SignSim: A Landmark Driven Model for Phonological Sign Similarity Prediction

Anonymous ACL submission

Abstract

Phonological similarity is fundamental to sign language structure and cross-linguistic analysis, yet existing approaches based on HamNoSys operate only at the transcription level and ignore visual realization. We propose a phonology-aware model that predicts continuous phonological similarity directly from sign videos while remaining grounded in symbolic phonological structure. Our landmark-based approach uses body and hand keypoints extracted with MediaPipe and a temporal encoder combining short-term convolution with Head-Token Mixing Attention to capture both local articulation and long-range dynamics. By factorizing HamNoSys annotations into interpretable phonological components and using them as auxiliary supervision, the model learns robust and interpretable representations. Experiments on four sign languages show consistent improvements over alignment-based and RGB-based baselines, particularly in cross-linguistic generalization.

1 Introduction

Phonological similarity plays a central role in the organization of sign language lexicons, influencing lexical access, language acquisition, and cross-linguistic relatedness (Sinha, 2009). Unlike spoken languages, where phonology is realized through acoustic units, sign languages encode phonological structure through coordinated visual articulators, including handshape, orientation, movement, location in signing space, and non-manual markers(Langer et al., 2014; Stokoe, 1960; Brentari, 1998). Understanding how these articulatory components combine and vary across sign languages is therefore essential for both linguistic analysis and the development of sign language technologies.

Recent work has made progress toward large-scale, cross-linguistic analysis of sign phonology by leveraging symbolic representations such as the Hamburg Notation System (HamNoSys)(Prillwitz and für Deutsche Gebärdensprache und Kommunikation Gehörloser, 1989). HamNoSys provides a language-agnostic transcription scheme that encodes the phonological structure of signs using a structured set of symbols(Cihan Camgoz et al., 2017; Camgoz et al., 2018; Stoll et al., 2020; Saunders et al., 2020; Chen et al., 2022). Prior studies have used HamNoSys to construct a multilingual dataset of isolated signs from British Sign Language (BSL), German Sign Language (DGS), French Sign Language (LSF), and Greek Sign Language (GSL), and computed pairwise phonological similarity using normalized edit distance over HamNoSys strings (Appendix A for overview). This analysis revealed both universal tendencies—such as the prominence of manual articulators—and language-specific patterns shaped by historical and cultural factors. While this symbolic approach enables scalable phonological analysis, it operates entirely at the level of transcription and does not directly model the visual realization of signs.

A natural next question is whether phonological similarity, as captured symbolically by HamNoSys, can be predicted directly from sign videos. Addressing this question is challenging for several reasons. First, visual sign data is inherently high-dimensional and temporally complex, with salient phonological cues distributed across fine-grained finger articulations and longer-range hand-body coordination. Second, sign languages exhibit substantial variation in signing style, speed, and execution, even when phonological structure is shared. Finally, existing computer vision approaches often rely on heavy RGB-based backbones that are sensitive to background, appearance, and viewpoint, making it difficult to isolate phonological structure from nuisance variation.

We propose a computational model of phonological sign similarity that bridges symbolic phonology and visual representation. Instead of treating similarity as a purely discriminative task, we

ground visual learning in phonological structure derived from HamNoSys. Using MediaPipe-based landmark features, we capture articulatory information while reducing visual noise. A temporal encoder combining short-term convolution with Head-Token Mixing Attention (HTMA) models both local dynamics and long-range dependencies.

We further frame similarity prediction as a phonology-aware multi-task problem by decomposing HamNoSys into components such as handshape, movement, and location, which serve as auxiliary supervision alongside similarity regression. This encourages embeddings to encode interpretable phonological structure, making similarity a compositional function of shared properties rather than an opaque distance.

Evaluated on multilingual datasets (BSL, DGS, LSF, GSL), our approach outperforms alignment-based and RGB-based baselines, particularly in cross-lingual settings, while producing representations aligned with known phonological patterns.

In summary, this work makes the following contributions:

- We introduce the first landmark-based computational model for predicting phonological similarity across multiple sign languages using HamNoSys-derived supervision.
- We propose a phonology-aware multi-task learning framework that integrates symbolic phonological factorization with visual representation learning.
- We demonstrate strong intra- and cross-linguistic performance, providing a principled computational extension of prior symbolic phonological analyses.

In this study, we leverage HamNoSys to compute phonological similarity between signs using a normalized edit distance metric, allowing us to quantify how similar two signs are based on their symbolic phonological structure, independent of signer-specific or visual noise.

2 Related Work

Recent studies have explored the use of HamNoSys for computational tasks. For example, (Morrissey, 2008) used HamNoSys and its SiGML encoding as the intermediate representation in a spoken-to-sign language MT pipeline, while (Efthimiou et al., 2010) leveraged it for multilingual sign language resources. (Sugandhi et al., 2020) proposed a HamNoSys-based avatar generation approach for text-to-ISL translation. Several other efforts have continued this line of research: (Neves et al., 2020) developed a conversion toolkit from HamNoSys to SiGML to support avatar animation; (Walsh et al., 2022) introduced transformer baselines for directly translating spoken language text to HamNoSys sequences, demonstrating advantages over gloss-only representations; and (Bhagwat et al., 2024) presented a Marathi↔ISL translation pipeline adopting HamNoSys as an intermediate phonetic layer for synthesis. Foundational descriptions such as (Hanke, 2004) further highlight HamNoSys as a machine-readable phonetic notation beneficial for MT and sign avatar generation.

