Preprocessing Mediapipe Keypoints with Keypoint Reconstruction and Anchors for Isolated Sign Language Recognition

Kyunggeun Roh, Huije Lee, Eui Jun Hwang, Sukmin Cho, Jong C. Park
School of Computing
Korea Advanced Institute of Science and Technology
{rohbrian, anqiquer, ehwa20, nelllpic, jongpark}@kaist.ac.kr

Abstract

Isolated Sign Language Recognition (ISLR) aims to classify signs into the corresponding gloss, but it remains challenging due to rapid movements and minute changes of hands. Pose-based approaches, recently gaining attention due to their robustness against the environment, are crucial against such challenging movements and changes due to the difficulty of capturing small joint movements from the noisy keypoints. In this work, we emphasize the importance of preprocessing keypoints to alleviate the risk of such errors. We employ normalization using anchor points to accurately track the relative motion of skeletal joints, focusing on hand movements. Additionally, we implement bilinear interpolation to reconstruct keypoints, particularly to retrieve missing information for hands that were not detected. Preprocessing methods proposed in this work show a 6.05% improvement in accuracy and achieved 83.26% accuracy with data augmentation on the WLASL dataset, which is the highest among pose-based approaches. The proposed methods show strengths in cases with signs having importance in the hand shape, especially when some frames have undetected hands.

Keywords: Sign Language Recognition, Keypoint Preprocessing, Transformer Architecture

1. Introduction

Sign language is the visual means of communication for the deaf, utilizing hand shapes, body movements, and facial expressions to convey messages. Like spoken languages, sign languages have their own diverse vocabulary and grammar. The difficulty of recognizing signs with detailed movements and diverse hand shapes remains as a barrier for hearing individuals to learn sign language. Sign Language Processing is an emerging field of machine learning that makes a bridge between the deaf and hearing individuals, by generating (Saunders et al., 2020), translating (Camgöz et al., 2020), and recognizing (Zhou et al., 2020) sign language expressions.

Isolated Sign Language Recognition (ISLR) focuses on translating sign language videos into the corresponding glosses, which are word-level representations of sign language expressions (Grobel and Assan, 1997; Jiang et al., 2021). ISLR shares similarities with video recognition tasks; however, the limited resources of ISLR datasets have been known as the main limitation, leading models to easily overfit on the dataset (Jang et al., 2022). Pose-based ISLR utilizes pose estimation models for keypoint extraction to overcome the challenges associated with the quantity and quality of datasets (Laines et al., 2023). The extracted keypoints remain independent of backgrounds and subjects, and since the keypoints are relatively lighter than RGB vidoes, they can also be easily augmented to prevent overfitting. Moreover, keypoints can be processed as sequential data with RNN or Transformer-based models or as graph representations with graph neural networks (Ko et al., 2018; de Amorim et al., 2019).

Hand shape is one of the most important components of sign language, containing dense information in a smaller area than the body. Despite the importance of hand shape, pose-based approaches struggle with the challenging task of recognizing hand shapes, which easily differs with that of identifying minute movements of hand keypoints. To address this, the previous methods have been applying normalization on keypoints or have been implementing an additional model seperately trained on the hands (Coster et al., 2020; Hu et al., 2021). The challenge becomes more difficult due to noisy keypoints from the failure of detection on the hands of the pose estimation model. For instance, Mediapipe (Lugaresi et al., 2019), a widely used pose estimation framework in the sign language domain, fails to detect over 50% of the hands appearing in each frame of the word-level American Sign Language dataset, WLASL. The noisy and undetected keypoints hinder hand shapes, leading to wrong predictions (Jiao et al., 2023).

In this work, we introduce a preprocessing framework, focused on hands, developed for pose-based ISLR. Our framework is based on the following strategies: anchor-based normalization, hand keypoint reconstruction, and fixing length. First, anchor-based normalization is applied to normalize the body and hands based on anchor points, which are set to clearly outline the hand shape by considering the relative distance between skeleton joints. Second, we employ keypoint reconstruc-

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 323–334

tion to recover the information of undetected hands by applying bilinear interpolation on surrounding frames. Additionally, the input sign language sequences are padded with frame duplication in a uniform distribution to train the model on stable data with a fixed length.

Finally, for evaluation, we validate our preprocessing framework on two representative ISLR datasets, WLASL-100 (Li et al., 2020a) and AUTSL (Sincan and Keles, 2020). The performance of the methods is assessed using both a Transformer encoder-decoder architecture and an encoder-only architecture to demonstrate the generality of the preprocessing methods. Our proposed methods improve the accuracy of recognizing sign language keypoints by 6.05%, and with basic augmentation, we achieve an accuracy of 83.26% on the WLASL-100 dataset, the highest among pose-based approaches. Further analysis demonstrates the significance of our normalization and reconstruction techniques in ISLR, and case studies show the effectiveness of our methods. We also discuss better input formats for sign language keypoints and handling highly undetected keypoints for future work.

