A Video-Based Reverse Dictionary for Sign Language Using Gesture Similarity

Batyrbek Orazumbekov, Daniyal Bayanov, Aruzhan Kaltay, Anara Sandygulova
School of Engineering and Digital Sciences, Nazarbayev University, Astana, Kazakhstan
{batyrbek.orazumbekov, daniyal.bayanov, aruzhan.kaltay, anara.sandygulova}@nu.edu.kz

Abstract

Sign language recognition systems are usually modeled as classification systems that map gesture videos to pre-defined glosses. But these systems do not allow similarity searches, where a user can search for similar gestures without knowing the corresponding gloss. This paper presents a pose-based video-to-video search framework for isolated signs, which acts as a reverse gesture dictionary. The system employs keypoints on the skeletal structure instead of RGB images. Two architectures are proposed for modeling temporal information: an encoder with self-attention in a Transformer architecture and a Spatial-Temporal Graph Convolutional Network (ST-GCN). The embedding space is optimized using metric learning objectives, including supervised contrastive learning and ArcFace angular margin loss. The performance of the retrieval system is evaluated on the WLASL dataset using ranking metrics like Recall@K and mean Average Precision (mAP). Experiments reveal that the temporal modeling using the Transformer architecture is an improvement over the graph-based modeling approach in the low-shot learning scenario. The attention-based temporal pooling approach further enhances the ranking quality, with the best-performing model achieving an mAP of 0.237 on the WLASL validation set. Cross-dataset evaluation on a 226-label AUTSL dataset reveals non-trivial generalization performance on the unseen dataset, despite training only on the WLASL dataset.

Keywords: gesture retrieval, sign language technology, metric learning, pose-based representation, cross-dataset generalization

1. Introduction

Sign languages are visual languages that are used by Deaf and hard-of-hearing people. They involve the use of hand and body movements to create meaning. Recent breakthroughs in computer vision and deep learning have been used extensively in the development of automatic sign language recognition systems. Most of the existing literature is based on classification tasks, where a video is classified into a predefined label (Li et al., 2020). Although these systems work well for translation and transcription tasks, they are not intended for similarity search.

In many real-world situations, a user may observe a gesture without knowing its corresponding gloss. In such cases, text-based queries are not possible. This motivates the concept of a reverse gesture dictionary: a system that takes a gesture video as input and retrieves visually similar gestures from a database (Hassan et al., 2025). Unlike classification, which produces discrete predictions, retrieval requires learning a continuous embedding space where similar gestures are located near each other.

Although cross-modal retrieval between sign video and text has been explored (Duarte et al., 2022; Cheng et al., 2023), purely visual video-to-video retrieval remains relatively underexplored. Many current approaches rely on textual supervision or gloss alignment (Duarte et al., 2022). Learning similarity directly from motion therefore requires robust visual representations and metric learning strategies capable of structuring embedding geometry.

The central problem addressed in this work is the development of a system that retrieves visually similar isolated signs given a query video. Unlike classification models, which optimize label prediction, retrieval models must learn embeddings where intra-class samples are compact and inter-class samples are well separated (Ghojogh et al., 2022; Cakir et al., 2019). This task is challenging due to variability across signers, motion speed differences, camera perspectives, and limited samples per gloss. In datasets such as WLASL, only three to four instances per class are available on average, creating a low-shot learning scenario (Li et al., 2023; Musgrave et al., 2020).

To address these challenges, this work proposes a pose-based gesture similarity learning framework. Instead of using raw RGB frames, structured skeletal keypoints are extracted to reduce background noise and appearance variability (Parian-Scherb et al., 2024). Temporal modeling is performed to encode motion dynamics into fixed-dimensional embeddings (Ramanathan et al., 2015). Two architectures are investigated: a Transformer-based model using self-attention to capture global temporal dependencies, and a Spatial-Temporal Graph Convolutional Network (ST-GCN) that models skeletal structure as a graph.

The embedding space is optimized using metric learning objectives, including supervised con-

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 398–407

trastive loss (Khosla et al., 2020) and ArcFace angular margin loss (Deng et al., 2019). Retrieval performance is evaluated using ranking-based metrics such as Recall@K and mean Average Precision (mAP), which align with the objectives of similarity search (Cakir et al., 2019).