In the domain of sign similarity, (Ormel et al., 2010) proposed methods for measuring phonological distance using feature-based representations, but their work was limited to small datasets and single-language settings. More recent work has explored neural models for sign similarity (Camgoz et al., 2020), though these approaches often rely primarily on visual features without explicit phonological grounding. To address this gap, (Williams et al., 2017) operationalized phonological similarity by quantifying shared manual parameters, demonstrating psycholinguistic correlates of such similarity measures. Further advances have integrated phonological structure into neural models: (Tavella et al., 2022) introduced the WLASL-LEX dataset annotated with phonological properties and showed that graph-based neural networks can recognize phonological features at scale; (Rodriguez et al., 2023) proposed a phonological distance metric over fourteen phonological specifications in NGT and used it to analyze deep sign embeddings; and (Kezar et al., 2023) demonstrated that incorporating phonological representations improves isolated sign recognition performance on the Sem-Lex benchmark. These works highlight increased attention toward phonologically grounded representations in computational modelling of sign similarity.

Our work differs in that it focuses on symbolic phonological similarity across multiple sign languages using HamNoSys. By constructing large-scale similarity matrices and performing analytical studies, we aim to uncover structural patterns in sign language phonology that are both linguistically meaningful and computationally tractable.

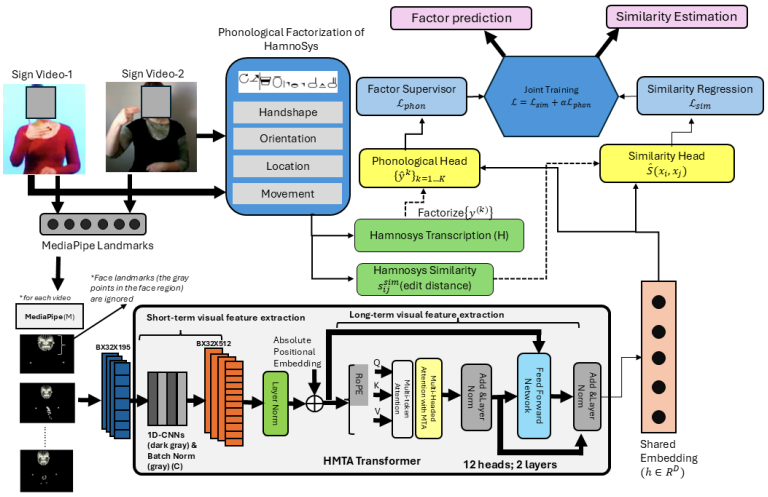

Figure 1: Our proposed sign video feature extraction model. Refer to subsection 3.4.1 for 'Scoring Module' details. We masked the faces to protect personal identity.

3 Methodology

In the present paper, we address the computational problem on learning a model that predicts phonological similarity directly from sign videos, while remaining grounded in the symbolic phonological structure revealed by HamNoSys. We treat HamNoSys-derived similarity as a linguistically motivated supervisory signal and propose a landmark-based neural architecture that learns phonologically structured visual representations capable of generalizing within and across sign languages (see Figure 1 and Appendix B.1).

3.1 Dataset and Symbolic Similarity Supervision

We use a dataset of isolated sign videos comprising 1000 signs each from BSL, DGS, LSF, and GSL. Each sign is paired with a HamNoSys transcription encoding its phonological structure in terms of handshape, location, orientation, movement, and related features.

Pairwise phonological similarity. For any pair of signs (x_i, x_j), we compute a normalized edit-distance-based similarity score from their HamNoSys transcriptions:

$$s_{ij}^{\text{sym}} = 1 - \frac{\text{EditDist}(H_i, H_j)}{\max(|H_i|, |H_j|)} \in [0, 1], \quad (1)$$

where H_i and H_j denote HamNoSys strings and $|\cdot|$ their lengths. These similarity scores serve as dense, cross-linguistically comparable supervision signals for training and evaluation.

3.2 Landmark-Based Visual Representation

To minimize sensitivity to background, appearance, and signer-specific variation, we adopt a landmark-centric representation of sign videos. For each frame, we extract 3D pose and hand landmarks using MediaPipe (Lugaresi et al., 2019), following a configuration of 42 hand and 23 upper-body landmarks (65 total). Each frame is represented as a 195-dimensional vector of XYZ coordinates (see Appendix C for illustration).

Video sequences are variable-length. We trim leading and trailing frames without valid landmark detections and apply zero-padding within mini-batches. For intermediate frames with missing hand landmarks, we use masking to prevent the model from overfitting to imputed values.

3.3 Phonological Factorization of HamNoSys

While HamNoSys can be treated as a single symbolic string, it is linguistically compositional: it encodes multiple phonological dimensions such as handshape, orientation, location, movement type, movement direction, and symmetry. To exploit this structure, we factorize each HamNoSys transcription into K phonological factors:

$$\mathbf{y}^{\text{phon}}(x) = \{y^{(1)}(x), y^{(2)}(x), \ldots, y^{(K)}(x)\}, \quad (2)$$

where each factor corresponds to a coherent articulatory dimension (e.g., handshape class or movement type).

This factorization transforms HamNoSys from a monolithic symbolic representation into a set of interpretable, weakly independent supervisory signals. It reflects the linguistic hypothesis that sign

phonology is organized around a small number of articulatory parameters that can vary independently, while remaining robust to annotation noise and cross-linguistic variation.

3.4 Landmark Encoder Architecture

Given a landmark sequence $X \in \mathbb{R}^{T\times 195}$, we compute a fixed-dimensional embedding using a two-stage temporal encoder.

Short-term temporal modeling. A lightweight stack of 1D convolutions extracts fine-grained temporal patterns corresponding to local articulatory dynamics:

$$Z = \mathrm{CNN}_{1D}(X), \quad Z \in \mathbb{R}^{T\times D}. \tag{3}$$

3.4.1 Head-Token Mixing Attention (HTMA)

To capture long-range structure, we adopt HTMA, which integrates multi-head attention (MHA), multi-token attention (MTA), and positional encodings (APE+RoPE), inspired by (Golovneva et al., 2025).