2. Related Work

With the development of machine learning, ISLR research has also been highlighted in recent years. The approaches handling sign language videos are divided into two streams: the RGB-based approach, which directly recognizes features extracted from the RGB video into gloss representations, and the pose-based approach, which extracts skeleton keypoints from the RGB videos and recognizes the keypoints into the corresponding gloss.

2.1. RGB-based Approaches

Early Sign Language Recognition began with applying the Hidden Markov Model (HMM) to ISLR (Grobel and Assan, 1997). These approaches required additional equipment, such as colored gloves. However, with the development of CNN-based models, machine learning models can now segment the hand area without such additional equipment and directly extract feature vectors from the visual representation (Koller et al., 2018; Pigou et al., 2016). With the advancement of language processing models, the sequential feature vectors extracted from the CNN models can be effectively recognized with RNN or LSTM-based models (Koller et al., 2020; Cui et al., 2019). The development of 3D CNN models has demonstrated the strength of a single model capable of extracting both spatial and temporal information from videos without information loss between different models (Tran et al., 2015). Specifically, research using the I3D model has shown that RGB-based methods can achieve reliable results in ISLR (Li et al., 2020a; Joze and Koller, 2019). Still, RGB-based approaches face limitations due to the constrained size of sign language video datasets. This leads models to develop biases towards the environments and appearances of the signers included in the training data. Recently, Jang et al. (2022) proposed a framework designed to augment the sign language video dataset by altering the background of the videos.

2.2. Pose-based Approaches

Pose-based ISLR has a significant advantage in that the pose estimation models are trained on a relatively large dataset compared to sign language datasets, making models more robust against different environments. Since the initial machine learning models with CNN architectures were not specifically designed to handle sequential keypoints, Pham et al. (2019) applied a transformation to the 3D skeleton keypoints to generate an image that contains both the spatial and temporal information of the keypoints, and a ResNet model was employed to recognize the generated image. With the enhancement of sequential models, the keypoints can be directly recognized with RNN or LSTM models (Ko et al., 2018; Liu et al., 2016; Papadimitriou et al., 2023). The skeleton keypoints can also be treated as graphs, and Graph Convolutional Networks (GCNs) have shown the strength of the architecture compared to the previous LSTM and RNN models (Maruyama et al., 2021). Especially, Jiang et al. (2021) have shown that pose-based architectures can outperform 3D CNN-based architectures with GCNs specialized for sign language. With the successful application of Transformer models to keypoints by Hu et al. (2021) and Bohácek and Hrúz (2022), recently, there has been an increasing focus among researchers on exploring the application of the Transformer model.

2.3. Preprocessing Pose Keypoints

One of the advantages of using skeleton keypoints is the lightweight nature compared to RGB videos, making preprocessing much easier. Normalization is a basic preprocessing method, and Transformer-based models have shown that the normalized keypoints can significantly improve the performance (Bohácek and Hrúz, 2022). With data augmentation, keypoint data can be augmented using basic approaches such as rotation and Gaussian noise to prevent the model from overfitting with limited data (Coster et al., 2020). Other approaches have shown that extracting additional features, such as movement of joints or bone information, can help

the recognition (Jiao et al., 2023). As shown in various studies, preprocessing methods enable models to learn effectively and overcome problems related to the limited amount of data.

The primary challenge with pose keypoints is that the pose estimation model can easily fail to detect the correct hand keypoints. To address such errors, researchers have been exploring better frameworks and attempting to combine different modalities (Zuo et al., 2023; Kanakanti et al., 2023). Masking keypoints is another preprocessing method aimed at reducing the risk from error keypoints and making the model more robust on such keypoints (Jiao et al., 2023; Hu et al., 2021). While current approaches focus on optimizing the use of the keypoints, there has not been as much exploration into recovering error keypoints.

In action recognition and other domains, several preprocessing approaches have been developed to improve the quality of noisy keypoints and reconstruct them using autoencoder models (Li et al., 2019; Wu et al., 2020; Zhou et al., 2021). However, these approaches face challenges when applied to sign language keypoints, particularly due to the frequent occurrences of undetected hands that such models cannot easily reconstruct. For instance, the Mediapipe framework, one of the major pose estimation frameworks for sign language, fails to detect almost 50% of the hands on the WLASL dataset. This high rate of undetection necessitates the adoption of alternative preprocessing techniques for hand pose reconstruction.

3. Methodology

The main goal of the proposed work is to concentrate on a more efficient method to normalize and reconstruct the hand keypoints, thereby facilitating the training of the model. In this section, we outline the designed experiments and provide details on how we handled and normalized the data.