The contributions of this work are threefold: (1) a pose-based video-to-video retrieval framework for isolated signs, (2) a comparative analysis of Transformer and ST-GCN temporal modeling under low-shot conditions, and (3) an empirical study of metric learning objectives for gesture similarity. The results demonstrate that purely visual gesture retrieval is feasible and effective without reliance on textual supervision, providing a foundation for reverse sign language dictionary systems and future scalable retrieval research.

In addition to the primary WLASL experiments, this research also assesses cross-dataset generalization by applying the best model to an AUTSL subset with 226 unseen labels, showing the effectiveness of the learned embedding on a different dataset (Sincan and Keles, 2020).

2. Literature Review

2.1. From Sign Recognition to Retrieval-Oriented Systems

Sign language technology has evolved significantly over the past decade. Early systems primarily addressed gesture recognition as a classification problem, where an input video is mapped to a predefined gloss label. Large-scale datasets such as WLASL (Li et al., 2020) enabled benchmarking of deep learning architectures for word-level recognition. These systems demonstrated strong performance in transcription and translation tasks but were inherently limited to closed-set classification.

However, recognition-based systems do not support similarity-driven exploration. In practical scenarios, users may observe a gesture without knowing its gloss. This limitation motivated the development of reverse sign language dictionaries, where users provide a video query and retrieve visually similar gestures. Such systems shift the problem from discrete prediction to continuous similarity modeling.

Earlier work on sign language retrieval predates deep learning-based embedding approaches. For example, the Dicta-Sign project (Efthimiou et al., 2012) introduced a search-by-example interface that allowed users to perform a sign using a Kinect device and retrieve corresponding or closest matching signs from a multilingual lexicon. While enabling early sign lookup based on motion input, these approaches relied on handcrafted features and did not incorporate learned embedding representations for scalable similarity modeling.

Recent advances reflect this transition toward retrieval-based paradigms. Duarte et al. (Duarte et al., 2022) introduced a cross-modal retrieval system that aligns visual sign embeddings with textual queries through joint embedding spaces. Their approach enabled bidirectional search between sign and spoken language. Similarly, Cheng et al. (Cheng et al., 2023) proposed Cross-lingual Contrastive Learning (CiCo), integrating sign and spoken languages into a shared vector space. CiCo demonstrated strong performance on multilingual sign datasets.

More recent work has further advanced cross-modal alignment between sign language and text. Jiang et al. (Jiang et al., 2024) proposed SignCLIP, a contrastive learning framework inspired by CLIP that learns a shared embedding space between sign videos and textual descriptions. This enables efficient text-to-video and video-to-text retrieval. However, such approaches rely on linguistic supervision and primarily optimize for semantic similarity rather than visual similarity between sign executions.

While these systems advanced retrieval capabilities, they remain largely dependent on textual supervision. Purely visual video-to-video retrieval, where similarity is learned directly from motion rather than gloss alignment, remains underexplored. Addressing this gap requires robust motion representation and metric learning techniques specifically optimized for ranking rather than classification.

2.2. Motion Representation and Temporal Modeling

Accurate gesture retrieval depends on how motion is represented and encoded. Most modern systems rely on landmark-based pose extraction, using frameworks such as MediaPipe or OpenPose to capture upper-body and hand keypoints. By focusing on skeletal trajectories instead of raw RGB frames, these representations reduce sensitivity to background noise, clothing variation, and lighting conditions.

Pose-based retrieval has shown promising results. Parian-Scherb et al. (Parian-Scherb et al., 2024) demonstrated that normalized keypoint trajectories combined with attention mechanisms can effectively model gesture similarity across signers. Their findings highlight the importance of structured motion encoding over purely appearance-based features.

Temporal modeling further enhances representation quality. Ramanathan et al. (Ramanathan et al., 2015) introduced temporal embeddings that preserve motion order and dynamic transitions. Their

work demonstrated that modeling temporal evolution rather than static posture improves discrimination between visually similar actions.

More recent architectures extend temporal modeling through sequence learning structures such as Transformers and graph-based networks. Self-attention mechanisms enable modeling of long-range dependencies across time, allowing systems to capture subtle differences in motion trajectories. Graph-based methods, including Spatial-Temporal Graph Convolutional Networks (ST-GCN), incorporate skeletal structure explicitly by representing joints as nodes and bones as edges. These approaches embed structural priors into motion representation.

The combination of normalized keypoint trajectories and advanced temporal modeling forms a strong foundation for gesture similarity learning.