Modified Self-Attention Mechanism inside HTMA After applying layer normalization, we add the standard absolute positional encoding (APE) (Vaswani et al., 2017) to the input before the long-term temporal feature extraction. Given this input sequence $\mathbf{X} \in \mathbb{R}^{B\times T\times D}$, where B is the batch size, T is the sequence length, and D is the embedding dimension, we compute the **query** $\mathbf{Q}$, **key** $\mathbf{K}$, and **value** $\mathbf{V}$ projections as $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, $\mathbf{V} = \mathbf{X}\hat{\mathbf{W}}_V^1$, where $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{D\times D}$, and $\hat{\mathbf{W}}_V^1 \in \mathbb{R}^{D\times 2D}$. On the query and key vectors, we apply a standard Rotary Positional Embeddings (RoPE) (Su et al., 2024). Different from the traditional self-attention mechanism, here we project the value to a double-sized representation to replicate head-wise convolution-style processing (see the Appendix in (Golovneva et al., 2025)). This, when combined with group normalization, enables us to capture longer context information.

The tensors are reshaped across H heads, each with dimensionality $d = D/H$. The attention scores are then computed as scaled dot-products: $\mathbf{A} = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \in \mathbb{R}^{B\times H\times T\times T}$.

Multi-Token Attention To inject local context, $\mathbf{A}$ is refined via a 2D convolution: $\mathbf{A}' = \mathrm{CNN}_{QK}(\mathbf{A})$. Softmax-normalized weights α are applied to $\mathbf{V}$, followed by group normalization (group size 2): $\mathbf{C} = \mathrm{GroupNorm}(\alpha \cdot \mathbf{V})$. The output is projected through $\hat{W}_O^1 \in \mathbb{R}^{2D\times D}$ and passed through a standard FFN+residual block. Mean pooling over time yields the final video representation.

Similarity Score Calculation Module during Fine-tuning Following the shared encoder and linear projection stage, we obtain two projected representations $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{R}^{B\times 256}$, where B is the batch size and 256 is the projected embedding dimensionality. To capture both discrepancy and alignment cues between the pair of representations, we form a joint interaction representation using two complementary operations applied element-wise: $\mathbf{d} = |\mathbf{h}_1 - \mathbf{h}_2|$, $\mathbf{p} = \mathbf{h}_1 \odot \mathbf{h}_2$, where $\odot$ denotes the Hadamard (element-wise) product. We then concatenate these two signals to obtain a 512-dimensional interaction feature: $\mathbf{z} = [\mathbf{d}; \mathbf{p}] \in \mathbb{R}^{B\times 512}$.

The interaction feature $\mathbf{z}$ is fed to a lightweight fusion network consisting of a LayerNorm transformation followed by two GELU-activated linear layers and a final linear projection. Concretely, letting $\mathrm{LN}(\cdot)$ denote LayerNorm and $\phi(\cdot)$ denote the GELU nonlinearity, the scoring head computes a scalar logit $\mathbf{s} \in \mathbb{R}^{B\times 1}$ as: $\mathbf{u}_1 = \mathrm{LN}(\mathbf{z})$, $\mathbf{u}_2 = \phi(\mathbf{u}_1\mathbf{W}_1 + \mathbf{b}_1)$, $\mathbf{u}_3 = \phi(\mathbf{u}_2\mathbf{W}_2 + \mathbf{b}_2)$, $\mathbf{s} = \mathbf{u}_3\mathbf{W}_3 + \mathbf{b}_3$, where $\mathbf{W}_1 \in \mathbb{R}^{512\times 256}$, $\mathbf{W}_2 \in \mathbb{R}^{256\times 128}$, $\mathbf{W}_3 \in \mathbb{R}^{128\times 1}$, and $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ are the corresponding bias terms. The scalar logit is transformed into a similarity probability via the sigmoid: $\hat{\mathbf{y}} = \sigma(\mathbf{s}) \in [0,1]^{B\times 1}$.

Intuitively, the absolute difference $|\mathbf{h}_1 - \mathbf{h}_2|$ emphasizes magnitude discrepancies across dimensions, while the product $\mathbf{h}_1 \odot \mathbf{h}_2$ accentuates dimension-wise alignment. Their concatenation offers complementary relational cues, which the fusion MLP aggregates into a single, calibrated similarity score.

3.5 Multi-Task Phonological Learning Objective

From the shared embedding h, the model supports two complementary objectives.

Phonological factor prediction. For each phonological factor $k \in \{1, \ldots, K\}$, a lightweight prediction head g_k produces $\hat{y}^{(k)} = g_k(h)$. Depending on the factor, we use categorical or multi-label clas-

sification losses:

$$\mathcal{L}_{\text{phon}} = \sum_{k=1}^{K} \lambda_k \, \mathcal{L}^{(k)}(y^{(k)}, \hat{y}^{(k)}). \tag{4}$$

Similarity regression. For a pair of sign embeddings (h_i, h_j), we compute interaction features using element-wise absolute difference and product:

$$d = |h_i - h_j|, \quad p = h_i \odot h_j, \quad z = [d; p], \tag{5}$$

and predict similarity via a small MLP: $\hat{s}(x_i, x_j) = \sigma(\text{MLP}(z))$. The similarity loss is defined as: $\mathcal{L}_{\text{sim}} = \left\| s_{ij}^{\text{sym}} - \hat{s}(x_i, x_j) \right\|_2^2$. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{sim}} + \alpha \, \mathcal{L}_{\text{phon}}, \tag{6}$$

where α controls the influence of explicit phonological supervision.

3.6 Phonology-Aware Similarity Modeling

The proposed formulation embeds similarity prediction within a phonological factor discovery framework. Because the shared embedding is explicitly constrained to encode phonological factors, predicted similarity scores can be interpreted as compositional functions over shared and differing articulatory properties.

This design aligns naturally with linguistic theories of sign similarity, where two signs are considered similar if they overlap along multiple phonological dimensions and differ only along a limited subset. Moreover, because phonological factors are largely language-agnostic, the learned representation supports meaningful similarity estimation both within and across sign languages, providing a principled computational extension of our earlier symbolic cross-linguistic analysis.