3.1. Anchor Based Normalization

Previous keypoint normalization techniques have been focusing on normalizing the keypoints based on the average position of the center of the body and rescaling lengths based on the shoulder length (Yoon et al., 2019). Especially, Coster et al. (2020) and Bohácek and Hrúz (2022) have normalized the keypoints with bounding boxes by aligning the keypoints in a segmented box region. Unlike such approaches, we envision that an anchor point could let the model learn better with a standard point. For this purpose, we normalized the keypoints by shifting them to position the neck (center of the body) fixed on the center $(0, 0)$. To normalize the length information, we divided all values by the length of the neck instead of the shoulders because the neck seemed to be moving less than the shoulders for the ISLR task which has less facial expressions. By centering and scaling, we normalized the skeleton keypoints against the position of the signer so that the model becomes robust no matter how close the signer is to the camera or aligned in some direction. The equation below outlines the normalization process where x_k and y_k are the x and y coordinates, respectively, for each skeleton keypoint. The zeroth keypoint is designated as the neck ($k = 0$), and the first keypoint ($k = 1$) is identified as the center of the head. The normalization formula is given by:

$$(x'_k, y'_k) = \frac{(x_k, y_k) - (x_0, y_0)}{|(x_1, y_1) - (x_0, y_0)|} \quad (1)$$

We also conducted separate normalization for the hands, utilizing anchors positioned on the palm. In sign language recognition, the significance of the hand primarily stems from its shape and position. Since the position of the hand is already incorporated into the body keypoints with the wrist keypoint, our focus for the hands should be on shape information rather than position. To achieve this, we chose to normalize the hands separately from the body, akin to the approach taken by Bohácek and Hrúz (2022), to reduce the weight of positional information and emphasize shape information. However, to capture hand shapes more efficiently, we introduced anchors to the palm and shifted the hands based on these anchors to eliminate positional information. The size of the hands, containing information such as the relative distance from the body, is not separately normalized as length.

3.2. Hand Keypoint Reconstruction

Sign language videos often include rapid hand movements, leading to blurry frames. Extracting keypoints from such blurry frames frequently results in failures in pose estimation. Additionally, signs involve occlusions due to overlapping hands, producing one of the most challenging cases to estimate accurately. To address these challenges, previous research has primarily focused on masking techniques to enhance the model's robustness against noisy keypoints (Hu et al., 2021; Jiao et al., 2023). While these approaches concentrate on making the model robust against noisy keypoints, Laines et al. (2023) have recovered positional information by placing undetected hand keypoints into the position of the palm.

Our approach focuses on recovering the basic information of the hand shape through keypoint reconstruction, as illustrated in Figure 1. We use bilinear interpolation to fill in the empty hand keypoints based on the surrounding skeleton keypoints. To apply bilinear interpolation to frames lacking keypoint data, we require at least one preceding and

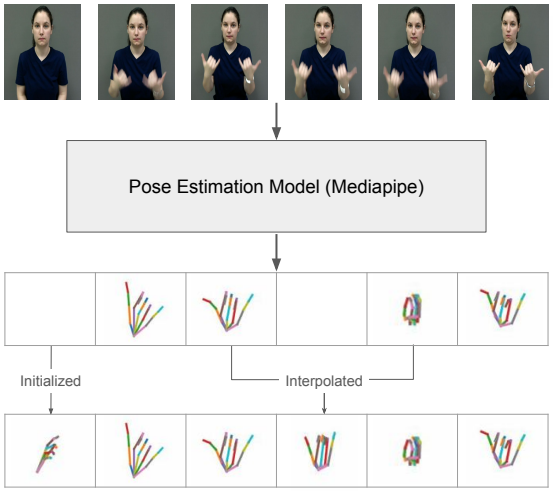

Figure 1: The process of initialization and reconstruction on a single hand. The average shape for the first and last frames is applied for initialization, and bilinear interpolation on other frames is used for reconstruction.

one succeeding frames with identified keypoints to serve as reference points. Therefore, we initiate our process by standardizing the keypoints of the first and last frames based on the average keypoint values, which typically represent the pose when the signer is waiting to start. This initialization step ensures that every empty frame is now sandwiched between frames populated with keypoints. Subsequently, we apply bilinear interpolation to these empty frames to recover the missing information. The provided equation for the normalized hand keypoints f_k from the kth frame incorporates a conditional mechanism to handle both the presence and absence of keypoint data. The equation is structured as follows:

$$f'_k = \begin{cases} \frac{\beta f_{k-\alpha} + \alpha f_{k+\beta}}{\alpha+\beta} & \text{, if } f_k = 0 \\ f_k & \text{, otherwise} \end{cases} \quad (2)$$

where α and β are the minimum numbers that the $k-\alpha$th and $k+\beta$th frames have hand keypoints detected, repectively, which means $f_{k-\alpha} \neq 0$.

3.3. Fixing Length

One of the main motivations of this work is to concentrate on training the model more effectively through data preprocessing. We considered that methods related to the input length could also affect the model's performance. Sign language videos exhibit various lengths, ranging from below 15 frames to over 200 frames for a single gloss. The variability in length is due not only to the difficulty of expressing the sign but also to different signing styles among signers. Typically, padding is applied to short sequences to facilitate training together with long sequences in a single batch (Vázquez-Enríquez et al., 2021). Instead of padding, an alternative approach of interest was extending the length of the input sequence. To do so, frame duplication with a uniform distribution was applied to each instance, fixing the length to 512 frames.

