electronics

MDPI

Article

Multi-Stream General and Graph-Based Deep Neural Networks for Skeleton-Based Sign Language Recognition

Abu Saleh Musa Miah [1], Md. Al Mehedi Hasan [2], Si-Woong Jang [3], Hyoun-Sup Lee [4,*] and Jungpil Shin [1,*]

[1] School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Japan
[2] Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology (RUET), Rajshahi 6204, Bangladesh; mehedi_ru@yahoo.com
[3] Department of Computer Engineering, Dongeui University, Busanjin-Gu, Busan 47340, Republic of Korea
[4] Department of Applied Software Engineering, Dongeui University, Busanjin-Gu, Busan 47340, Republic of Korea
* Correspondence: lhskmj@deu.ac.kr (H.-S.L.); jpshin@u-aizu.ac.jp (J.S.)

Abstract: Sign language recognition (SLR) aims to bridge speech-impaired and general communities by recognizing signs from given videos. However, due to the complex background, light illumination, and subject structures in videos, researchers still face challenges in developing effective SLR systems. Many researchers have recently sought to develop skeleton-based sign language recognition systems to overcome the subject and background variation in hand gesture sign videos. However, skeleton-based SLR is still under exploration, mainly due to a lack of information and hand key point annotations. More recently, researchers have included body and face information along with hand gesture information for SLR; however, the obtained performance accuracy and generalizability properties remain unsatisfactory. In this paper, we propose a multi-stream graph-based deep neural network (SL-GDN) for a skeleton-based SLR system in order to overcome the above-mentioned problems. The main purpose of the proposed SL-GDN approach is to improve the generalizability and performance accuracy of the SLR system while maintaining a low computational cost based on the human body pose in the form of 2D landmark locations. We first construct a skeleton graph based on 27 whole-body key points selected among 67 key points to address the high computational cost problem. Then, we utilize the multi-stream SL-GDN to extract features from the whole-body skeleton graph considering four streams. Finally, we concatenate the four different features and apply a classification module to refine the features and recognize corresponding sign classes. Our data-driven graph construction method increases the system's flexibility and brings high generalizability, allowing it to adapt to varied data. We use two large-scale benchmark SLR data sets to evaluate the proposed model: The Turkish Sign Language data set (AUTSL) and Chinese Sign Language (CSL). The reported performance accuracy results demonstrate the outstanding ability of the proposed model, and we believe that it will be considered a great innovation in the SLR domain.

Keywords: sign language recognition (SLR); large scale dataset; American Sign Language; Turkish Sign Language; Chinese Sign Language; AUTSL; CSL

Citation: Miah, A.S.M.; Hasan, M.A.M.; Jang, S.-W.; Lee, H.-S.; Shin, J. Multi-Stream General and Graph-Based Deep Neural Networks for Skeleton-Based Sign Language Recognition. *Electronics* **2023**, *12*, 2841. https://doi.org/10.3390/electronics12132841

Academic Editor: Chunjie Zhang

Received: 5 May 2023
Revised: 24 June 2023
Accepted: 25 June 2023
Published: 27 June 2023

1. Introduction

Sign language is a spatial type of visual language based on dynamic gesture movement, including hand, body, and facial gestures [1–7]. It is the language used by those communities who do not speak or hear anything spatially, including the deaf and speech-impaired people. Due to the difficulties and complexity of sign language, such as the considerable time required to understand and utilize it, the non-deaf community is generally not eager to learn this language to establish communication with those specialized disabled people. In addition, teaching this language to the non-deaf community to communicate with the minor community is not practical or feasible. Moreover, there are no common international versions of sign language, and it differs with respect to several languages, such as

Bangla [3], Turkish [8], Chinese [9], and English [10], as well as culture [1,4,11]. To establish effective communication between the non-deaf community and the deaf community, a sign language translator is needed; however, it is rare to find expert sign language interpreters. In this context, researchers believe that automatic sign language recognition (SLR) can effectively address these problems [3–6].

Researchers have worked to develop SLR systems with the help of computer vision [3,4,12], sensor-based methods [5,6,12–19], and artificial intelligence, in order to facilitate communication for deaf and hearing impaired people. Many researchers have recently proposed skeleton-based SLR systems that mainly use specific skeleton points instead of the pixels of images and/or sensors [20–24]. The main advantage of skeleton-based SLR systems is that they can increase the attention paid to signs and have strong adaptability to complicated backgrounds and dynamic circumstances. However, there are still some deficiencies in extracting the skeleton points for SLR, for example, due to the high computational complexity of ground truth skeleton annotation. Many motion capture systems (e.g., Microsoft Kinetic, Microsoft Oak-D, and Intel RealSense, among other systems) provide the main body coordinates and their skeleton annotations; however, it is difficult to obtain skeleton annotations for various gestures [25]. Shin et al. extracted 21 hand key points from an American sign language data set using the MediaPipe system. After extracting the distance and angular features, they applied an SVM for recognition [26]. The use of hand skeleton information alone is sometimes insufficient to correctly represent the exact meaning of a sign due to a lack of emotion and bodily expression. As such, researchers have recently considered that the use of a full-body skeleton may be more effective for SLR systems [27].

Xia et al. extracted the hand skeleton and body skeleton with different approaches and achieved good performance using an RNN-based model [28]. Their main problems were the unreliability of hand key points and that the RNN did not perform well with respect to the dynamics of the skeleton. Perez et al. extracted 67 key points, including those related to the face, body, and hand gestures, using an OpenCV AI Kit with Depth (OAK-D) camera. They recorded 3000 skeleton samples for Mexican sign language (MSL) by considering 30 different signs, where each sample was constructed using the spatial coordinates of the body, face, and hands in 3D. They mainly calculated the motion from the extracted skeleton key points, and, finally, they reported the performance accuracy obtained by an LSTM with gated recurrent units (GRUs) [29]. Jiang et al. applied a different approach to a multimodal data set including full-body skeleton points, and achieved good performance accuracy [8,30]. They also considered reducing the number of skeleton points to decrease the computational complexity of the model. The main problem is that this method did not seem capable of achieving good performance and generalization for the SLR task when compared to existing systems.

In addition, researchers have focused on skeleton-based SLR due to the high complexity of pixel-based systems. For example, 67 landmark points have been considered, including those related to body, face, and hand gestures, which led to computational complexity problems when forming a graph. Furthermore, although the use of a full-body skeleton including 133 points may decrease the performance accuracy compared to when using a pixel-based approach, the computational complexity should be further lowered by selecting effective skeleton points from these 133 key points. To overcome the above-mentioned challenges, in this paper, we propose the multi-stream graph-based deep neural networks (SL-GDN) approach to recognize sign language using fewer skeleton points selected to represent whole-body information more effectively.