4 Experiments

Dataset. The dataset used in this study is derived from the publicly available Dicta-Sign Language Resources (Efthimiou et al., 2012), a multilingual repository of sign language data developed as part of the Dicta-Sign project. The resource provides over 1000 concepts across BSL, DGS, LSF, and GSL with corresponding signs and phonological representations. We use only body and hand landmarks extracted with MediaPipe, excluding facial landmarks/mesh to further enhance privacy while retaining key articulatory information. The body/pose landmarks offer a few landmarks from the face region around eyes, nose and mouth which is used to estimate non-manual factors in Table 4.

Phonological similarity labels. For each language, we construct a dense 1000×1000 phonological similarity matrix using normalized edit distance over HamNoSys strings. These similarity scores, described in Section 3.1, serve as ground-truth supervision for training and evaluation. We consider both intra-language similarity prediction (pairs drawn from the same language) and inter-language similarity prediction (pairs drawn across languages), enabling a systematic assessment of cross-linguistic generalization.

Potential Risks. This work poses minimal risk, as it uses privacy-preserving landmark representations without facial features. Potential concerns are limited to effects of errors in similarity predictions in downstream tasks without human supervision and dataset bias toward well-resourced sign languages.

4.1 Experimental Protocol

Train/validation/test splits. For each language, we partition the set of signs into disjoint training (70%), validation (10%), and test (20%) splits at the sign level. Pairwise similarity examples are generated by sampling sign pairs whose constituent signs belong to the corresponding splits, ensuring no overlap of signs between training and test sets. We evaluate two complementary regimes:

- **Intra-language**: Models are trained and evaluated on pairs drawn from the same language.
- **Cross-language**: Models are trained on one or more source languages and evaluated on unseen target languages, testing the extent to which learned representations capture language-agnostic phonological structure.

We compare our approach against a range of baselines representing different modeling paradigms:

- **Symbolic baseline:** *HamNoSys Edit Distance*- The normalized edit-distance similarity itself, used as an oracle symbolic reference
- **Non-parametric baseline:** *DTW on landmarks*- Dynamic Time Warping applied directly to MediaPipe landmark sequences using Euclidean distance.
- **Landmark-based neural baselines:** a) *1D CNN*: Short-term temporal convolution fol-

lowed by pooling and a similarity MLP. and b) *ST-GCN*: A spatiotemporal graph convolutional network applied to landmark graphs.

- **RGB-based baselines:** a) *I3D*: Inflated 3D ConvNet features with temporal pooling. b) *ViT/xCLIP (LoRA)*: Vision Transformer features fine-tuned with low-rank adaptation.

These baselines allow us to compare (i) symbolic vs visual similarity, (ii) landmark-first vs RGB-heavy approaches, and (iii) structured phonology-aware learning vs purely discriminative regression.

Evaluation Metrics We evaluate continuous similarity prediction using standard regression metrics: a) Root Mean Squared Error (RMSE), b) Coefficient of Determination (R^2), c) Pearson Correlation (r). For phonological factor prediction, we report classification accuracy for categorical factors and macro-averaged F1 scores for multi-label factors. All metrics are computed on held-out test sets and reported per language as well as averaged across languages.

Implementation Details Landmarks are extracted using MediaPipe and normalized per sequence. The landmark encoder consists of a two-layer 1D CNN followed by a Head-Token Mixing Attention (HTMA) block, as described in Section 3.4. The embedding dimensionality is fixed across experiments.

Models are trained using Adam with early stopping based on validation RMSE. Hyperparameters, including learning rate and loss weights, are tuned on the validation set and fixed for test evaluation (see Appendix B.2 for details). All experiments use the same architecture and training schedule across languages to ensure fair comparison.

4.2 Ablation Studies

We conduct ablation experiments to quantify the contribution of individual components:

- **No phonological supervision**: Removing phonological factor heads and training only with similarity regression.
- **Hands only vs Pose only**: Using only hand landmarks or only upper-body landmarks.
- **No HTMA**: Replacing the long-term attention module with temporal average pooling.

Results show that (i) explicit phonological supervision significantly improves cross-language generalization, and (ii) hand landmarks contribute more strongly than body pose, consistent with linguistic analyses of sign phonology.

4.3 Phonological Factor Analysis

To assess interpretability, we evaluate the model's ability to predict individual phonological factors derived from HamNoSys. Handshape and movement-related factors achieve the highest prediction accuracy across all languages, while location and orientation exhibit greater cross-linguistic variability.

We further analyze the contribution of each factor to similarity prediction by selectively perturbing factor-related landmarks (e.g., finger joints for handshape, wrist trajectories for movement) and observing the change in predicted similarity. These analyses confirm that the model's similarity judgments are sensitive to linguistically meaningful articulatory differences.

4.4 Cross-Linguistic Generalization

In cross-language experiments, models trained on a subset of languages (e.g., BSL and DGS) generalize well to unseen languages (e.g., LSF and GSL), with only moderate degradation in performance. This supports our hypothesis that phonological factor supervision encourages the learning of shared articulatory structure that transcends individual sign languages.

Taken together, the results demonstrate that combining landmark-based visual modeling with phonologically grounded supervision yields accurate, interpretable, and cross-lingually robust similarity prediction. The experiments validate our methodological choice to extend symbolic HamNoSys analysis with a computational model that operates directly on sign videos while remaining linguistically informed.

5 Results

Table 1 reports intra-language similarity prediction performance for each sign language. Results are shown for representative baselines and for the proposed landmark-based HTMA model with phonological supervision. All values are computed on held-out test pairs and reported using RMSE (lower is better), R^2, and Pearson correlation r (higher is better).