Dataset	# Glosses	# Videos	Detect %
WLASL (2020a)	100	2k	46.56
AUTSL (2020)	226	36k	78.83

Table 1: Statistics related to the two datasets, WLASL and AUTSL. Detect % stands for the detection rate on hands, using the Mediapipe framework.

4. Experiments

Experimental setups are introduced in this section. We provide information about the datasets used, the pose estimation frameworks employed for the experiments, and details regarding the settings.

4.1. Datasets

The datasets chosen to evaluate the proposed approaches are the WLASL and AUTSL datasets. WLASL is a Word-Level American Sign Language dataset that aligns with the task of ISLR (Li et al., 2020a). The dataset is structured with subsets of varying class sizes, 100, 300, 1,000, and 2,000 classes, which are ordered by the number of instances per class. Due to the difficulty of recognizing large subsets, which are unbalanced on the number of instances per class, we decided to use the smallest but richest subset, WLASL-100, for this experiment. The WLASL-100 dataset is composed of 2,038 instances from 97 different signers. With a relatively large number of signers, WLASL exhibits strength in diversity; however, this diversity makes recognition challenging due to the varying signing styles, speeds, and expressions.

The Ankara University Turkish Sign Language Dataset (AUTSL) is a Turkish Sign Language dataset with 226 classes, 36,302 instances, and 43 different signers (Sincan and Keles, 2020). The dataset is relatively balanced regarding the number of instances per class. However, with a smaller number of signers than WLASL, AUTSL exhibits limited diversity concerning the environment. These two datasets were selected for their distinct characteristics so that we can evaluate the efficacy of the proposed methods in diverse settings. As the datasets already include train/dev/test annotations, we apply the annotations for the experiment.

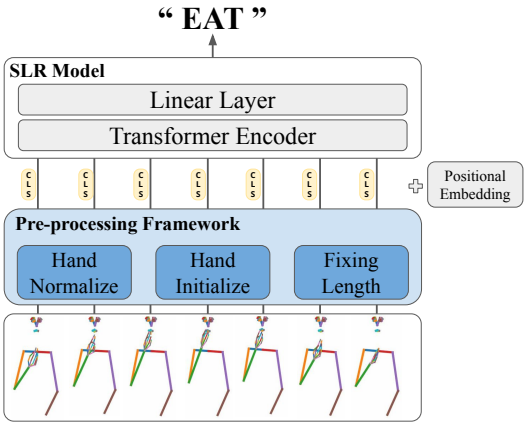

Figure 2: The baseline Transformer encoder architecture framework with preprocessing. Positional Embedding (PE) is added, and the CLS tokens are concatenated to the feature vectors.

4.2. Keypoint Representation

For the datasets mentioned in Section 4.1, we utilized Mediapipe Holistic to extract keypoints from the sign language videos (Lugaresi et al., 2019). Similar to the previous methods, we also decided not to use the z-axis data, as the Mediapipe documentation mentions that it is unreliable. The keypoints we used follow the work of Laines et al. (2023) for fair compatibility. As marked in Table 1, it is quite difficult to detect hands from sign language videos. To retain the positional information of the hands, the palm keypoints were duplicated to be included in the body, resulting in 20 keypoints for the face, 8 keypoints for the body (including the palms) and 21 keypoints for each hand, totaling 70 keypoints. For the baseline settings, the keypoints for undetected hands were set to the position of the palm, and for other settings, we preprocessed the keypoints as mentioned in Section 3.2.

4.3. Model Architecture and Setups

Transformer Encoder. As previous studies have demonstrated the effectiveness of applying Transformer architectures (Vaswani et al., 2017) to ISLR, we have also decided to utilize a Transformer architecture in this work (See Figure 2). However, our approach differs in that we only applied the encoder model, which appeared to be more efficient. The vanilla Transformer encoder model with 4 layers was applied, which shows reliable performance with low complexity that seemed to be more efficient than using more layers. Positional embedding was incorporated with learnable parameters to train the model with the awareness of spatial information, which indicates that each skeleton joint contains distinct information. Similar to the classification based Transformer models by Devlin et al. (2019) and Dosovitskiy et al. (2021), class tokens are concatenated to the features as a parameter. Finally, a linear layer is applied to the output class tokens, and accuracy is measured. We set the Transformer encoder architecture as the baseline and demonstrate the effects of the proposed methods.

To compare the Transformer encoder model with previous researches based on a different Transformer model, we also employ the architecture of SPOTER (Bohácek and Hrúz, 2022). SPOTER is based on a Transformer encoder-decoder architecture with 6 layers and positional embeddings on every feature that contain both spatial and temporal information.

Training Details. The learning rate was fixed at 1e-5, and the models were trained for 200 epochs. Batch size differed between datasets, with WLASL-100 trained on batch size 4, while AUTSL, which has a relatively larger size, was trained with batch size 16. Adam Optimizer was used for optimization. Cross-entropy loss was employed for the training loss, and the top-1 accuracy score was measured for evaluation. All results shared in the results are the average scores from 5 or more attempts with a random seed, as the results may vary depending on the seed number.