For this study, we designed a new skeleton graph for SLR, including spatial and temporal features, using a graph and a neural network to model the embedded dynamics.

The major contributions of the work are detailed below:

- We construct a skeleton graph for large-scale SLR using 27 key points selected among the whole-body key points. The main purpose of this graph is to construct a unified graph to dynamically optimize the nodes and edges based on different actions due

to the minimum number of the skeleton key points being selected among the whole-body points, and extract features from four streams that can be solved to increase the model's performance accuracy and generalizability.
- We extract hybrid features from the multiple streams, including joints, joint motion, bones, and bone motion of the skeleton by combining the graph-based SL-GDN and general neural network features. After concatenating the features, we use a classification module to refine the concatenated features for prediction.
- We use two large-scale data sets with four modalities (i.e., joint, joint motion, bone, and bone motion) to evaluate the model, and our model presents superior performance when compared to an existing system.

The remainder of this paper is organized as follows: Section 2 summarizes the existing research work and problems related to the presented work. Section 3 describes the benchmark and proposed Korean sign language data sets, and Section 4 describes the architecture of the proposed system. Section 5 details the evaluation performed, including a comparison with a state-of-the-art approach. In Section 6, our conclusions and directions for future work are discussed.

2. Related Work

Many researchers have worked to develop automatic sign language recognition systems using various approaches, including segmentation, semantic detection, feature extraction, and classification [9,31–37]. Some studies have considered the use of scale-invariant feature transform (SIFT) [34] or histogram of oriented gradients (HOG), ref. [35] for handcrafted feature extraction, followed by machine learning approaches such as support vector machine (SVM) or *k*-nearest neighbors (*k*NN) [9,36,37] for classification. The main drawback of segmentation–semantic detection methods is that they may face difficulties in producing a good performance for video or large-scale data sets. To overcome the challenges, researchers have recently focused on the various deep-learning-based approaches to improve the potential features and SLR classification accuracy from video and large-scale data sets [1–4,31,32,38–43]. Existing SLR systems still face many difficulties in achieving good performance, due to the high computational cost of the potential information, considerable gestures for SLR, and potential features. One of the most common challenges is capturing the global body motion skeleton at the same time as local arm, hand, and facial expressions. Neverova et al. employed a ModDrop framework to initialize individual and gradual fusion modalities for capturing spatial information [38]. They achieved good performance in terms of spatial and temporal information for multiple modalities. However, one of the drawbacks of their approach is that they applied data augmented with audio, which is not effective at all times.

Pu et al. employed connectionist temporal classification (CTC) for sequence modeling and a 3D convolutional residual network (3D-ResNet) for feature learning [39]. The employed LSTM and CTC decoder were jointly trained with a soft dynamic time warping (soft-DTW) alignment constraint. Finally, they employed 3D-ResNet for training labels with loss and validated the developed model on the RWTHPHOENIX-Weather and CSL data sets, obtaining a word error rate (WER) of 36.7% and 32.7%, respectively. Koller et al. employed a hybrid CNN-HMM model to combine the two kinds of features; namely, the discriminative features of the CNN with the sequence features of the hidden Markov model (HMM) [31]. They claimed that they achieved good recognition accuracy for three benchmark sign language data sets, reducing the WER by 20%. Huang et al. proposed an attention-based 3D-convolutional neural network (3D-CNN) for SLR, in order to extract the spatio-temporal features, and selected highlighted information using an attention mechanism [44]. Finally, they evaluated their model on the CSL and ChaLearn 14 benchmark data sets, and achieved 95.30% accuracy on the ChaLearn data set.

Pigou et al. proposed a simple temporal feature pooling-based method and showed that temporal information is more important for deriving discriminative features for video classification-related research [45]. They also focused on the recurrence information with

temporal convolution, which can improve the significance of the video classification task. SINCAN et al. proposed a hybrid method combining an LSTM, feature pooling, and a CNN method to recognize isolated sign language [46]. They included the VGG-16 pre-trained model with the CNN part and two parallel architectures for learning RGB and depth information. Finally, they achieved 93.15% accuracy on the Montalbano Italian sign language data set. Huang et al. applied a continuous sign language recognition approach to eliminate temporal segmentation in pre-processing, which they called hierarchical attention network with latent space (LS-HAN) [47]. They mainly included a two-stream CNN, LS, and a HAN for video feature extraction, semantic gap bridging, and latent space-based recognition, respectively. The main drawback of their work is that they mainly extracted pure visual features, which are not effective for capturing hand gestures and body movements. Zhou et al. proposed a holistic visual appearance-based approach and a 2D human pose-based method to improve the performance of large-scale sign language recognition [48]. They also applied a pose-based temporal graph convolution network (Pose-TGCN) to extract the temporal dependencies of pose trajectories and achieved 66% accuracy on 2000-word glosses. Liu et al. applied a feature extraction approach based on a deep CNN with stack temporal fusion layers with a sequence learning model (i.e., Bidirectional RNN) [49].

Guo et al. employed a hierarchical LSTM approach with word embedding, including visual content for SLR [50]. First, spatio-temporal information is extracted using a 3D CNN, which is then compacted into visemes with the help of an online key based on the adaptive variable length. However, their approach is may not good for capturing motion information. The main drawback of image and video pixel-based work is the high computational complexity. To overcome these drawbacks, researchers have considered the joint points, instead of pixels of full images, for hand gesture and action recognition [51–53]. Various models have been proposed for skeleton-based gesture recognition, including LSTMs [24] and RNNs [54]. Yan et al. applied a graph-based method—namely, ST-GCN—to construct a dynamics pattern for skeleton-based action recognition using a graph convolutional network (GCN) [24]. Following the previous task, many researchers have employed modified versions of the ST-GCN to improve the performance accuracy for hand gestures and human activity recognition. Li et al. employed an encoder and decoder for extracting action-specific latent information [53]. They included two links for this purpose, and finally employed a GCN-based approach (action-structured GCN) to learn temporal and spatial information. Shi et al. have employed a two-stream-based GCN for action recognition [55] and a multi-stream GCN for action recognition [21]. In the multi-stream GCN, they integrated the GCN with a spatio-temporal network to extract the more important joints and features from all of the features. Zhang et al. proposed a decoupling GCN for skeleton-based action recognition [20].