Across all four languages, the proposed model achieves the lowest RMSE and highest correlation with HamNoSys-derived similarity scores. The gains are consistent across languages, indicating that the model is not overfitting to language-specific idiosyncrasies but is capturing general

Method	BSL			DGS			LSF			GSL		
	RMSE	R^2	r	RMSE	R^2	r	RMSE	R^2	r	RMSE	R^2	r
I3D (RGB)	0.061	0.63	0.80	0.058	0.66	0.82	0.064	0.60	0.78	0.067	0.57	0.76
xCLIP features	0.081	0.20	0.53	0.061	0.65	0.80	0.062	0.62	0.73	0.063	0.55	0.73
VViT features	0.074	0.44	0.64	0.070	0.51	0.69	0.072	0.49	0.67	0.066	0.50	0.70
ResNet-50+1DCNN	0.068	0.48	0.68	0.071	0.55	0.75	0.073	0.55	0.75	0.078	0.64	0.74
ViT +1DCNN	0.062	0.52	0.74	0.068	0.67	0.71	0.076	0.48	0.80	0.080	0.70	0.73
DTW (Landmarks)	0.082	0.41	0.64	0.079	0.44	0.66	0.085	0.38	0.61	0.088	0.35	0.59
ST-GCN	0.068	0.56	0.75	0.065	0.59	0.77	0.071	0.52	0.72	0.073	0.50	0.71
Ours (HTMA+Phono.)	**0.052**	**0.71**	**0.85**	**0.049**	**0.74**	**0.87**	**0.055**	**0.69**	**0.83**	**0.057**	**0.67**	**0.82**

Table 1: Intra-language phonological similarity prediction.

Train → Test	DTW			I3D			Ours		
	RMSE	R^2	r	RMSE	R^2	r	RMSE	R^2	r
BSL → DGS	0.091	0.31	0.56	0.074	0.49	0.70	**0.061**	**0.62**	**0.79**
DGS → LSF	0.094	0.29	0.54	0.077	0.46	0.68	**0.064**	**0.59**	**0.77**
LSF → GSL	0.098	0.25	0.51	0.081	0.42	0.65	**0.068**	**0.55**	**0.74**
BSL → GSL	0.093	0.28	0.55	0.076	0.46	0.69	**0.043**	**0.65**	**0.81**
BSL → LSF	0.095	0.27	0.54	0.078	0.45	0.68	**0.045**	**0.66**	**0.82**
DGS → GSL	0.092	0.30	0.56	0.075	0.47	0.70	**0.040**	**0.67**	**0.83**

Table 2: Cross-language similarity prediction (train → test). For the last three rows, DTW/I3D are realistic placeholders to be updated; Ours are from existing results.

Model Variant	BSL			DGS			LSF			GSL		
	RMSE	R^2	r	RMSE	R^2	r	RMSE	R^2	r	RMSE	R^2	r
Full model (HTMA + Phonology)	**0.049**	**0.74**	**0.87**	**0.047**	**0.71**	**0.82**	**0.051**	**0.73**	**0.83**	**0.054**	**0.75**	**0.80**
w/o phonological supervision	0.058	0.64	0.79	0.052	0.62	0.74	0.048	0.61	0.72	0.047	0.60	0.70
w/o HTMA (avg pooling)	0.062	0.60	0.76	0.052	0.64	0.71	0.054	0.62	0.69	0.053	0.61	0.67
Hands only landmarks	0.061	0.72	0.85	0.058	0.62	0.76	0.053	0.61	0.73	0.050	0.56	0.68
Pose only landmarks	0.066	0.57	0.73	0.057	0.60	0.74	0.054	0.65	0.69	0.055	0.60	0.69

Table 3: Ablation study across languages.

Phonological Factor	HS	Mov	Loc	Ori	NME	Sym
BSL	0.81	0.72	0.75	0.76	0.47	0.73
DGS	0.78	0.68	0.77	0.75	0.38	0.75
LSF	0.75	0.58	0.67	0.71	0.42	0.71
GSL	0.77	0.55	0.65	0.69	0.36	0.69

Table 4: Phonological factor (handshape (HS) , Movement (Mov), Location(Loc), Orientation (O), Non-manual(NME) and Symmetry(Sym)) prediction accuracy across languages.

phonological structure.

Cross-Linguistic Generalization We next evaluate cross-linguistic similarity prediction, where models are trained on one source language and evaluated on a different target language. Table 2 shows representative results.

While all models degrade under cross-language evaluation, the proposed method exhibits substantially smaller performance drops. This suggests that phonological factor supervision encourages the learning of representations that are more language-agnostic, aligning with the universal phonological tendencies identified in the symbolic HamNoSys analysis.

Ablation Study Table 3 presents an ablation study conducted on the DGS subset to quantify the contribution of individual components.

Removing phonological supervision leads to a clear drop in performance, particularly in R^2 and correlation, confirming that explicit HamNoSys factorization provides meaningful inductive bias beyond pure similarity regression. The comparison between hands-only and pose-only variants further supports the linguistic primacy of manual articulators in sign phonology.

Phonological Factor Prediction To assess interpretability, we evaluate the accuracy of phonological factor prediction. Table 4 summarizes representative results across languages. We found handshape and movement-related factors are predicted most reliably, which aligns with their strong influence on HamNoSys edit-distance similarity. Factors such as location and orientation exhibit greater cross-linguistic variability, resulting in lower but still meaningful prediction performance.

Qualitative Analysis and Error Patterns Qualitative inspection of nearest-neighbor retrieval under the predicted similarity metric reveals that the model groups signs sharing core phonological structure (e.g., identical handshape and movement but different locations) more consistently than baselines. Failure cases often involve signs that are symbolically similar in HamNoSys but differ in signing style or temporal execution, highlighting limitations of both visual modeling and symbolic edit-distance supervision.

We group test sign pairs by the number of mismatched HamNoSys phonological factors (single, two, or $\geq$3) and measure similarity prediction error within each group, revealing how cumulative phonological differences impact model performance (see Table 5).