2.3. Metric Learning for Gesture Retrieval

While representation learning encodes motion patterns, the organization of embeddings in feature space determines retrieval performance. Metric learning provides a principled framework for structuring this embedding space.

Unlike classification objectives that optimize discrete label prediction, metric learning focuses on relative similarity relationships. Ghojogh et al. (Ghojogh et al., 2022) categorize metric learning methods into spectral, probabilistic, and deep approaches, unified by the objective of minimizing intra-class distances and maximizing inter-class distances.

Deep metric learning extends this idea using neural networks trained with specialized loss functions. Contrastive and triplet losses explicitly enforce distance constraints between positive and negative pairs. Cakir et al. (Cakir et al., 2019) further proposed Deep Metric Learning to Rank, directly optimizing ranking-based objectives aligned with metrics such as Recall@K and mAP. Such ranking-aware losses are particularly relevant for retrieval tasks.

Margin-based objectives such as ArcFace introduce angular separation between classes, improving embedding discriminability. Supervised contrastive learning (Khosla et al., 2020) enhances intra-class compactness by leveraging multiple positive samples within each batch. These objectives are especially important in low-shot settings where limited examples per class make generalization challenging (Li et al., 2023).

However, Musgrave et al. (Musgrave et al., 2020) emphasize that improvements in metric learning often depend heavily on fair experimental settings. Their study demonstrates that careful hyperparameter tuning and standardized evaluation are critical for meaningful comparisons. This insight underscores the importance of methodological rigor in gesture retrieval research.

2.4. Robustness and Domain Generalization

Gesture retrieval systems are required to handle large intra-class variability due to differences in signers, speed of performance, camera orientation, and environments. The previous study on adversarial metric learning (Duan et al., 2020) and domain-invariant representations in person re-identification (Zahra et al., 2023) emphasizes the significance of robustness to variations. In the context of gesture retrieval, it is crucial to have domain-invariant representations that are robust to diverse visual conditions, and attention and structured normalization can be used to enhance generalization.

2.5. Research Gaps

Although some progress has been made, there are still some gaps that need to be filled. Video-to-video search purely based on visual information is still a challenge because most current models are based on text or multi-modal alignment. Low-shot learning is also a challenge because the number of samples per gloss is very low. In addition, the evaluation metric in sign language studies is more focused on accuracy than Recall@K and mAP.

2.6. Positioning of the Present Work

This research aims to address the less-explored task of video-to-video similarity learning without textual supervision, specifically in pose-based motion modeling and retrieval-driven metric learning. The proposed system integrates normalized skeletal features, a Transformer, and ST-GCN temporal models, along with ranking-driven metric learning. This research can help develop reverse sign language dictionary systems and provide an understanding of cross-dataset gesture retrieval.

3. Datasets

3.1. Overview

The objective of this work is to develop a video-to-video gesture retrieval system that learns similarity relationships between isolated signs. Unlike classification systems that assign videos to predefined gloss labels, retrieval systems must organize gestures in a continuous embedding space according to visual similarity (Duarte et al., 2022; Cheng et al., 2023). This distinction requires datasets containing multiple instances per class to support metric learning objectives.

Deep metric learning relies on constructing meaningful positive and negative pairs within each class (Ghojogh et al., 2022; Musgrave et al., 2020). Therefore, the dataset must contain more than one sample per gesture. Based on these requirements, the Word-Level American Sign Language (WLASL) dataset was selected as the primary dataset for training and evaluation (Li et al., 2020). In addition, an internal dataset of Russian and Kazakh isolated signs was examined for potential generalization experiments.

3.2. WLASL Dataset

The WLASL dataset consists of 2,000 isolated glosses with an average of three to four video instances per gloss (Li et al., 2020). The dataset contains recordings from multiple signers, introducing variability in motion speed, execution style, and recording conditions. This diversity makes it suitable for evaluating robustness in gesture retrieval.

After preprocessing and removal of unusable samples, the dataset contains 11,980 videos divided into official training, validation, and test splits (8,313 / 2,253 / 1,414). Each video represents a single isolated sign lasting between one and five seconds.

Although the number of samples per class is relatively small, this setting reflects a realistic low-shot learning scenario. Few-shot metric learning literature highlights the difficulty of structuring embedding spaces under limited class repetition (Li et al., 2023). Nonetheless, having at least two samples per gloss satisfies the minimal requirement for supervised contrastive and margin-based losses (Ghojogh et al., 2022; Cakir et al., 2019).