Song et al. proposed ResGCN integrated with part-wise attention (PartAtt) to improve the performance and computational cost of skeleton-based action recognition [22]. However, the main drawback was that their performance was not significantly higher than that of the existing ResNet. Amorin et al. proposed a human skeleton movement-based sign language recognition using ST-GCN, where they proposed to select potential key points from the whole-body key points. Finally, they achieved 85.0% accuracy on their data set (named ASLLVD) [56]. The disadvantage of this work was that they considered only one hand with the body key points. Perez et al. extracted 67 key points, including face, body, and hand gestures, using a special camera. Finally, they achieved good performance with an LSTM [29]. In the same way, many researchers have considered 133 points from the whole body to recognize sign language [8]. Jiang et al. applied a different approach with a multimodal data set including full-body skeleton points and achieved good performance accuracy [8]. To improve upon the performance and decrease the computational cost of these models, we propose the multi-stream graph-based deep neural network (SL-GDN) to recognize sign language using potential skeleton points from whole-body information.

3. Data Sets

We use two large-scale data sets in this study, which are detailed in Table 1. In Section 3.1, the AUTSL data set is described, whereas Section 3.2 describes the CSL data set.

Table 1. Data sets utilized in this study.

Data Set	Language	Year	Signs	Subjects	Total Sample	Each Sign Word
AUTSL [57]	Turkish	2020	226	43	38,336	169 (on average)
CSL [48,58]	Chinese	2019	500	50	125,000	250

3.1. AUTSL Data Set

One of the sign language data sets considered in this paper is the Turkish Sign Language data set (AUTSL), which was collected from diverse, challenging backgrounds including real-life scenarios. To record the data set, a Microsoft Kinect V2 was used, including RGB, depth, and skeleton modalities [57]. This data set was collected from 43 people considering 226 signs. They recorded 38,336 video clips in total for the 226 signs at 30 frames per second, collected with 20 different challenging backgrounds. In the background, they considered the camera's field-of-view, increasing or decreasing the appearance by adding a new object or removing an object in the background. In addition, moving trees, various lighting conditions, sunlight, artificial light, people passing behind the signer, and some bright, dark, or shadowed areas. For the selected signs, they considered the most-used words in the Turkish language, including push, wait, shoe, face, wait, help, danger, doctor, hospitals, building, and signs. Figure 1 shows sample signs from the AUTSL data set.

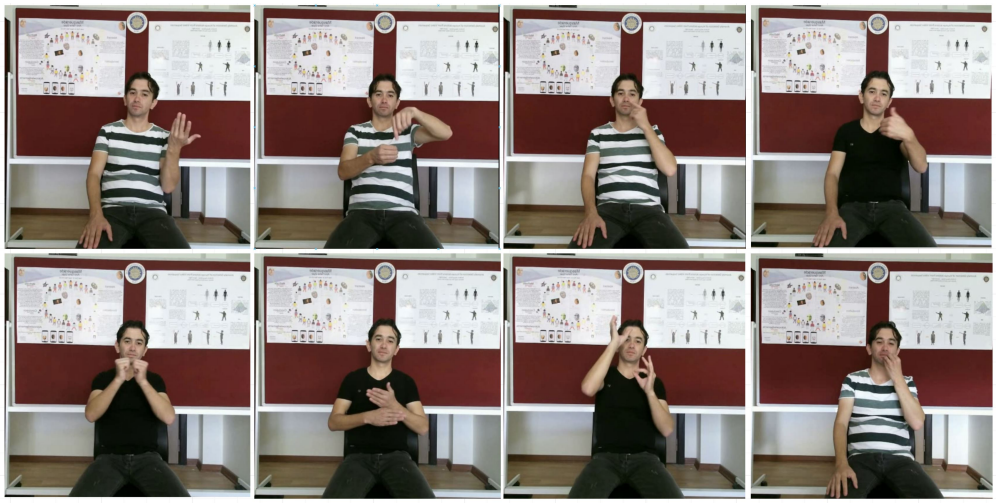

Figure 1. Samples from AUTSL data set.

3.2. CSL Data Set

CSL, in the context of a large-scale data set, refers to a collection of videos depicting Chinese Sign Language, along with their corresponding transcriptions and annotations. There are 50 subjects, with the data having Depth, RGB, and Skeleton modalities. The videos were recorded at 30 FPS with 1280 $\times$ 720 RGB resolution, having a duration of 2–4 s [48,58]. They selected 500 different words for the labels and recorded 2–4 videos for each word. In total, 125,000 videos were recorded from 50 people. These data sets are often used for training and evaluating machine learning models for sign language recognition, translation, and other tasks, and is an important resource for research and development in sign language processing, including sign language recognition, translation, and other tasks. The large-scale CSL data set typically includes many sign language samples recorded from a diverse group of signers, covering a wide range of signs and variations in signing styles. This diversity ensures that machine learning models trained on the data set can generalize to

real-world signing scenarios. The annotations in the CSL data set include information such as the signs being performed, each sign's start and end times, and the signing posture and movement patterns. This information can be used to train and evaluate machine learning models that aim to recognize and transcribe sign language. Such large-scale data sets are crucial in advancing sign language processing technology and making it more accessible to the deaf and hard-of-hearing communities. These data sets provide a foundation for building more accurate and effective sign language recognition and translation systems, which can help to bridge the communication gap between the hearing and non-hearing worlds. Figure 2 shows a sample sign from the CSL data set.

Figure 2. Sample from CSL data set.

4. Proposed Methodology

For this study, we developed the multi-stream graph-based deep neural network (SL-GDN) approach to recognize sign language using potential skeleton points from whole-body information. This idea was inspired by the concept of Jiang [8], and we extracted joint motion, bone, and bone motion information from the joint skeleton data. The key idea is applying a neural network (NN) with a fully connected layer to construct a fully connected graph from the selected whole-body key points. The objective is to dynamically learn edge and node features through the use of a sequential graph a and general convolutional network, which is performed using both spatial and temporal information. The graph is mainly constructed as a spatiotemporal model for recognizing hand gestures based on dynamic human body skeleton information. We also adopted a multi-stream approach considering various information, in order to further improve the performance of the model. Although researchers have developed an SLR system with 21 key points extracted using a MediaPipe-based approach [26]. It was found that the use of only hand information cannot fully express a sign's exact meaning and emotion. Therefore, researchers have come to believe that the use of full-body skeleton data is more impactful for SLR systems [27]. For this purpose, some researchers have considered both the body and the hand, with the aim of localizing the key points or joints in the human body from a single image or video. Besides traditional approaches—such as pictorial structures [59] and probabilistic models [29]—for estimating single-person poses, many researchers have developed systems using the ground truth skeleton derived from motion capture devices such as the Kinect version 2 [60].