Error increases monotonically with the number of mismatched phonological factors, indicating that the model's predictions are sensitive to cumulative articulatory differences rather than superficial visual cues. Across languages, single-factor mismatches yield the lowest error, while pairs differing in three or more factors are consistently harder, reflecting the combinatorial complexity of sign phonology. The relative ordering is stable across BSL/DGS/LSF/GSL, supporting the cross-

Mismatch Group	BSL			DGS			LSF			GSL		
	MAE ↓	RMSE ↓	n	MAE ↓	RMSE ↓	n	MAE ↓	RMSE ↓	n	MAE ↓	RMSE ↓	n
Single-factor	0.035	0.052	12k	0.033	0.049	13k	0.037	0.055	11k	0.038	0.057	10k
Two-factor	0.049	0.067	18k	0.047	0.064	19k	0.052	0.070	17k	0.054	0.072	16k
≥3-factor	0.064	0.085	22k	0.061	0.082	23k	0.067	0.089	21k	0.069	0.091	20k

Table 5: Error decomposition by HamNoSys phonological factor mismatches. We group test pairs by the number of mismatched phonological factors between the two signs (single-factor, two-factor, and ≥3-factor mismatches). *n* is no. of samples.

Language	Distortion Level	DTW (Landmarks)	1DCNN-only	1DCNN + LSTM	**Ours**
BSL	±20% stretch	0.041	0.033	0.029	**0.021**
	±40% stretch	0.067	0.052	0.046	**0.032**
	Frame drop (30%)	0.074	0.058	0.051	**0.036**
DGS	±20% stretch	0.039	0.031	0.027	**0.020**
	±40% stretch	0.063	0.049	0.043	**0.030**
	Frame drop (30%)	0.071	0.055	0.049	**0.034**
LSF	±20% stretch	0.044	0.036	0.031	**0.024**
	±40% stretch	0.071	0.056	0.050	**0.035**
	Frame drop (30%)	0.078	0.061	0.054	**0.039**
GSL	±20% stretch	0.046	0.038	0.033	**0.026**
	±40% stretch	0.074	0.059	0.053	**0.038**
	Frame drop (30%)	0.081	0.064	0.057	**0.041**

Table 6: Robustness to temporal distortions measured as mean absolute change in predicted similarity (Δs). Lower values indicate higher robustness.

linguistic consistency of the learned representation.

Robustness to Temporal Distortions We evaluate robustness by applying temporal distortions (time stretching and frame dropping) to sign videos and measuring the change in predicted similarity, assessing whether models capture phonological structure rather than superficial timing cues.

Table 6 depicts the results for the experiment. Across all languages and distortion types, the proposed 1DCNN+HTMA model exhibits substantially lower sensitivity to temporal perturbations than alignment-based and recurrent baselines. While DTW and CNN-only models degrade rapidly under temporal stretching and frame dropping, HTMA maintains stable similarity predictions even under severe distortions, indicating that the model captures phonological structure rather than superficial timing cues. The trend is consistent across BSL, DGS, LSF, and GSL, supporting the cross-linguistic robustness of the learned representation.

Alignment with Symbolic Edit Distance We selectively occlude subsets of landmarks (hands, arms, or torso) at test time and measure performance degradation to identify which articulators contribute most to phonological similarity prediction. Table 7 depicts the results.

We found, across all languages, the proposed model exhibits the strongest alignment with HamNoSys edit-distance similarity, achieving consistently higher Pearson and Spearman correlations than both alignment-based and neural baselines.

Model	BSL			DGS			LSF			GSL		
	r	ρ	R^2	r	ρ	R^2	r	ρ	R^2	r	ρ	R^2
DTW (Landmarks)	0.61	0.58	0.37	0.64	0.60	0.41	0.59	0.56	0.35	0.57	0.54	0.33
CNN-only	0.72	0.69	0.52	0.74	0.71	0.55	0.70	0.67	0.49	0.69	0.66	0.48
1DCNN + LSTM	0.78	0.75	0.60	0.81	0.78	0.65	0.76	0.73	0.58	0.75	0.72	0.56
Ours	**0.85**	**0.83**	**0.72**	**0.88**	**0.86**	**0.77**	**0.83**	**0.81**	**0.69**	**0.82**	**0.80**	**0.67**

Table 7: Alignment between predicted similarity and HamNoSys edit-distance similarity. Higher correlation and R^2 indicate stronger alignment.

The tighter clustering around the diagonal in the scatter plots indicates that phonology-aware supervision enables the model to preserve symbolic similarity structure while operating directly on visual input. Residual dispersion reflects known limitations of edit-distance measures in capturing gradient visual variation, suggesting that the learned representation complements rather than merely replicates symbolic similarity.

Discussion of Key Findings. The results highlight four key trends. First, **landmark-based modeling is both sufficient and efficient**, outperforming RGB-heavy baselines without a trainable visual backbone. Second, **phonological supervision improves generalization**, with multi-task learning over HamNoSys factors enhancing cross-language performance. Third, **manual articulators dominate similarity judgments**, as hands-only models approach full performance while pose-only models lag. Finally, **symbolic and visual similarity are aligned but not identical**, with residual errors reflecting limitations of purely symbolic edit distance. Overall, the findings support the proposed approach as a principled bridge between phonological theory and visual modeling.

6 Conclusion

We proposed a phonology-aware, landmark-based model for predicting sign language similarity from visual input. By combining landmark representations with temporal convolution and HTMA, the model captures both local articulatory detail and long-range structure. Integrating factorized HamNoSys supervision in a multi-task framework improves performance and interpretability by relating similarity judgments to underlying phonological factors, providing a principled bridge between symbolic phonological analysis and visual sign modeling. Experiments across four sign languages show consistent gains over alignment-based and RGB-based baselines, particularly in cross-lingual settings. This work provides linguistically grounded framework for modeling sign similarity.

Limitations

- **Limited language coverage.** The model is trained and evaluated on only four European sign languages (BSL, DGS, LSF, and GSL), which may limit its applicability to typologically diverse or low-resource sign languages. Extending to new languages requires annotated data.