The choice of WLASL is motivated by three factors. First, it provides multiple instances per class, which is essential for metric learning. Second, signer diversity supports learning invariant representations, similar to challenges studied in person re-identification research (Zahra et al., 2023). Third, it contains isolated signs, aligning directly with the objective of reverse dictionary-style retrieval systems (Hassan et al., 2025). While sentence-level datasets exist, they introduce additional segmentation and alignment challenges beyond the scope of isolated gesture retrieval (Duarte et al., 2022; Martins, 2024).

3.3. Preprocessing Consistency

To ensure representation consistency, identical preprocessing steps were applied to all datasets. Each video was converted into pose and hand keypoint sequences, temporally normalized to a fixed frame length, and spatially normalized.

Learning embeddings from normalized temporal sequences aligns with prior work on temporal representation learning (Ramanathan et al., 2015). Keypoint-based modeling also follows established gesture retrieval approaches that rely on structured motion representations (Parian-Scherb et al., 2024). Maintaining a consistent feature space is essential for stable metric learning behavior (Ghojogh et al., 2022).

3.4. Summary

The WLASL dataset was selected as the primary dataset due to its suitability for supervised metric learning and isolated gesture retrieval. Despite its low-shot nature, it provides sufficient structure for embedding-based similarity modeling. The internal Russian and Kazakh dataset cannot support supervised training but offers opportunities for future cross-dataset evaluation. Together, these datasets provide a solid foundation for studying pose-based gesture similarity learning under realistic data constraints (Li et al., 2023). To better analyze the generalization ability across datasets, a subset of the AUTSL dataset with 226 distinct labels was created in a controlled gallery-query split, where four training videos per label and one test video per label were chosen for analysis. This subset was then passed through the same pose pipeline to be compatible with the trained WLASL model.

4. Methods

4.1. Overview of the Proposed Framework

The objective of the research is to discover an embedding space for gestures where similar isolated signs are placed near each other while dissimilar gestures are far apart based on their visual characteristics. In the proposed work, the representation is based on poses with the incorporation of temporal information and metric learning (Ghojogh et al., 2022; Musgrave et al., 2020).

The proposed system has several stages: video input, keypoint extraction, preprocessing and normalization, feature building, training the embedding network, and evaluation with the retrieval task. The key difference between the proposed work and other gesture recognition systems is that the proposed system is based on optimizing the structure of the embedding space to enable the retrieval task based on ranking (Cakir et al., 2019; Duarte et al., 2022).

4.2. Pose and Hand Keypoint Extraction

Instead of working with RGB values, the system utilizes the pose-based representation, which helps eliminate background noise and appearance variations. In the system, 75 keypoints are detected

per frame, with 33 keypoints detected for the upper body and 21 keypoints detected for each hand. Each of these keypoints is represented by three-dimensional coordinates, resulting in 225 features per frame.

This representation only considers movement and pose, without any consideration of texture and color. The structured skeletal representation helps the system cope with changes in illumination, background, and clothing worn by the signer, which is consistent with gesture retrieval approaches that rely on structured motion representations (Parian-Scherb et al., 2024).

4.3. Temporal and Spatial Normalization

For the videos, several preprocessing steps are taken. Missing landmark points are addressed by interpolating the valid frames. Spatial normalization is carried out by normalizing the coordinates with respect to the midpoint of the shoulders and the length of the torso. Temporal normalization of the videos is carried out by fixing the length of all videos at 100 frames.

Since the videos are of different lengths, it is important to have the same length so that the videos can be batch-trained and the embeddings can be extracted. Learning embeddings from temporally aligned sequences follows prior work on temporal embedding learning for video analysis (Ramanathan et al., 2015).

4.4. Feature Representation

Several feature components are built to represent static posture as well as dynamic motion.

The primary feature representation is a set of keypoint coordinates after normalization. To explicitly represent motion, first-order temporal derivatives of the keypoint coordinates, i.e., velocity features, are computed by subtracting consecutive frames. The velocity features help in differentiating gestures with similar static posture but different motion trajectories.

For graph-based experiments, bone vectors are computed by subtracting parent joint coordinates from child joint coordinates based on skeletal structure. Additionally, bone velocities are computed to represent relative joint motion, and this extended feature representation is beneficial for exploiting structural information inherent in the human skeleton using spatial-temporal graph convolutional networks. Such structured modeling aligns with graph-based gesture modeling approaches (Parian-Scherb et al., 2024).