At present, many deep-learning-based techniques can extract the key points for the whole body. Although these deep-learning-based pose estimation approaches generate key body points, they may be insufficient due to their spatial dependencies on the extracted key points. One researcher [8] extracted 133 key points for the whole body, including for the body and face, whereas others have followed a different approach [12], taking 42 key

points from the left and right hands and the rest of the key points from the upper body [8,9]. Among the 133 key points for the whole body, we selected the 27 most effective key points using the graph reduction approach, considering 20 key points from the left and right hands and 7 from the upper body. Figure 3 shows the detailed workflow for the architecture of the proposed model. We first took the 27 whole-body joint key points, then extracted joint motion, bone, and bone motion key points from them based on a previously described formula [8].

In each of the skeleton streams, we applied an NN with a fully connected layer, in order to form a fully connected graph in which the node and edge features learn through graph convolution in the deep neural network. We then extracted spatial features using the graph convolutional network and fed them to the convolutional, batch normalization, ReLU, and dropout layers of the neural network to produce a feature vector. In the same way, we extracted features from the four streams and concatenated them to produce the final feature vector. We fed the final feature vector into the classification module to refine the final features and, after converting the feature matrix into a vector, we used a classification layer. Figure 3 shows the workflow of the proposed study, where JM denotes joint motion and BM denotes bone motion. Figure 4 depicts the NN, SL-GDN, and classification modules separately. Figure 4a includes a fully connected (FC) layer, a ReLU activation layer, a normalization layer, and a dropout layer. Figure 4b includes a convolutional graph layer, a batch normalization (BN) layer, two ReLU activation layers, a convolutional layer, and a dropout layer. Figure 4c includes a ReLU activation layer, a neural network (NN) module, an averaging layer, and a fully connected (FC) layer. Figure 4d shows the channel attention layer in detail.

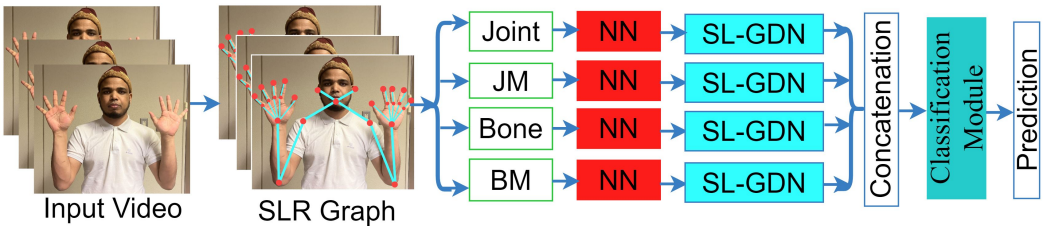

Figure 3. Workflow of the proposed architecture.

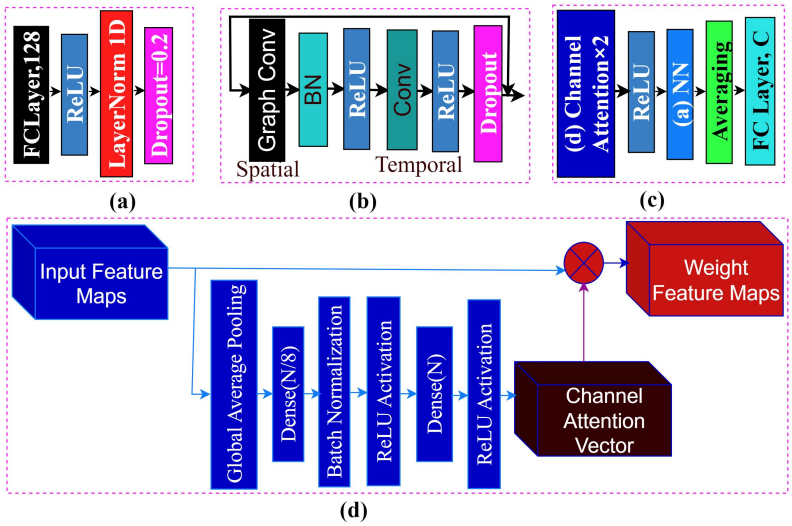

Figure 4. Details of: (**a**) neural network (NN); (**b**) SL-GDN; (**c**) classification module; and (**d**) channel attention module.

4.1. Key Point Selection and Graph Construction

A sequence of vectors is derived from each frame, which is considered a raw skeleton with the individual vectors representing the 2D coordinates of the individual human joints.

Moreover, a full hand gesture sign consists of a certain amount of frames, based on the number of frames and samples. We constructed a spatio-temporal graph by considering the natural connections among the adjacent skeleton points. First, we constructed a node set $V = v(i,t) \mid i = 1,\ldots,N, t = 1,\ldots,T$, consisting of body, face, and hand pose skeleton points. To construct the adjacency matrix for the graph, we used the following formula (Equation (1)):

$$f(x) = \begin{cases} 1 & \text{if the nodes are adjacent} \\ 0 & \text{if they are not adjacent.} \end{cases} \quad (1)$$

Here, the adjacency condition is defined in terms of the minimum distance or shortest path calculated between two nodes. As mentioned, there were 133 key points in our pose, including body, face, and hand points. Due to the large number of nodes and edges, unnecessary noise may be introduced. In addition, in any case, if two nodes are far from each other, then it is not easy to extract the relation between them. Due to this complexity, and as all key points produce noise, it is difficult to improve the performance accuracy [8,61]. Therefore, based on the visualization of the spatiotemporal graph, we selected 27 nodes based on graph reduction algorithms. These 27 nodes included ten nodes for each hand and seven key points from the upper body (shown in Figure 1) as an SLR graph. With this reduction, the performance is increased with the low computational cost.

4.2. Neural Network (NN)

We constructed a graph from the whole-body skeleton and then extracted features from the skeleton-based graph using graph convolution and a general neural network. To increase the ability to modify the unified graph dynamically, based on the different actions, we employed an NN for the skeleton. The main purpose of this NN is to achieve generalizability in constructing skeleton graphs, in a manner not dependent on the number of skeleton points. We employed the NN to produce the initial feature from the skeleton points, in which we first employed a fully connected layer along with the ReLU function, followed by normalizing with a normalization layer and a dropout layer to reduce overfitting, finally producing the initial feature F1 [1].

4.3. Graph Convolution

We consider the spatiotemporal graph based on the strategy of spatially partitioning the dynamic skeleton model in order to extract the potential patterns embedded in the whole-body skeleton graph [8,24]. To construct the spatial graph for the whole-body points, we use Equation (2) as follows:

$$G_{out} = D^{-(1/2)}(A + I)D^{-(1/2)} \times W, \quad (2)$$

where A, I, D, and W denote the intra-body connection, self-connection (or identity matrix), the diagonal matrix is $(A + I)$,

and trainable weight matrix for convolution, respectively. For the implementation of graph convolution, we performed 2D convolution and multiplied it with $DD^{-(1/2)}(A + I)D^{-(1/2)}$, which can be considered as a spatial graph convolution. We also conducted a 2D convolution with a kernel of size $k_t \times 1$ to implement the temporal graph convolution. We adopted a neural network (NN) architecture consisting of a fully connected network to boost the network's capacity. The fully connected output of the NN is fed to the SL-GDN network to produce the final features.