4.4. Data Augmentation

Data augmentation is considered as one of the distinct strengths of pose-based ISLR (Alyami et al., 2024; Selvaraj et al., 2022). Previous research has consistently demonstrated that data augmentation significantly enhances performance, especially on limited datasets with unbalanced instances (Zuo et al., 2023). By implementing data augmentation, we show that the proposed preprocessing methods are independent of data augmentation, which means that the methods can be utilized together with different data augmentation techniques from previous and future works.

In this study, we implemented widely adopted augmentation techniques, rotation and Gaussian noise. We adopted the augmentation settings as utilized by Bohácek and Hrúz (2022), applying rotation with angles randomly chosen between -13 and 13 degrees and adding Gaussian noise to each keypoint, following a distribution with a mean of 0 and a standard deviation of 10^{-3}.

5. Results and Analysis

5.1. Main Results

The results of applying the proposed methods appear in Table 2 on the two datasets. With the encoder-only model that we proposed, we can see

Dataset	Model	Method			Acc. (%)
		Hand Normalize	Hand Initialize	Fixing Length	
WLASL	Transformer Encoder-Decoder (SPOTER)	✗	✗	✗	71.63
		✓	✗	✗	79.38
		✓	✓	✗	**80.31**
		✓	✗	✓	78.68
		✓	✓	✓	<u>79.46</u>
	Transformer Encoder-only (Baseline)	✗	✗	✗	76.12
		✓	✗	✗	79.85
		✓	✓	✗	80.62
		✓	✗	✓	<u>81.16</u>
		✓	✓	✓	**82.17**
AUTSL	Transformer Encoder-only (Baseline)	✗	✗	✗	90.40
		✓	✗	✗	90.76
		✓	✓	✗	90.77
		✓	✗	✓	<u>90.95</u>
		✓	✓	✓	**91.15**

Table 2: Comparative results on WLASL and AUTSL between SPOTER and our Transformer encoder ISLR model under three preprocessing settings. Results with the best accuracy score are bold, and the following best results are underlined.

that normalizing hands based on anchors significantly improves accuracy with a 3.73% improvement on the WLASL dataset. Moreover, initializing the keypoints with bilinear interpolation and fixing the input length has also enhanced the performance. Applying all of the methods together, the encoder-only model has shown a 6.05% improvement. The performance change is relatively small in the AUTSL dataset; however, we notice that each method is improving the performance and showing a similar tendency with the results of WLASL.

Results from the Transformer encoder-decoder model show that anchor-based normalization and reconstruction of hands give rise to a significant improvement, which shows the generality of the two methods on a different model architecture. Unlike other methods, fixing the length seemed to be bothering the training process on the encoder-decoder model. The difference of the model based on the SPOTER architecture and the encoder-only model is that the positional embedding of the SPOTER has considered both the spatial and temporal embeddings together, while the baseline model has only been focusing on embedding spatial information. As the length of the input sequences has been extended and fixed by duplication, it seemed that the inconsistent information with the temporal embedding resulted in a lower performance.

5.2. Comparison with Other Methods

WLASL. Results conducted on WLASL are presented in Table 3. Our proposed method outperforms previous pose-based methods. SPOTER† is the result of the SPOTER model trained on Mediapipe keypoints. Our approach outperforms previous RGB-based methods and most of the multimodal methods that use pose and RGB data together. While we still cannot reach the performance of the NLA-SLR model by Zuo et al. (2023), the results highlight the importance of the proposed preprocessing methods.

Method	Modality	Acc. (%)
I3D (2020a)	RGB	65.89
TK-3DConvNet (2020b)		77.55
Full Transformer (2022)		80.72
GCNBERT (2021)	Pose	60.15
SPOTER (2022)		63.18
SPOTER†		71.63
SignBERT (2021)		79.07
SL-TSSI† (2023)		81.47
I3D+ST-GCN (2021)	Multi.	81.38
SignBERT (2021)		82.56
NLA-SLR (2023)		**92.64**
Ours†	Pose	82.17
Ours† w/ Augment		<u>83.26</u>

Table 3: Accuracy comparison on WLASL with previous methods using different modalities. Note that the dagger(†) mark refers to researches based on Mediapipe keypoints and Multi. refers to the multimodal approaches.

AUTSL. In contrast to the results related to WLASL, the results presented in Table 4 indicate

Method	Modality	Acc. (%)
VTN-PF (2021)	RGB	92.92
I3D (2022)		93.53
MViT-SLR (2023)		95.72
SL-TSSI† (2023)	Pose	93.13
MS-G3D (2021)		95.38
SL-GCN (2021)		96.47
SAM-SLR (2021)	Multi.	**98.53**
Ours†	Pose	91.15
Ours† w/ Augment		91.66

Table 4: Accuracy comparison on AUTSL-100 with previous methods with different modalities. Note that the dagger(†) mark refers to researches based on Mediapipe keypoints and Multi. refers to the multimodal approaches.