4.5. Embedding Architectures

Two types of architectures are considered, namely, the transformer-based approach to modeling temporality and the Spatial-Temporal Graph Convolutional Networks.

For the former, a higher-dimensional embedding of the input feature space is performed, followed by a positional encoding to preserve temporality. Multiple self-attention layers are employed to learn long-range temporality, enabling the model to learn global motion patterns. Rather than using a mean pooling strategy, attention-based pooling is adopted to compute a weighted aggregation of frame-level representations. This helps to focus on certain segments of temporality. Transformer-based temporal embedding learning builds upon prior work on learning temporal embeddings for video representation (Ramanathan et al., 2015).

For the latter, a skeleton is modeled as a graph, where joints are represented as nodes, and edges are formed by bones. Spatial graph convolutions are employed to capture joint dependencies, and temporally, convolutions are adopted to capture motion evolution. Two-stream variants are also considered to separately process pose and hand information before feature fusion. Graph-based modeling has been widely used for structured motion representation in visual analysis tasks (Zahra et al., 2023).

4.6. Metric Learning Objectives

Several metric learning objectives are considered to organize the embedding space (Ghojogh et al., 2022; Musgrave et al., 2020). In supervised contrastive loss, embeddings of different classes are forced to be far apart, and embeddings of the same class are encouraged to be close (Khosla et al., 2020). ArcFace adds a penalty term for angular margin, which helps to improve separation for classification-based training (Deng et al., 2019). A combination of both objectives is also considered. These objectives directly affect the embedding space, which is essential for retrieval (Cakir et al., 2019; Duan et al., 2020).

4.7. Retrieval Evaluation Setup

After training, all test set samples are mapped to a feature embedding space. The similarity between two gestures is calculated using cosine similarity. For each query, its k-nearest neighbors are retrieved and ranked according to their similarity score.

The performance of the system is measured using various ranking-based metrics, such as Recall@K and mean Average Precision (mAP). These metrics are based on the system's ability to re-

trieve relevant gestures within its top-ranked results, which is in line with the purpose of a reverse gesture dictionary and ranking-based metric learning approaches (Cakir et al., 2019; Duarte et al., 2022).

For the cross-dataset experiment, the same process of retrieval was done on the AUTSL dataset by representing all the gallery and query sequences in the WLASL-trained model, allowing the assessment of how well the embedding space transfers to a completely unseen sign vocabulary.

4.8. Summary

This chapter has provided the complete methodological framework for pose-based gesture similarity learning. The system combines keypoint-based representation, temporal representation using the Transformer and ST-GCN architectures, and metric learning objectives optimized for ranking performance (Ghojogh et al., 2022). The experimental results and the effect of the architecture and loss function choices will be provided in the next chapter.

5. Results

5.1. Experimental Overview

This chapter discusses the experimental evaluation of the proposed gesture retrieval framework. All architectures were trained using the WLASL training set (Li et al., 2020), and the preprocessing technique explained in the Datasets chapter was utilized. It is worth noting that only gesture classes with at least two instances were included in the experiment to facilitate metric learning objectives (Ghojogh et al., 2022; Musgrave et al., 2020).

The experiment was conducted using ranking-based evaluation metrics. For a given query in the validation set, the cosine similarity was calculated between the query embedding and the embeddings in the gallery set. The obtained results were then ranked based on the similarity score. The evaluation metrics employed in the experiment included Recall@1, Recall@5, Recall@10, Recall@50, and mean Average Precision (mAP), which are standard in retrieval-oriented metric learning (Cakir et al., 2019; Duarte et al., 2022). Table 1 presents the performance of all evaluated architectures.

5.2. Quantitative Results

The results demonstrate clear performance differences across architectural and loss configurations.

Table 1: Retrieval Performance Comparison on the WLASL Validation Set. The table reports Recall@K and mean Average Precision (mAP) for different architectures and metric learning configurations. The best results are highlighted in bold.