4.4. SL-GDN Architecture Block

The proposed SL-GDN takes the output of the NN as an input, then performs a $k_t \times 1$ convolution on the $N \times T \times C$ initial feature map, where N, T, and C denote the number of vertices, temporal length, and the number of channels, respectively. The SL-GDN architecture is mainly constructed with a graph convolution layer, and the batch normalization (BN) layer, and relies on a convolutional layer, ReLU layer, and dropout

layer. After that, this temporal feature is concatenated with the initial feature to produce the final feature.

Figure 5 depicts the backbone [53] of the proposed SL-GDN method. Notably, we stacked 13 SL-GDN modules as a backbone network, providing a powerful backbone.

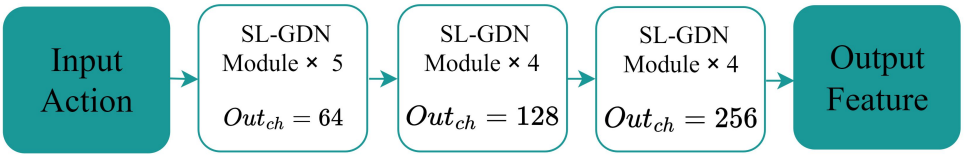

Figure 5. The backbone of the proposed SL-GDN network, including 13 SL-GDN blocks.

Finally, we employed a classification module, including the NN to refine the final feature. After averaging the refined features, the feature matrix is converted into a vector and a fully connected layer is employed, based on the number of classes, which is denoted by C in Figure 4c.

4.5. Four-Stream Approach

To overcome the problems related to the irrelevant features in skeleton-based SLR systems, we employed first- and second-order representations of the skeleton points, namely, joint coordinates and bone coordinates, as well as their respective motion vectors [11,52,53]. Figure 3 depicts the multi-four stream SL-GDN, using joint, joint motion, bone, and bone motion data. From the joint data (in vector form), bone data are indicated from source to target joints based on the natural connections in the human body. Here, we consider the nose a zero-number joint—known as a root joint of the human body—and the bone data for the nose is 0. Assume that the source and target joints can be expressed as $v^J_{p,t} = (x_{p,t}, y_{p,t}, S_{p,t})$ and $v^J_{q,t} = (x_{q,t}, y_{q,t}, S_{q,t})$, where the $x - y$ score and confidence score are represented by x, y, and S. The bone vectors can be calculated by subtracting the source joint and target joint, as $v^B_{p,t} = (x_{p,t} - x_{q,t}, y_{p,t} - y_{q,t}, S_{p,t})$. Here, (p, q) denotes the set of key point joints with respect to face, body, and hand pose. The difference between adjacent frames is used to obtain the motion data for both joints and bones [8]. Based on the mentioned formula, we can calculate the joint motion as follows: $v^{J,M}_{p,t} = (x_{p,t} - x_{p,t+1}, y_{p,t} - y_{p,t+1}, S_{p,t})$. In the same way, bone motion can be calculated as $v^{B,M}_{p,t} = v^B_{p,t} - v^B_{p,t+1}$. We trained each stream with individual data to produce relevant features and, finally, we concatenated all four features to produce the final feature vector [20,21]

4.6. Classification Module

After concatenating the four stream features, we constructed a final feature vector and applied a classification module for prediction. The classification module consists of two parts. The first part includes a channel attention module [62] to refine the temporal features, as demonstrated in detail in Figure 4d. In the second part, a fully connected layer with ReLU and dropout layers is applied, including averaging of the matrix into a vector, and a fully connected layer with several classes for classification is employed.

5. Experimental Results

We conducted sign language classification experiments on two large-scale data sets, in order to investigate the superiority and effectiveness of the proposed model. We detail its performance accuracy first, then provide a comparison with a state-of-the-art model.

5.1. Experimental Setting

We used two large-scale data sets, CSL and AUTSL, to evaluate the proposed model. To divide the data into training and testing sets, we followed the strategy detailed in a recent study [8]. For the AUTSL data set, we used 28,146 videos as a training set and 3742 videos for the testing set, where each video contained 150 frames. For the CSL

data set, we used 100,000 videos for training and 25,000 videos for testing, where each video contained 160 frames. Each data set provided four streams: joint, bone, joint motion, and bone motion. Each stream was used to produce individual features through the SL-GDN model, following which the four features were concatenated. Finally, we refined and classified the final feature using the classification module. To implement the proposed system, we used the Google Colab environment and the Python programming language. For the framework, we used the Pytorch [63] library for Python in the Google Colab Pro edition environment, which provided us with a Tesla 100 machine (Tesla, Inc., Austin, TX, USA) having 25 GB of GPU processing power [64]. Pytorch is an open-source library that effectively provides attention, transformer, and deep learning models, requiring low computational cost while maintaining high compatibility and adaptability properties with minimum resources. In addition, we used the OpenCV, open pose, pickle, and csv packages for the initial processing [65,66]. The main goal of the pickle package is to convert a data set into a byte stream for portable storage. We used the Numpy and Pandas packages as they provide increased flexibility regarding matrix multiplication and other operations facilitating statistical and mathematical procedures. We used initial learning to reduce the high fluctuation rate and sped up the convergence of the training and testing processes using the Adam optimizer [66]. We set 1000 epochs for tuning the model, considering various parameter tuning operations for the learning rate and optimization with respect to the multiple classes considered in this study.

5.2. Performance Accuracy of SL-GDN on Benchmark Data Set

Table 2 details the classification accuracy of the proposed model, including its performance on the AUTSL and CSL data sets. We tested the performance under the individual four streams and the multi-stream of the proposed model. Table 2 reports 96.00% testing accuracy for the AUTSL data with joint information and 95.00%, 94.00%, 93.00%, and 96.00% testing accuracy when using the joint motion, bone, bone motion, and multi-stream key point data, respectively. In the same way, on CSL, 88.70%, 87.00%, 86.00%, and 89.45% testing accuracy was achieved with the joint, joint motion, bone, bone motion, and multi-stream key points, respectively.

Table 2. Performance of the proposed model on the two data sets.