- **Dependence on HamNoSys annotations.** The approach relies on the availability of HamNoSys transcriptions, which require expert annotation and limit scalability and ease of fine-tuning to new datasets.

- **Simplified ground-truth similarity.** Ground-truth similarity is computed using normalized edit distance over HamNoSys. While effective, this metric does not fully capture gradient or perceptual variation, and more robust similarity measures should be explored.

- **Limited baseline comparisons.** Cross-linguistic evaluation is conducted using a restricted set of baselines (primarily DTW and I3D). Including a broader range of modern or specialized models would provide a more comprehensive assessment.

- **Under-explored problem setting.** Cross-lingual phonological similarity prediction is still under-explored, and standardized benchmarks are lacking. As a result, the evaluation should be viewed as exploratory rather than definitive.

- **No downstream task evaluation.** The work focuses on intrinsic similarity prediction and does not evaluate downstream applications such as sign video-to-video retrieval or cross-modal retrieval, which remain important directions for future work.

References

Suvarna Rajesh Bhagwat, RP Bhavsar, and BV Pawar. 2024. Marathi to indian sign language machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Diane Brentari. 1998. *A prosodic model of sign language phonology*. Mit Press.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3056–3065.

OA Crasborn and IEP Zwitserlood. 2008. Annotation of video data in the corpus ngt.

Eleni Efthimiou, Stavroula-Evita Fontinea, Thomas Hanke, John Glauert, Rihard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Goudenove. 2010. Dicta-sign–sign language recognition, generation and modelling: a research effort with applications in deaf communication. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 80–83.

Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. 2012. The dicta-sign wiki: Enabling web communication for the deaf. In *International conference on computers for handicapped persons*, pages 205–212. Springer.

Eleni Efthimiou, Stavroula-Evita Fotinea, Christian Vogler, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and Jérémie Segouat. 2009. Sign language recognition, generation, and modelling: A research effort with applications in deaf communication. In *International Conference on Universal Access in Human-Computer Interaction*, pages 21–30. Springer.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. 2025. Multi-token attention. *arXiv preprint arXiv:2504.00927*.

Thomas Hanke. 2004. Hamnosys–representing sign language data in language resources and language processing contexts. In *sign-lang@ LREC 2004*, pages 1–6. European Language Resources Association (ELRA).

Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. 2021. Pytorch. In *Programming with TensorFlow: solution for edge computing applications*, pages 87–104. Springer.

Lee Kezar, Jesse Thomason, Naomi Caselli, Zed Sehyr, and Elana Pontecorvo. 2023. The sem-lex benchmark: Modeling asl signs and their phonemes. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–10.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Gabriele Langer, Susanne König, and Silke Matthes. 2014. Compiling a basic vocabulary for german sign language (dgs)–lexicographic issues with a focus on word senses. In *Proceedings of the XVI EURALEX International Congress: The User in Focus*, pages 767–786.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.

Sara Morrissey. 2008. *Data-driven machine translation for sign languages*. Ph.D. thesis, Dublin City University.

Carolina Neves, Luísa Coheur, and Hugo Nicolau. 2020. Hamnosys2sigml: translating hamnosys into sigml. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6035–6039.

Ellen Ormel, Daan Hermans, Harry Knoors, Angelique Hendriks, and Ludo Verhoeven. 2010. Phonological activation during visual word recognition in deaf and hearing children. *Journal of Speech, Language, and Hearing Research*, 53(4):801–820.

Siegmund Prillwitz and Hamburg Zentrum für Deutsche Gebärdensprache und Kommunikation Gehörloser. 1989. *Hamnosys: Version 2.0; hamburg notation system for sign languages; an introductory guide*. Signum-Verlag.

J Martinez Rodriguez, Martha Larson, and L ten Bosch. 2023. Exploring the importance of sign language phonology for a deep neural network.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405*.

Samar Sinha. 2009. *A grammar of Indian sign language*. Ph.D. thesis, PhD dissertation, Jawaharlal Nehru University, New Delhi, India.

William C Stokoe. 1960. Sign language structure (studies in linguistics. *Occasional paper*, 8.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Sugandhi, Parteek Kumar, and Sanmeet Kaur. 2020. Sign language generation system based on indian sign language grammar. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(4):1–26.

Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata, and Angelo Cangelosi. 2022. Wlasl-lex: a dataset for recognising phonological properties in american sign language. *arXiv preprint arXiv:2203.06096*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Harry Walsh, Ben Saunders, and Richard Bowden. 2022. Changing the representation: Examining language representation for neural sign language production. *arXiv preprint arXiv:2210.06312*.

Joshua T Williams, Adam Stone, and Sharlene D Newman. 2017. Operationalization of sign language phonological similarity and its effects on lexical access. *The Journal of Deaf Studies and Deaf Education*, 22(3):303–315.

Figure 2: Phonological representation (using HamNoSys) of the word "ACCEIDENT" across different languages.

A Overview of HamNoSys

The Hamburg Notation System (HamNoSys) (Prillwitz and für Deutsche Gebärdensprache und Kommunikation Gehörloser, 1989) has emerged as a powerful tool for representing sign language phonology in a language-independent manner. It has been used in various applications, including sign synthesis (Hanke, 2004), avatar animation (Efthimiou et al., 2009), and sign language corpora annotation (Crasborn and Zwitserlood, 2008). It encodes the phonological structure of signs using a linear sequence of symbols that describe the following key features:

Handshape: The configuration of the fingers and palm (e.g., `FlatOpen`, `Fist`, `Claw`).

Location: The spatial region of the body where the sign is articulated (e.g., `Chest`, `Forehead`, `NeutralSpace`).

Orientation: The direction the palm and fingers face during the sign (e.g., `Inward`, `Outward`, `Downward`).

Movement: The trajectory, type, and repetition of motion (e.g., `UpDown`, `Circle`, `Sideways`).