Model	Architecture	Loss Function	R@1	R@5	R@10	R@50	mAP
M1	Transformer	SupCon	0.178	0.455	0.576	0.771	0.212
M2	Two-Stream ST-GCN	ProxyNCA	0.105	0.311	0.424	0.704	0.098
M3	Two-Stream ST-GCN	ArcFace	0.160	0.396	0.510	0.756	0.160
M4	Two-Stream ST-GCN	ArcFace + SupCon	0.177	0.446	0.559	0.769	0.192
M5	Transformer (Attention Pooling)	SupCon	**0.183**	0.433	0.554	0.732	**0.237**

5.3. Transformer vs ST-GCN Architectures

In all cases, the performance of the Transformer-based models surpassed that of the ST-GCN-based models, as measured by mean Average Precision. The baseline Transformer, even with supervised contrastive loss (Khosla et al., 2020), achieved a value of 0.212, which was already higher than that of ST-GCN with ProxyNCA and ArcFace (Deng et al., 2019).

Despite the explicit modeling of skeletal structure using graph convolutions, ST-GCN failed to achieve a higher retrieval accuracy than the Transformer-based model. This may indicate that global temporal modeling using self-attention is a more powerful approach for modeling fine-grained motion similarity in isolated sign sequences, consistent with temporal embedding learning principles (Ramanathan et al., 2015). The best ST-GCN architecture, ArcFace + SupCon, achieved a value of 0.192, still lower than that of the baseline Transformer.

5.4. Impact of Loss Functions

The impact of the selection of the loss function is substantial. The performance of the ST-GCN model with ProxyNCA was the lowest among all the configurations. Switching the ProxyNCA loss function to the ArcFace loss function (Deng et al., 2019) boosted the performance. This shows the importance of angular margin-based objectives, as they improve the inter-class distance in the embedding space (Ghojogh et al., 2022). Further performance improvements were observed when the supervised contrastive loss function (Khosla et al., 2020) was added to the ArcFace loss function. This shows the importance of metric learning objectives for the task of gesture retrieval, particularly ranking-oriented embedding optimization (Cakir et al., 2019).

5.5. Effect of Attention Pooling

In the last experiment, attention-based temporal pooling was integrated with the Transformer model, along with mild data augmentation and P-K batch sampling (K=3), which is commonly used in metric learning batch construction (Musgrave et al., 2020). This resulted in the best performance in

terms of mean Average Precision (0.237), as well as Recall@1 (0.183), compared to mean pooling. It was found that, compared to mean pooling, using attention-based temporal pooling resulted in a higher weightage being given to informative frames of a gesture sequence. This resulted in a better representation of the gesture, as captured by the learned embeddings, as a whole rather than averaging out across the entire sequence of frames. While there was a minor decline in Recall@5 and Recall@10 compared to the baseline Transformer model, there was a significant improvement in terms of ranking quality, as captured by mAP, which aligns with retrieval-oriented evaluation principles (Cakir et al., 2019).

5.6. Summary of Findings

Based on the experiments, the following conclusions are made:

- Transformer-based temporal modeling achieves higher retrieval performance than graph-based modeling for isolated gesture retrieval.
- Angular margin losses improve embedding separation compared to proxy-based objectives (Deng et al., 2019).
- The combination of ArcFace and supervised contrastive loss enhances retrieval accuracy (Khosla et al., 2020).
- Attention-based temporal pooling yields higher-quality embeddings than mean pooling.

Among all the configurations, the transformer with attention pooling and supervised contrastive loss outperforms the other configurations for retrieval performance. Further analysis of the experiment and its implications, as well as the behavior of the models, are discussed in the next chapter.

5.7. Cross-Dataset AUTSL Evaluation

To test the generalization ability on other datasets, the best model, which was a Transformer with attention pooling, was tested on the AUTSL dataset with 904 gallery and 226 query samples, and it got a Recall@1 of 3.54%, Recall@10 of 29.20%, and mAP of 0.0914. This shows that the model is able to maintain the partial structure in the embedding space despite not being trained on the AUTSL dataset.

6. Discussion

6.1. Overview

This chapter examines the experimental results and discusses their implications for gesture similarity learning. The intent is not to repeat the numerical results, but rather to understand why some architectures and loss functions were better performing and what this means in terms of isolated sign retrieval (Cakir et al., 2019; Duarte et al., 2022). The experiments show that the quality of the embedding is affected by the temporal modeling strategy, loss function, and pooling method (Ghojogh et al., 2022). The Transformer-based architecture with the attention pooling method yielded the best overall retrieval performance, which indicates the strength of global modeling in this problem.