Stream	Testing Accuracy on AUTSL [%]	Testing Accuracy on CSL [%]
Joint	96.00	88.70
Joint Motion	95.00	87.00
Bone	94.00	86.00
Bone Motion	93.00	86.50
Multi-Stream	96.00	89.45

5.3. Ablation Study

Our proposed model is composed of three main modules with four stream structures, including a neural network module, an SL-GDN module, and a classification module. In the SL-GDN module, we used many SL-GDN blocks (as shown in Figure 5), and there were various stages in the channel attention block of the classification module (as shown in Figure 4c). From the table below, it can be seen that optimizing the number of SL-GDN and channel module attention blocks could be beneficial for sign language recognition accuracy, whereas other configurations may have a detrimental effect on performance. Table 3 details the average performance obtained on both data sets in the ablation study. We focused the ablation study by following the concept of the two existing skeleton-based GCN architecture [8,55] methods and a combination of channel attention method [67]. The author in [55] applied GCN with nine series blocks of the spatial-temporal architecture, and the author in [8] applied GCN with ten series blocks of the spatial-temporal architecture. They did not include an attention model in their architecture. On the other hand, we achieved maximum accuracy with 13 series of spatial-temporal GDN blocks with two-

channel attention modules. The performance table proves that our proposed combination reported high performance compared to the existing module.

Table 3. Validation accuracy when tuning the number of SL-GDN and channel attention blocks in series.

Method Name	No. of SL-GDN in Series	No.of Channel Attention Newly Added	Datasets	Performance Accuracy [%]
Two Stream [55]	9	1	AUTSL	94.00
Two Stream [55]	9	1	CSL	88.00
Four Stream [8]	10	2	AUTSL	96.00
Four Stream [8]	10	2	CSL	88.50
Proposed Four Stream	13	2	AUTSL	96.45
Proposed Four Stream	13	2	CSL	89.45

5.4. Comparison of the Proposed Model with State-of-the-Art on the AUTSL Data Set

Table 4 compares the proposed model with the state-of-the-art model proposed by Jiang [8]. The proposed model obtained 96.00% accuracy under the multi-stream modality, whereas the state-of-the-art model obtained 95.54% accuracy. Jiang et al. have proposed various models with various criteria and key points based on the one method detailed in [8]. They tested these various models, one of which achieved 95.02%, 94.70%, 93.10%, 92.49%, and 95.45% accuracy with the use of joint, joint motion, bone, bone motion, and multi-stream key points, respectively.

Table 4. Comparison with a state-of-the-art approach on the AUTSL data set.

Data Set Type	Method Name	Sign Recognition Accuracy [%]
RGB+Depth	CNN+FPM+LSTM+Attention [57]	83.93
Skeleton	Two Stream CNN [55]	93.70
Skeleton Joint	Jiang [8]	95.02
Skeleton Joint Motion	Jiang [8]	94.70
Skeleton Bone	Jiang [8]	93.10
Skeleton Bone Motion	Jiang [8]	92.49
Skeleton Multi-Stream	Jiang [8]	95.45
Skeleton Joint	Proposed Model	96.00
Skeleton Joint Motion	Proposed Model	95.00
Skeleton Bone	Proposed Model	94.00
Skeleton Bone Motion	Proposed Model	93.00
Skeleton Multi-Stream	Proposed Model	96.45

5.5. Comparison of the Proposed Model with State-of-the-Art on the CSL Data Set

Table 5 provides a comparison of the proposed model with a state-of-the-art model on the CSL dataset, where the reported performance accuracy for our proposed model was 89.45%, whereas that for the existing 3D-CNN [44] model was 88.70%. According to Tables 2, 4 and 5, we can conclude that our approach seems to establish new baselines for state-of-the-art sign recognition performance on the AUTSL and CSL data sets.

Table 5. Comparison with a state-of-the-art approach on the CSL data set.

Data Set Name	Data Set Type	Methodology	Sign Recognition Accuracy [%]
CSL	RGB-D+Skeleton	3D-CNN [44]	88.70
Proposed Model	Skeleton	SL-GDN	89.45

6. Conclusions

In this study, we proposed a multi-stream graph-based deep neural network (MSL-GDN) for a skeleton-based SLR system, in which we consider four modalities derived from a skeleton-based sign language data set. Specifically, we constructed a graph—known as a

skeleton graph—for the whole-body pose key points, then applied MSL-GDN to compute the spatial and temporal features. We extracted individual features from each stream and, finally, concatenated the features and applied the classification module to refine the feature vector and carry out classification. The performance testing results demonstrated the superiority and generalizability of the proposed model, considering its high accuracy on the large-scale AUTSL and CSL data sets. The reason for the high generalizability of the proposed model is that we selected 27 whole-body key points among the 133 body pose key points and extracted features from bone data and joint and bone motion streams. The main limitation of this study is that we only used four streams of the network to improve the effectiveness of the features and performance accuracy. We plan to combine the skeleton's final features with other modalities present in the data sets, including RGB and depth data. In addition, we intend to work towards calculating the inverse dynamics from videos with different pose models, allowing us to apply the MSL-GDN model to the inverse dynamic features.

Author Contributions: Conceptualization, A.S.M.M.; methodology, A.S.M.M., S.-W.J., H.-S.L., M.A.M.H. and J.S.; investigation, A.S.M.M., M.A.M.H. and J.S.; data curation, A.S.M.M., M.A.M.H., S.-W.J., H.-S.L. and J.S.; writing—original draft preparation, A.S.M.M. and J.S.; writing—review and editing, A.S.M.M. and J.S.; visualization, A.S.M.M. and M.A.M.H.; supervision, J.S.; funding acquisition, S.-W.J., H.-S.L. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research work is supported by the Ministry of Science and ICT (MSIT), Korea, under the Grand Information Technology Research Center support program (IITP-2023-2020-0-01791) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work was supported by the Competitive Research Fund of The University of Aizu, Japan.