Method	None	Gauss.	Rotate	Both
Accuracy	82.17	82.24	82.63	**83.26**

Table 5: Accuracy score with different data augmentation methods, Gaussian noise, rotation, and applying both.

that our model underperforms compared to previous methods based on pose and RGB data. The limitation seemed to be due to the smaller number of parameters than the previous methods, as it is highlighted in the earlier work of Laines et al. (2023). SL-TSSI employs 7.2M parameters, and SL-GCN employs around 19.2M parameters, whereas our method works on 5.3M parameters. Moreover, the difference based on the pose estimation frameworks shows that only SL-TSSI has been using the Mediapipe keypoints, which produces a relatively similar result compared to others.

5.3. Analysis

Data Augmentation. We also show that our methods can be enhanced with basic data augmentation skills mentioned in Section 3.4. Table 5 shares the results of applying each augmentation skill. Both augmentation methods are showing improvements, especially when they are applied simultaneously. These results demonstrate that the proposed methods and data augmentation complement each other and show the possibilities with more complicated augmentation methods, such as augmentation based on speed or joint rotation (Bohácek and Hrúz, 2022; Laines et al., 2023).

Normalization Comparison. To show the importance of the anchor-based normalization, we share the results of normalizing our model and the

Method	SPOTER	TF Encoder
Bounding Box	76.59	78.06
Anchor-based	**79.38**	**79.85**

Table 6: Accuracy score of the two models, SPOTER and our Transformer encoder model, with the two different normalization methods of setting bounding boxes and normalizing based on anchors.

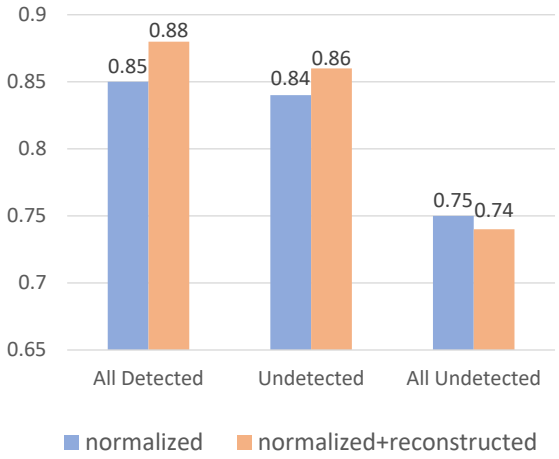

Figure 3: Accuracy scores on all detected, undetected, and all undetected cases. All detected stands for the instances that have all of the hands detected, undetected for those with some hands undetected, and all undetected for those that have at least one hand undetected in all of the frames.

SPOTER model based on our anchor-based normalization and the bounding box-based normalization in Table 6. As we can see, the normalization with anchors on the hands shows better performance on the two different models. The use of anchor keypoints suggests that the model learns more effectively based on the relative distance between skeleton joints.

Reconstruction Effectiveness. The proposed methods have shown improvements in the model performance. To clearly see that the model is recovering the information of keypoints, we divided the WLASL test dataset according to whether the hand detection fails or not. Instances with all hands well detected are checked as “all detected”, some frames having undetected hands are checked as “undetected”, and those with all frames having at least one hand undetected are checked as “all undetected”. For comparison, we analyzed our proposed methods trained with the hands normalized and having the hands reconstructed.

Results are shared in Figure 3, where we observe that the model trained on reconstructed hands exhibits the strength in instances where at least some

hands are detected. The reconstruction not only seemed to be improving the performance based on the recovered information but also seemed to be alleviating the difficulty of training the model with different keypoint representations, some of which have all keypoints detected while others are missing many of the keypoints. However, instances with almost no hands detected seemed to be struggling with reconstructed keypoints that do not possess much information, resulting in a slight decrease in performance. Still, the trade-off is smaller than the improvements noticing that the keypoint reconstruction recovers some information and alleviates the problems coming from undetected hands.

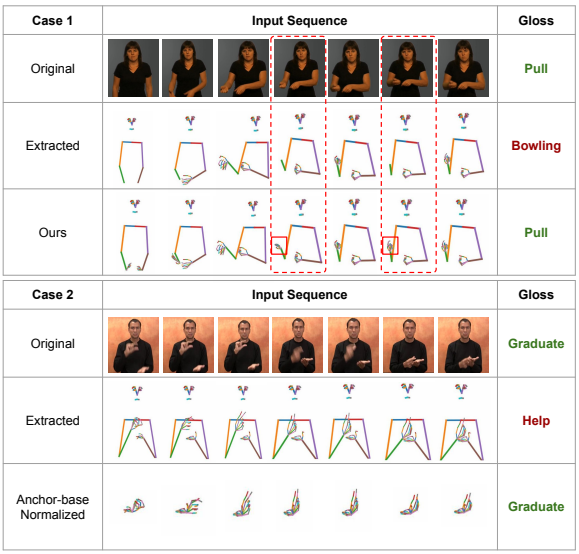

Figure 4: Case studies on the WLASL dataset. Hand keypoints successfully reconstructed are highlighted with red boxes.