6.2. Transformer vs Graph-Based Modeling

One of the key findings from the current study is that the performance of the Transformer-based model surpassed the performance of the Spatial-Temporal Graph Convolutional Networks (ST-GCN) model in the isolated sign retrieval task. The ST-GCN model incorporates the skeletal structure of the hand through the use of a graph convolutional network. This has been a popular approach in action recognition, where the spatial relationships between the joints are crucial. However, in the current low-shot retrieval setting, the ST-GCN model did not achieve superior performance in comparison to the Transformer-based encoder.

There are a number of reasons why the ST-GCN model did not achieve superior performance in the current study. Firstly, the WLASL dataset contains only three to four instances per class on average (Li et al., 2020). The graph-based model is based on the ability of the model to learn robust spatial-temporal patterns from multiple instances.

Secondly, the Transformers incorporate self-attention mechanisms that capture global temporal dependencies across the whole sequence (Ramanathan et al., 2015). Contrary to the graph convolutional approach that emphasizes local joint relationships, the self-attention mechanisms emphasize all time steps relative to each other. The ability to capture global temporal dependencies could help capture the nuanced differences in motions between visually similar gestures.

Finally, the Transformer architecture is less restricted by the skeletal structure. For the graph models, the structure is heavily influenced by the joint adjacency matrix. In the attention mechanisms, the ability to capture long-range dependencies is flexible and not restricted to the skeletal structure. In a retrieval scenario where the differences matter, the flexibility could offer an advantage.

These results are also in line with the recent trends in sequence modeling where self-attention models have achieved remarkable performance in modeling long-range dependencies across different

domains.

6.3. Influence of Metric Learning Objectives

As mentioned earlier, the experiments also demonstrate the significance of the learning objectives used for the metric learning (Ghojogh et al., 2022). ProxyNCA yielded the poorest performance among all the losses used for evaluation. While proxy-based approaches ease the optimization process by learning the proxy representation for all classes, they may not capture the intra-class clustering adequately, especially under low-shot conditions (Musgrave et al., 2020). Substituting ProxyNCA with the ArcFace loss significantly boosted the performance (Deng et al., 2019). ArcFace incorporates the angular margin, thus improving the inter-class discriminability (Deng et al., 2019). This shows the effectiveness of angular discriminability in the gesture retrieval task.

Further performance gains were observed when the supervised contrastive loss was added with the ArcFace loss (Khosla et al., 2020). While the angular margin improves the inter-class discriminability, the addition of the supervised contrastive loss strengthens the intra-class clustering. This shows the effectiveness of the retrieval-oriented optimization objectives (Cakir et al., 2019).

6.4. Impact of Attention Pooling

The greatest improvement in the overall quality of the retrieval process was achieved by the attention pooling method. The reason for this is the ability of the model to learn which parts of the video contribute the most to the gesture identity, unlike the mean pooling method. Isolated signs have different phases: preparation, execution, and return. Not all frames of the gesture are equally discriminative. The mean pooling method treats all the frames equally, which may reduce the discriminative information of the gesture. The better mAP values of the attention pooling method imply that the model has learned the importance of the frames, which is a crucial component of the gesture similarity modeling process, consistent with attention-based modeling approaches (Parian-Scherb et al., 2024).

6.5. Low-Shot Learning Considerations

WLASL dataset is a low-shot learning scenario, i.e., there are only a few examples per word sense (Li et al., 2020). This also presents a number of challenges for representation learning. Less repetition of classes makes it difficult for the model to learn intra-class variation. The better performance of the Transformer-based model might be due to the robustness of global temporal modeling, as compared to graph-based structural modeling, in low-shot learning scenarios (Li et al., 2023).

Additionally, using P-K sampling with K=3 was beneficial for the stability of contrastive learning, as there were enough positive pairs in each batch (Musgrave et al., 2020). This also emphasizes the role of batch formation in metric learning, especially in low-shot learning.

The cross-dataset experiment on AUTSL further emphasizes that the learned embedding captures transferable motion patterns even when the sign vocabulary and signer distribution are very different from WLASL. While the absolute values of the retrieval scores are not high, the fact that it is able to retrieve the correct matches in a 226-class open-set scenario is an indication that it has learned gesture-level similarities rather than learning dataset-specific information.

6.6. Limitations

Although the results are promising, several aspects are to be noted. Firstly, the study only considers isolated sign retrieval. Continuous sign sequences are not considered. More segmentation mechanisms would be needed to tackle the retrieval of sentences (Martins, 2024).