Data Availability Statement: In the study we used two datasets which are available online: **AUTSL**: https://drive.google.com/drive/folders/1VUQsh_nf70slT4YsC-UzTCAZ3jB_uFKX. **CSL**: hagjie@mail.ustc.edu.cn. **Preprocess data of WLASL, AUTSL, CSL**: https://drive.google.com/drive/folders/1VUQsh_nf70slT4YsC-UzTCAZ3jB_uFKX.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Miah, A.S.M.; Hasan, M.A.M.; Shin, J. Dynamic Hand Gesture Recognition using Multi-Branch Attention Based Graph and General Deep Learning Model. *IEEE Access* **2023**, *11*, 4703–4716. [CrossRef]
2. Miah, A.S.M.; Hasan, M.A.M.; Shin, J.; Okuyama, Y.; Tomioka, Y. Multistage Spatial Attention-Based Neural Network for Hand Gesture Recognition. *Computers* **2023**, *12*, 13. [CrossRef]
3. Miah, A.S.M.; Shin, J.; Hasan, M.A.M.; Rahim, M.A. BenSignNet: Bengali Sign Language Alphabet Recognition Using Concatenated Segmentation and Convolutional Neural Network. *Appl. Sci.* **2022**, *12*, 3933. [CrossRef]
4. Miah, A.S.M.S.J.; Hasan, M.A.M.; Rahim, M.A.; Okuyama, Y. Rotation, Translation And Scale Invariant Sign Word Recognition Using Deep Learning. *Comput. Syst. Sci. Eng.* **2023**, *44*, 2521–2536. [CrossRef]
5. Miah, A.S.M.; Shin, J.; Islam, M.M.; Molla, M.K.I. Natural Human Emotion Recognition Based on Various Mixed Reality (MR) Games and Electroencephalography (EEG) Signals. In Proceedings of the 2022 IEEE 5th Eurasian Conference on Educational Innovation (ECEI), Taipei, Taiwan, 10–12 February 2022; IEEE: New York, NY, USA, 2022; pp. 408–411.
6. Miah, A.S.M.; Mouly, M.A.; Debnath, C.; Shin, J.; Sadakatul Bari, S. Event-Related Potential Classification Based on EEG Data Using xDWAN with MDM and KNN. In Proceedings of the International Conference on Computing Science, Communication and Security, Gujarat, India, 6–7 February 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 112–126.
7. Emmorey, K. *Language, Cognition, and the Brain: Insights from Sign Language Research*; Psychology Press: London, UK, 2001.
8. Jiang, S.; Sun, B.; Wang, L.; Bai, Y.; Li, K.; Fu, Y. Skeleton aware multi-modal sign language recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3413–3423.
9. Yang, Q. Chinese sign language recognition based on video sequence appearance modeling. In Proceedings of the 2010 5th IEEE Conference on Industrial Electronics and Applications, Taichung, Taiwan, 5–17 June 2010; IEEE: New York, NY, USA, 2010; pp. 1537–1542.
10. Valli, C.; Lucas, C. *Linguistics of American Sign Language: An Introduction*; Gallaudet University Press: Washington, DC, USA, 2000.
11. Mindess, A. *Reading between the Signs: Intercultural Communication for Sign Language Interpreters*; Nicholas Brealey: Boston, MA, USA, 2014.
12. Shin, J.; Musa Miah, A.S.; Hasan, M.A.M.; Hirooka, K.; Suzuki, K.; Lee, H.S.; Jang, S.W. Korean Sign Language Recognition Using Transformer-Based Deep Neural Network. *Appl. Sci.* **2023**, *13*, 3029. [CrossRef]

13. Miah, A.S.M.; Shin, J.; Hasan, M.A.M.; Molla, M.K.I.; Okuyama, Y.; Tomioka, Y. Movie Oriented Positive Negative Emotion Classification from EEG Signal using Wavelet transformation and Machine learning Approaches. In Proceedings of the 2022 IEEE 15th International Symposium on Embedded Multicore/Many-Core Systems-on-Chip (MCSoC), Penang, Malaysia, 19–22 December 2022; IEEE: New York, NY, USA, 2022; pp. 26–31.
14. Miah, A.S.M.; Rahim, M.A.; Shin, J. Motor-imagery classification using Riemannian geometry with median absolute deviation. *Electronics* **2020**, *9*, 1584. [CrossRef]
15. Miah, A.S.M.; Islam, M.R.; Molla, M.K.I. Motor imagery classification using subband tangent space mapping. In Proceedings of the 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 22–24 December 2017; IEEE: New York, NY, USA, 2017; pp. 1–5.
16. Zobaed, T.; Ahmed, S.R.A.; Miah, A.S.M.; Binta, S.M.; Ahmed, M.R.A.; Rashid, M. Real time sleep onset detection from single channel EEG signal using block sample entropy. *Iop Conf. Ser. Mater. Sci. Eng.* **2020**, *928*, 032021. [CrossRef]
17. Kabir, M.H.; Mahmood, S.; Al Shiam, A.; Musa Miah, A.S.; Shin, J.; Molla, M.K.I. Investigating Feature Selection Techniques to Enhance the Perfor-mance of EEG-Based Motor Imagery Tasks Classification. *Mathematics* **2023**, *11*, 1921. [CrossRef]
18. Miah, A.S.M.; Islam, M.R.; Molla, M.K.I. EEG classification for MI-BCI using CSP with averaging covariance matrices: An experimental study. In Proceedings of the 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 11–12 July 2019; IEEE: New York, NY, USA, 2019; pp. 1–5.
19. Joy, M.M.H.; Hasan, M.; Miah, A.S.M.; Ahmed, A.; Tohfa, S.A.; Bhuaiyan, M.F.I.; Zannat, A.; Rashid, M.M. Multiclass mi-task classification using logistic regression and filter bank common spatial patterns. In Proceedings of the Computing Science, Communication and Security: First Interna-tional Conference, COMS2 2020, Gujarat, India, 26–27 March 2020; Revised Selected Papers; Springer: Berlin/Heidelberg, Germany, 2020; pp. 160–170.
20. Cheng, K.; Zhang, Y.; Cao, C.; Shi, L.; Cheng, J.; Lu, H. Decoupling gcn with dropgraph module for skeleton-based action recognition. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXIV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 536–553.
21. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Trans. Image Process.* **2020**, *29*, 9532–9545. [CrossRef] [PubMed]
22. Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Stronger, faster and more explainable: A convolutional graph baseline for skeleton-based action recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1625–1633.
23. Wang, H.; Wang, L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 499–508.
24. Yan, S.; Xiong, Y.; Lin, D. Spatial, temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Orleans, LA, USA, 2–7 February 2018; Volume 32.
25. Oberweger, M.; Lepetit, V. Deepprior++: Improving fast and accurate 3d hand pose estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 585–594.
26. Shin, J.; Matsuoka, A.; Hasan, M.A.M.; Srizon, A.Y. American sign language alphabet recognition by extracting feature from hand pose estimation. *Sensors* **2021**, *21*, 5856. [CrossRef]
27. Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; Luo, P. Whole-body human pose estimation in the wild. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IX 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 196–214.
28. Xiao, Q.; Qin, M.; Yin, Y. Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Netw.* **2020**, *125*, 41–55. [CrossRef]
29. Mejía-Peréz, K.; Córdova-Esparza, D.M.; Terven, J.; Herrera-Navarro, A.M.; García-Ramírez, T.; Ramírez-Pedraza, A. Automatic recognition of Mexican Sign Language using a depth camera and recurrent neural networks. *Appl. Sci.* **2022**, *12*, 5523. [CrossRef]
30. Jiang, S.; Sun, B.; Wang, L.; Bai, Y.; Li, K.; Fu, Y. Sign language recognition via skeleton-aware multi-model ensemble. *arXiv* **2021**, arXiv:2110.06161.
31. Lim, K.M.; Tan, A.W.C.; Lee, C.P.; Tan, S.C. Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. *Multimed. Tools Appl.* **2019**, *78*, 19917–19944. [CrossRef]
32. Shi, B.; Del Rio, A.M.; Keane, J.; Michaux, J.; Brentari, D.; Shakhnarovich, G.; Livescu, K. American sign language fingerspelling recognition in the wild. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; IEEE: New York, NY, USA, 2018; pp. 145–152.
33. Li, Y.; Wang, X.; Liu, W.; Feng, B. Deep attention network for joint hand gesture localization and recognition using static RGB-D images. *Inf. Sci.* **2018**, *441*, 66–78. [CrossRef]
34. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Corfu, Greece, 20–25 September 1999; IEEE: New York, NY, USA, 1999; Volume 2, pp. 1150–1157.
35. Zhu, Q.; Yeh, M.C.; Cheng, K.T.; Avidan, S. Fast human detection using a cascade of histograms of oriented gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; IEEE: New York, NY, USA, 2006; Volume 2, pp. 1491–1498.

