing them to the total number of non-zero occurrences. The rates are sorted by the magnitude of failed keypoints per sign, but may not align across different subplots (e.g., the sign characterized by the first bar in the left view may not correspond to the sign characterized by the first bar in the front view). As observed, keypoint extraction from both human and synthetic data performs well, indicating acceptable dataset quality. Sources of failed keypoints may include occlusion of hands and other body parts, as well as rapid motion across consecutive frames.
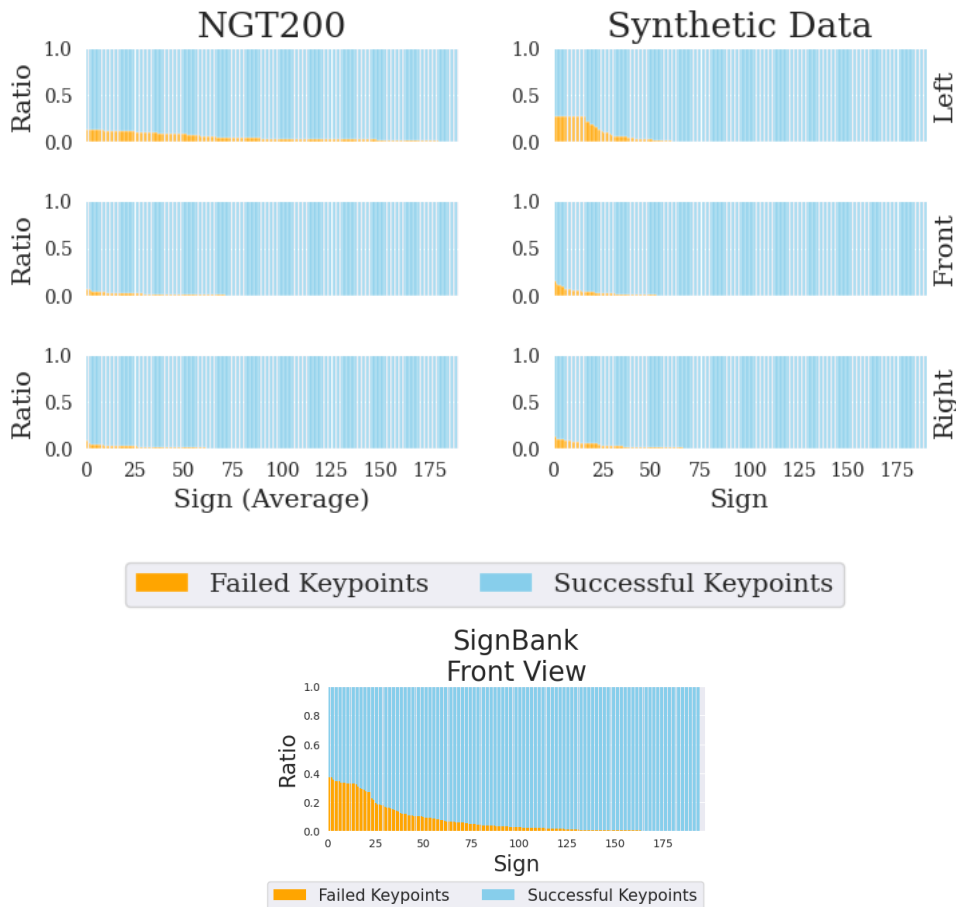

Figure 7. The average ratio of successful to failed keypoint detections for each sign across different views (Left, Front, Right) in both NGT200 and Synthetic datasets. Each bar represents the ratio for an individual sign, sorted by the magnitude of the ratio of failed to successful keypoint detections. The blue bars indicate successful keypoint detections, whereas the orange bars represent instances where keypoints failed, with MediaPipe returning a value of zero.

D. Comparison with Other Isolated Sign Datasets and State-Of-The-Art Methods

Table 9 compares the NGT200 dataset and the temporal-PONITA method with other commonly used isolated sign datasets and ISR State-Of-The-Art (SOTA) methods. We do not list continuous datasets, even though some sentence-level datasets do offer multiple views for advanced studies, such as Duarte et al. (2021) and Forster et al. (2014). Note that this list is not exhaustive, but simply aims to achieve some commonly considered datasets. Among these, we are only familiar that the NGT200 and a subset of WLASL-2000 have access to 3D-ground truth. Other ASL datasets may overlap with the 3D-LEX ASL dataset, making 3D ground truth available, but this has not yet been evaluated.

Table 9. Overview of Datasets Used in ISR Research, and State-Of-The-Art methods. *Pub. Acc.* refers to Publicly Accessible, and indicates whether the dataset is publicly available at the time of publication. In the columns *Vids/Sign* we present the number of videos *P/V* for multi-view datasets, meaning *Videos per Sign per View*. The letter *U* denotes that the information is unknown, and *HH* stands for hard of hearing. If multiple views are present, the SOTA is averaged across all views. Consent is marked as unknown if the paper does not clarify whether informed consent was obtained.