5.4. Case Studies

Finally, case studies were conducted to determine if the proposed methods were successfully applied to specific cases. Figure 4 illustrates two cases of when our method has been applied successfully. The first case contains an example where some of the hands are undetected, leading to incorrect predictions. Empty hand keypoints confuse the model, causing it to predict the input sequence into glosses having similar body movements but different hand shapes. Pull and bowling serve as examples of such difficult cases with similar body movements. The loss of keypoints seemed to be leading the model to incorrect predictions. Keypoint reconstruction applied in the proposed research reconstructs the missing hand keypoints and leads the model to correct predictions.

The second case contains an example with a sign that has a particular hand shape containing some important information while the body does not move so much. When the hands are not separately normalized based on anchors, the model struggles to predict similar signs having similar motions even though all hand keypoints are detected. Anchor-based normalization seemed to help the model recognize the shape of hands, leading to correct predictions.

6. Conclusions and Limitations

In this work, we proposed preprocessing methods for Isolated Sign Language Recognition (ISLR). First, we have applied anchor-based normalization, which normalizes the body and hands based on anchor points. Particularly, anchors from the hands remove unnecessary positional information and emphasize the distance between keypoints that effectively retains the shape information. Second, undetected hand keypoints were reconstructed using bilinear interpolation, showing that the reconstructed keypoints recover the shape information of hands. Finally, the length of the sign language sequence was fixed to relieve the difficulty of training a model on data with diverse input lengths. We argue that the methods show the generality across different model architectures and datasets. The application of basic data augmentation methods has improved the performance, demonstrating that the preprocessing methods are independent of data augmentation.

Still, we have several tasks to explore in the future. Fixing the length of the input sequence has been interrupting the training process when we applied the Transformer encoder-decoder model which has both spatial and temporal embeddings. We assume that the temporal embeddings have inconsistent information with the duplicated frames, and leave the question of implementing a better format instead of duplicating the frames for stable training on diverse models for future work. Additionally, the proposed methods still face challenges in cases with highly undetected keypoints, which need to be addressed as well in subsequent work by applying other preprocessing methods or better pose estimation frameworks specialized on hands (Ivashechkin et al., 2023).

7. Acknowledgements

This work was supported by the Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00010, Development of Korean sign language translation service technology for the deaf in medical environment).

8. Bibliographical References

Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. 2024. Isolated Arabic Sign Language recognition using a transformer-based model and landmark keypoints. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(1).

Matyás Bohácek and Marek Hrúz. 2022. Sign pose-based transformer for word-level sign language recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops, Waikoloa, HI, USA, January 4-8, 2022*, pages 182–191. IEEE.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10020–10030. Computer Vision Foundation / IEEE.

Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2020. Sign language recognition with transformer networks. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6018–6024. European Language Resources Association.

Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2021. Isolated sign recognition from RGB video using pose flow and self-attention. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3441–3450. Computer Vision Foundation / IEEE.

Runpeng Cui, Hu Liu, and Changshui Zhang. 2019. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Trans. Multim.*, 21(7):1880–1891.

Konstantinos M. Dafnis, Evgenia Chroni, Carol Neidle, and Dimitri N. Metaxas. 2022. Bidirectional skeleton-based isolated sign recognition using graph convolutional networks. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 7328–7338. European Language Resources Association.

Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. 2019. Spatial-temporal graph convolutional networks for sign language recognition. In *Artificial Neural Networks and Machine Learning - ICANN 2019 - 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17-19, 2019, Proceedings - Workshop and Special Sessions*, volume 11731 of *Lecture Notes in Computer Science*, pages 646–657. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yao Du, Pan Xie, Mingye Wang, Xiaohui Hu, Zheng Zhao, and Jiaqi Liu. 2022. Full transformer network with masking future for word-level sign language recognition. *Neurocomputing*, 500:115–123.

Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1110–1118. IEEE Computer Society.

Kirsti Grobel and Marcell Assan. 1997. Isolated sign language recognition using hidden markov models. In *1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation*, volume 1, pages 162–167. IEEE.

Al Amin Hosain, Panneer Selvam Santhalingam, Parth H. Pathak, Huzefa Rangwala, and Jana Kosecká. 2021. Hand pose guided 3D pooling for word-level sign language recognition. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 3428–3438. IEEE.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. SignBERT: Pre-training of hand-model-aware representation for sign language recognition. In

2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 11067–11076. IEEE.

Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. 2023. Denoising diffusion for 3d hand pose estimation from images. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 3128–3137. IEEE.

Youngjoon Jang, Youngtaek Oh, Jae-Won Cho, Dong-Jin Kim, Joon Son Chung, and In So Kweon. 2022. Signing outside the studio: Benchmarking background robustness for continuous sign language recognition. page 322.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3413–3423. Computer Vision Foundation / IEEE.

Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. 2023. CoSign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20619–20629. IEEE.