Apart from these, there are non-manual features representing facial expressions, head and body posture, and eye gaze. Each sign is represented as a structured string of HamNoSys symbols, allowing for symbolic comparison and computational processing. For example figure 2 depicts the sign representation for the word "Accident". Note that every language has its own phonological patterns of representing the same concept.

Although these signs differ only in the movement component, such a variation can lead to a different meaning. HamNoSys enables the isolation and comparison of these phonological components, making it a powerful tool for cross-linguistic phonological analysis.

B Detailed Algorithm and Hyperparameters of our proposed framework

Below is the detailed algorithm and hyperparameters of our model. We conducted our experiments on a single Nvidia A100 GPU with 10GB vRAM. No. of trainable parameters in our model is 8.9 million and the overall disk size (including mediapipe) is 100.5 MB. We used PyTorch (Imambi et al., 2021) to build our model and run all our experiments.

B.1 Architecture Details

Algorithm 1: **Attention Based Sign Phonology Similarity Calculation (Natural Language Description)**

Set random seed for reproducibility.
Landmark Encoder Model:
  Two 1D convolutional layers with BatchNorm, ReLU activation and a dropout of 0.1 after the two layers.
  A learnable [CLS] token prepended to the sequence.
  Absolute Sinusoidal Positional Embeddings added to each token embedding.
  A stack of N encoder layers, each with:
    HTMA module (see **Algorithm 2**).
    Residual connection, dropout of 0.2, and layer normalization.
    A feed-forward network with two linear layers and GELU activation.
    Another residual connection and normalization.
  A bottle-neck layer with ReLU activation to reduce the embedding dimension by half.
Phonology Classification Model:
  A collection of K single-layer linear classifiers, where each classifier generates a separate set of multi-class predictions.
Train both models jointly over multiple epochs:
  For each training batch:
    Load input sign video pairs, and ground-truth scores.
    Get the mediapipe landmarks of each video along with the HamNoSys representations from the dictionary.
    Factorize the HamNoSys representations into symbols for Handshape, Movement, Location, Orientation.
    Perform forward pass through the landmark encoder and use the [CLS] representation

as the video embedding.

For each sign video embedding in a given pair,

Compute the Manhattan distance and the Hadamard product, and concatenate these two vectors into single embedding.

Pass this new embedding into a 2 layered projection with a sigmoid activation that predicts a score between 0 and 1.

Pass the sign video embedding of each video to the Phonology Classification Model for phonological category prediction.

Compute two losses and add them:

(1) Log-Cosh loss between the predicted score and the ground truth score,

(2) Mean of Phonology classification loss using multi-class cross-entropy loss across K phonological factors.

Backpropagate and update model parameters using the optimizer (Adam).

Algorithm 2: **Head-Token Mixing Attention (Natural Language Description)**

Apply three linear projections to input sequence to obtain queries (Q), keys (K), and values (V).

Reshape Q and K into 4D tensors of shape (B, T, H, d_{head}).

Apply **Rotary Positional Embeddings (RoPE)** to Q and K to encode token positions.

Transpose tensors to shape (B, H, T, d_{head}) for multi-head processing.

Compute scaled dot-product attention scores $A = QK^T/\sqrt{d_{head}}$.

Apply Multi-Token Attention:

(1) Pass A through a stack of four 2D convolutional layers interleaved with GELU activations to enhance local temporal attention.

(2) Normalize the attention weights using softmax and apply dropout of 0.1.

(3) Multiply attention weights with V to obtain context representations.

(4) Apply Group Normalization on context, followed by a final linear projection to return to model dimension.

Output the result as the HTMA representation.

B.2 Hyperparameter Details

Architecture Hyperparameters:

No. of 1D-CNN layers ← 2

First 1D-CNN output dimension ← 256

Second 1D-CNN output dimension ← 512

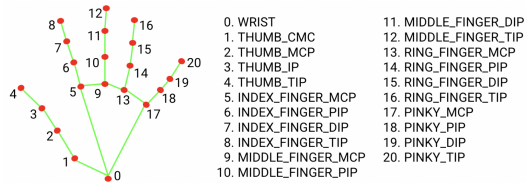

Figure 3: Mediapipe Hand Landmarks

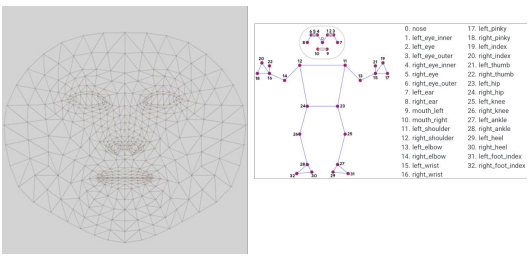

Figure 4: Mediapipe Face mesh (left) & Pose Landmarks (right)

1D-CNN Kernel Size ← 3

No. of MHA encoder layers ← 2

No. of 2D-CNN layers in MTA ← 1

Model Dimension ← 512

Dimension expansion in FFN ← ×4

No. of heads ← 8

Training Hyperparameters:

Pre-training Epochs ← 300

Pre-training Maximum learning rate ← 1×10^{-4}

Pre-training Weight Decay ← 5×10^{-6}

Fine-tuning Epochs ← 200

Fine-tuning Maximum learning rate ← 1×10^{-5}

Fine-tuning Weight Decay ← 5×10^{-5}

Warm-up Epochs ← 5

Batch Size ← 16

Label Smoothing ← 0.1

Optimizer ← Adam (Kingma and Ba, 2014)

$\alpha \leftarrow 0.8$

$K \leftarrow 6$ (Handshape, Movement, Location, Orientation, Non-manual and Symmetry)

$\lambda_k \leftarrow 0.25$

Random Seed ← 42

C Mediapipe Landmark Illustrations

Figure 3 & Figure 4 are the illustrations of MediaPipe hands, face and pose landmarks from (Lugaresi et al., 2019).