DATASET NAME	PUB. ACC.	SOURCE LANG.	VOCAB. SIZE	VIDS/ SIGN	SIGNERS	# OF VIEWS	COLLECTION METHOD	CONS. ACCESS	SOTA A@1	MODEL
NGT200	✓	NGT	200	12 (4 P/V)	4 DEAF	3	CURATED / LAB	✓	.56	TEMPORAL PONITA*
MVSL (GAO ET AL., 2023)	×	CSL	50	210 (70 P/V)	14 U	3	CURATED / LAB	✓	.95	MVTK (GAO ET AL., 2023)
WLASL-2000 (LI ET AL., 2020B)	✓	ASL	2,000	10.5	119 U	1	SCRAPED	×	.56	TMS-NET (DENG ET AL., 2024)
SEM-LEX (KEZAR ET AL., 2023B)	✓	ASL	3,149	21	41 DEAF	1	CURATED / WEB-CAM	✓	.87	SL-GCN (KEZAR ET AL., 2023A)
ASL CITIZEN (DESAI ET AL., 2024A)	✓	ASL	2,731	30.5	52 DEAF / HH	1	CROWD	✓	.63	I3D (DESAI ET AL., 2024A)
MS-ASL-1000 (JOZE & KOLLER, 2018)	✓	ASL	1,000	25	222 U	1	SCRAPED	×	.70	MASA (ZHAO ET AL., 2024)
INCLUDE (SRIDHAR ET AL., 2020)	✓	INDIAN SL	263	16	7 DEAF	1	CURATED / MIXED SCENES	U	.81	VGG-19 + BILSTM (DAS ET AL., 2022)
AUTSL (SINCAN & KELES, 2020)	✓	TURKISH SL	226	170	42 MIXED[1]	1	CURATED/ MIXED SCENES	✓	.97	TMS-NET (DENG ET AL., 2024)
MINDS-LIBRAS (REZENDE ET AL., 2021)	✓	BRAZILIAN SL	20	60	12 MIXED	U	CURATED / LAB	1	.97	MIPA-RESGCN (NAZ ET AL., 2023)

---

[1]Mixed includes a varying selection of instructors, translators, code, new signers, and trained signers.

E. Model performance comparison

Training times were recorded across 10 runs for both the SL-GCN and temporal-PONITA models trained on all three views. The average time-cost calculations are provided in Table 10. Although the temporal-PONITA model is computationally more demanding, leading to a higher average time per epoch, it is approximately 40% faster in terms of total running time compared to the SL-GCN.

MODEL	SL-GCN	TEMPORAL-PONITA
AVERAGE TIME PER EPOCH	8.0 S	11.5 S
AVERAGE # EPOCHS BEFORE STOPPING	357	145
TOTAL TIME COST	47M	28M

Table 10. Training time details of Temporal-PONITA and the SL-GCN.

Figure 8 shows the training and validation learning curves for 10 runs in experiments involving all three views. Comparing the curves between the two models, temporal-PONITA exhibits a more stable training profile, unlike the oscillating profile of SL-GCN. Additionally, temporal-PONITA converges faster in terms of each global step to approximately the same validation accuracy as SL-GCN. However, as found in Section 6, temporal-PONITA outperforms SL-GCN in test accuracy, indicating better generalization.
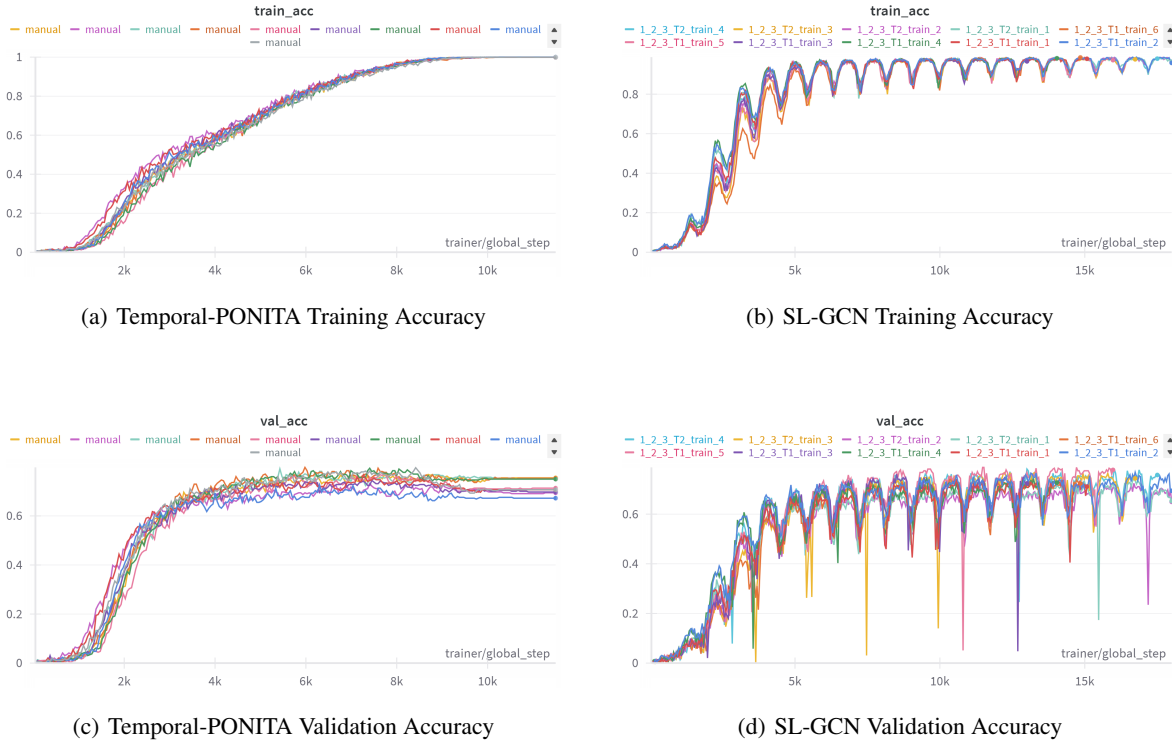

(a) Temporal-PONITA Training Accuracy

(b) SL-GCN Training Accuracy

(c) Temporal-PONITA Validation Accuracy

(d) SL-GCN Validation Accuracy

Figure 8. Comparison of learning performance between temporal-PONITA and the SL-GCN model.