Secondly, large-scale indexing techniques like FAISS or HNSW were not fully explored (Malkov and Yashunin, 2018; Emanuilov and Dimov, 2024). Although the evaluation of the efficiency of the embeddings with the help of the cosine similarity is sufficient for the analysis of the embeddings' quality, the use of the above-mentioned techniques would be necessary for the effective deployment of the system.

Thirdly, the study is mostly based on the WLASL dataset (Li et al., 2020). The cross-lingual and cross-dataset evaluation of the sign language dataset for the Russian and Kazakh languages is an important task to be validated.

6.7. Implications and Future Directions

The results obtained in the thesis have shown the viability of video-to-video gesture retrieval based on poses, and the effectiveness of attention-based transformer models in low-shot gesture retrieval. The possible ways for the extension of the suggested approach are:

First, a highly efficient method for scaling up the retrieval, such as FAISS or HNSW, might be used (Malkov and Yashunin, 2018; Emanuilov and Dimov, 2024). Second, experiments on gesture retrieval in a different domain, using the internal Russian and Kazakh dataset, might be performed. Third, a new approach to modeling time might enable the extension of the approach to sign language sequences (Ramanathan et al., 2015).

The suggested approach is a contribution to the development of visual gesture similarity systems, and it creates a foundation for the development of a reverse sign language dictionary (Hassan et al., 2025).

7. Acknowledgements

This text was modified with the assistance of ChatGPT.

8. Bibliographical References

Fatih Cakir, Kaiming He, Xunxia Xia, Brian Kulis, and Stan Sclaroff. 2019. Deep metric learning to rank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1861–1870.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. CiCo: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19016–19026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Yueqi Duan, Jiwen Lu, Wenzhao Zheng, and Jie Zhou. 2020. Deep adversarial metric learning. *IEEE Transactions on Image Processing*, 29:2037–2051.

Amanda Duarte, Samuel Albanie, Xavier Giró i Nieto, and Gül Varol. 2022. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14094–14104.

Simeon Emanuilov and Aleksandar Dimov. 2024. Billion-scale similarity search using a hybrid indexing approach with advanced filtering. *Cybernetics and Information Technologies*, 24(4):45–58.

Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. 2022. Spectral, probabilistic, and deep metric learning: Tutorial and survey. *arXiv preprint arXiv:2201.09267*.

Saad Hassan, Matyáš Boháček, Chaelin Kim, and Denise Crochet. 2025. Towards an AI-driven video-based American Sign Language dictionary: Exploring design and usage experience with learners. *arXiv preprint arXiv:2504.05857*.

Zifan Jiang, Gerard Sant Muniesa, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Xiang Li, Xinyu Yang, Zhen Ma, and Ji-Hong Xue. 2023. Deep metric learning for few-shot image classification: A review of recent developments. *Pattern Recognition*, 138:109381.

Yury A. Malkov and Dmitry A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Gonçalo Vinagre Martins. 2024. SLVideo: A sign language video moment retrieval framework. Master's thesis, Universidade NOVA de Lisboa (Portugal).

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Proceedings of the European Conference on Computer Vision*, pages 681–699.

Mahnaz Parian-Scherb, Peter Uhrig, Luca Rossetto, Stéphane Dupont, and Heiko Schuldt. 2024. Gesture retrieval and its application to the study of multimodal communication. *International Journal on Digital Libraries*, 25:585–601.

Vignesh Ramanathan, Kevin Tang, Greg Mori, and Li Fei-Fei. 2015. Learning temporal embeddings for complex video analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4479.

Asmat Zahra, Nazia Perwaiz, Muhammad Shahzad, and Muhammad Moazam Fraz. 2023. Person re-identification: A retrospective on

domain-specific open challenges and future trends. *Pattern Recognition*, 142:109669.

9. Language Resource References

Eleni Efthimiou and Stavroula-Evita Fotinea and Thomas Hanke and John Glauert and Richard Bowden and Annelies Braffort and Christophe Collet and Petros Maragos and François Lefebvre-Albaret. 2012. *Dicta-Sign Wiki*. Dicta-Sign Project. https://doi.org/10.1007/978-3-642-31534-3_32.

Dongxu Li and Cristian Rodriguez Opazo and Xin Yu and Hongdong Li. 2020. *WLASL: Word-Level American Sign Language Dataset*. Australian National University. https://doi.org/10.1109/WACV45572.2020.9093512.

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. *AUTSL: Ankara University Turkish Sign Language Dataset*. Ankara University. https://doi.org/10.1109/ACCESS.2020.3028072.