36. Dardas, N.H.; Georganas, N.D. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Trans. Instrum. Meas.* **2011**, *60*, 3592–3607. [CrossRef]
37. Memiş, A.; Albayrak, S. A Kinect based sign language recognition system using spatio-temporal features. In Proceedings of the Sixth International Conference on Machine Vision (ICMV 2013), London, UK, 16–17 November 2013; Volume 9067, pp. 179–183.
38. Rahim, M.A.; Miah, A.S.M.; Sayeed, A.; Shin, J. Hand gesture recognition based on optimal segmentation in human-computer interaction. In Proceedings of the 2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII), Kaohsiung, Taiwan, 21–23 August 2020; IEEE: New York, NY, USA, 2020; pp. 163–166.
39. Tur, A.O.; Keles, H.Y. Isolated sign recognition with a siamese neural network of RGB and depth streams. In Proceedings of the IEEE EUROCON 2019-18th International Conference on Smart Technologies, Novi Sad, Serbia, 1–4 July 2019; IEEE: New York, NY, USA, 2019; pp. 1–6.
40. Cai, Z.; Wang, L.; Peng, X.; Qiao, Y. Multi-view super vector for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 596–603.
41. Neverova, N.; Wolf, C.; Taylor, G.; Nebout, F. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1692–1706. [CrossRef]
42. Pu, J.; Zhou, W.; Li, H. Iterative alignment network for continuous sign language recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4165–4174.
43. Koller, O.; Zargaran, S.; Ney, H.; Bowden, R. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. *Int. J. Comput. Vis.* **2018**, *126*, 1311–1325. [CrossRef]
44. Huang, J.; Zhou, W.; Li, H.; Li, W. Attention-based 3D-CNNs for large-vocabulary sign language recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 2822–2832. [CrossRef]
45. Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; Saenko, K. Sequence to sequence-video to text. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4534–4542.
46. Pigou, L.; Van Den Oord, A.; Dieleman, S.; Van Herreweghe, M.; Dambre, J. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *Int. J. Comput. Vis.* **2018**, *126*, 430–439. [CrossRef]
47. Huang, J.; Zhou, W.; Zhang, Q.; Li, H.; Li, W. Video-based sign language recognition without temporal segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Orleans, LA, USA, 2–7 February 2018, Volume 32.
48. Li, D.; Rodriguez, C.; Yu, X.; Li, H. 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, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1459–1469.
49. Cui, R.; Liu, H.; Zhang, C. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Trans. Multimed.* **2019**, *21*, 1880–1891. [CrossRef]
50. Guo, D.; Zhou, W.; Li, H.; Wang, M. Hierarchical LSTM for sign language translation. In Proceedings of the AAAI Conference on Artificial Intelligence, Orleans, LA, USA, 2–7 February 2018; Volume 32.
51. Parelli, M.; Papadimitriou, K.; Potamianos, G.; Pavlakos, G.; Maragos, P. Exploiting 3D Hand Pose Estimation in Deep Learning-Based Sign Language Recognition from RGB Videos. In Proceedings of the Computer Vision—ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Bartoli, A., Fusiello, A., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 249–263.
52. Cai, J.; Jiang, N.; Han, X.; Jia, K.; Lu, J. JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF winter conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2021; pp. 2734–2743. [CrossRef]
53. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3595–3603.
54. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5457–5466.
55. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
56. de Amorim, C.C.; Macêdo, D.; Zanchettin, C. Spatial-temporal graph convolutional networks for sign language recognition. In Proceedings of the Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, 17–19 September 2019; Proceedings 28; Springer: Berlin/Heidelberg, Germany, 2019; pp. 646–657.
57. Sincan, O.M.; Keles, H.Y. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access* **2020**, *8*, 181340–181355. [CrossRef]
58. Huang, J. Chinese Sign Language Recognition Dataset. 2017. Available online: http://home.ustc.edu.cn/~hagjie/ (accessed on 23 June 2023).
59. Sincan, O.M.; Tur, A.O.; Keles, H.Y. Isolated sign language recognition with multi-scale features using LSTM. In Proceedings of the 2019 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, 24–26 April 2019; IEEE: New York, NY, USA, 2019; pp. 1–4.

60. Pagliari, D.; Pinto, L. Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors. *Sensors* **2015**, *15*, 27569–27589. [CrossRef]
61. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part III 14; Springer: New York, NY, USA, 2016; pp. 816–833.
62. Hirooka, K.; Hasan, M.A.M.; Shin, J.; Srizon, A.Y. Ensembled Transfer Learning Based Multichannel Attention Networks for Human Activity Recognition in Still Images. *IEEE Access* **2022**, *10*, 47051–47062. [CrossRef]
63. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems 32, Vancouver, BC, Canada, 8–14 December 2019.
64. Tock, K. Google CoLaboratory as a platform for Python coding with students. *RTSRE Proc.* **2019**, *2*. Available online: https://www.rtsre.org/index.php/rtsre/article/view/63 (accessed on 23 June 2023).
65. Gollapudi, S. *Learn Computer Vision using OpenCV*; Springer: Berlin/Heidelberg, Germany, 2019.
66. Dozat, T. Incorporating Nesterov Momentum into Adam 2016. Available online: https://cs229.stanford.edu/proj2015/054_report.pdf (accessed on 23 June 2023).
67. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.