Cristina Luna Jiménez, Manuel Gil-Martín, Ricardo Kleinlein, Rubén San Segundo, and Fernando Fernández Martínez. 2023. Interpreting sign language recognition using transformers and Mediapipe landmarks. In *Proceedings of the 25th International Conference on Multimodal Interaction, ICMI 2023, Paris, France, October 9-13, 2023*, pages 373–377. ACM.

Hamid Reza Vaezi Joze and Oscar Koller. 2019. MS-ASL: A large-scale data set and benchmark for understanding American Sign Language. page 100.

Mounika Kanakanti, Shantanu Singh, and Manish Shrivastava. 2023. MultiFacet: A multi-tasking framework for speech-to-sign language generation. In *International Conference on Multimodal Interaction, ICMI 2023, Companion Volume, Paris, France, October 9-13, 2023*, pages 205–213. ACM.

Sang-Ki Ko, Jae Gi Son, and Hye Dong Jung. 2018. Sign language recognition with recurrent neural network using human keypoint detection. In *Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems, RACS 2018, Honolulu, HI, USA, October 09-12, 2018*, pages 326–328. ACM.

Oscar Koller, Necati Cihan Camgöz, Hermann Ney, and Richard Bowden. 2020. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(9):2306–2320.

Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2018. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms. *Int. J. Comput. Vis.*, 126(12):1311–1325.

David Laines, Miguel González-Mendoza, Gilberto Ochoa-Ruiz, and Gissella Bejarano. 2023. Isolated sign language recognition based on tree structure skeleton images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 276–284. IEEE.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020a. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 1448–1458. IEEE.

Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. 2020b. Transferring cross-domain knowledge for video sign language recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6204–6213. Computer Vision Foundation / IEEE.

Shujie Li, Yang Zhou, Haisheng Zhu, Wenjun Xie, Yang Zhao, and Xiaoping Liu. 2019. Bidirectional recurrent autoencoder for 3d skeleton motion data refinement. *Comput. Graph.*, 81:92–103.

Tao Liu, Wengang Zhou, and Houqiang Li. 2016. Sign language recognition with long short-term memory. In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, pages 2871–2875. IEEE.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172.

Mizuki Maruyama, Shuvozit Ghose, Katsufumi Inoue, Partha Pratim Roy, Masakazu Iwamura, and Michifumi Yoshioka. 2021. Word-level sign language recognition with multi-stream neural networks focusing on local regions. *CoRR*, abs/2106.15989.

Gokul NC, Manideep Ladi, Sumit Negi, Prem Selvaraj, Pratyush Kumar, and Mitesh M. Khapra. 2022. Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Maxim Novopoltsev, Leonid Verkhovtsev, Ruslan Murtazin, Dmitriy Milevich, and Iuliia Zemtsova. 2023. Fine-tuning of sign language recognition models: a technical report. *CoRR*, abs/2302.07693.

Katerina Papadimitriou, Gerasimos Potamianos, Galini Sapountzaki, Theodoros Goulas, Eleni Efthimiou, Stavroula-Evita Fotinea, and Petros Maragos. 2023. Greek Sign Language recognition for an education platform. *Universal Access in the Information Society*, pages 1–18.

Huy-Hieu Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio A. Velastin. 2019. Learning to recognise 3D human action from a new skeleton-based representation using deep convolutional neural networks. *IET Comput. Vis.*, 13(3):319–328.

Lionel Pigou, Mieke Van Herreweghe, and Joni Dambre. 2016. Sign classification in sign language corpora with deep neural networks. In *sign-lang@LREC 2016*, pages 175–178. European Language Resources Association (ELRA).

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 687–705. Springer.

Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133.

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355.

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2022. Using motion history images with 3D convolutional networks in isolated sign language recognition. *IEEE Access*, 10:18608–18618.

Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3D convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4489–4497. IEEE Computer Society.

Anirudh Tunga, Sai Vidyaranya Nuthalapati, and Juan P. Wachs. 2021. Pose-based sign language recognition using GCN and BERT. In *IEEE Winter Conference on Applications of Computer Vision Workshops, WACV Workshops 2021, Waikoloa, HI, USA, January 5-9, 2021*, pages 31–40. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Manuel Vázquez-Enríquez, José Luis Alba-Castro, Laura Docío Fernández, and Eduardo Rodríguez Banga. 2021. Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3462–3471. Computer Vision Foundation / IEEE.

Zhize Wu, Thomas Weise, Le Zou, Fei Sun, and Ming Tan. 2020. Skeleton based action recognition using a stacked denoising autoencoder with constraints of privileged information. *CoRR*, abs/2003.05684.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, pages 4303–4309. IEEE.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2020. Spatial-temporal multi-cue network for continuous sign language recognition.

In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13009–13016. AAAI Press.

Kanglei Zhou, Zhiyuan Cheng, Hubert P. H. Shum, Frederick W. B. Li, and Xiaohui Liang. 2021. STGAE: spatial-temporal graph auto-encoder for hand motion denoising. In *IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2021, Bari, Italy, October 4-8, 2021*, pages 41–49. IEEE.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural language-assisted sign language recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14890–14900. IEEE.