GLA-DenseNet: An Advanced Convolutional Approach for ISLR Leveraging 3D Pose Estimation and Attention Mechanisms

Chengyong Cui, Xiangjie Kong, *Senior Member, IEEE,* and Wenyi Zhang

Abstract—Isolated sign language recognition (ISLR) seeks to convert gesture sequences from videos into text or glosses, bridging communication between hearing individuals and those who are deaf or hard-of-hearing without sign language proficiency. We introduce GLA-DenseNet, a convolutional neural network blending the DenseNet architecture with attention mechanisms, global long-sequence attention (GLAN), and 3D human pose estimation to enhance ISLR robustness. While DenseNets effectively parse representations from skeleton image sequences, our attention blocks hone in on crucial spatial regions, further augmented by GLAN. Dual-stream spatio-temporal transformer (DSTformer) elevates 2D poses into a 3D space, amplifying geometric intricacies. Moreover, a tree-structured skeleton image (TSSI) encoding preserves spatial joint relationships through a depth-first search order. Testing on three prominent sign language datasets—WLASL, AUTSL, and LSM—revealed that GLA-DenseNet surpasses contemporary skeleton-centric methods. With data augmentation, it even outperforms certain RGB-based techniques, all while maintaining a streamlined architecture. Ablation tests affirm the value added by 3D pose estimation and augmentation. In essence, our findings underscore the viability of CNN-centric models in delivering top-tier ISLR results when amalgamating diverse methodologies such as attention, 3D conversion, and intricate skeleton encoding, making ISLR systems more ripe for real-world application.

Index Terms—Sign language recognition, Attention, CNN, 3D human pose estimation

1 Introduction

In the past few years, Human Action Recognition (HAR) has garnered attention in the research community, prominently led by advances in Sign Language Recognition (SLR). Specifically, Isolated Sign Language Recognition (ISLR) aims to translate videos of gesture sequences into real-world language and glosses (the term referring to individual sign language words). The importance of ISLR extends beyond academic interest, having substantial societal implications[1]. For instance, incorporating a proficient ISLR model into video conferences could bridge the communication barrier between hearing individuals and those not versed in sign language.

Deep learning (DL) techniques have substantially advanced the field of ISLR. Among various methods, Convolutional Neural Networks (CNN) have emerged as particularly influential[2][3]. Their exceptional ability to learn from 3D arrays has inspired researchers to represent skeleton sequences as grayscale images, a move that facilitates a comprehensive grasp of the spatio-temporal dynamics[4][5]. Within these images, each row encapsulates the spatial distribution of coordinates for a specific moment in time, while each column tracks the temporal progression of particular joints. Each image then represents a specific axis of joint coordinates; for instance, the x-axis of the skeleton sequence is visualized as a single-channel image. This transformation yields what is known as a "skeleton image"—a multi-channel representation seamlessly processed by CNN techniques like the Dense Convolutional Network (DenseNet)[6], which boasts a stellar reputation in image classification[7].

Nevertheless, a limitation arises as some CNN-based methods might not adequately focus on crucial spatial or temporal segments, potentially compromising recognition accuracy[8]. A remedy to this oversight lies in integrating attention mechanisms. Incorporating specific tools like attention masks[9][10] can enhance the emphasis on salient features within skeleton images.

Leveraging skeleton-based features presents a host of advantages. Foremost among them is their resilience in challenging scenarios, such as intricate backgrounds and inconsistent lighting, all while retaining a streamlined structure that necessitates fewer parameters[11]. Another intriguing dimension in this realm is 3D human pose estimation, a cutting-edge research avenue focused on extracting skeletal data from visual inputs. By adding a third dimension to the conventional 2D skeletons, this technique amplifies their geometric intricacies, bestowing an added layer of rich information for CNN-based models[12]. The Dual-stream Spatio-temporal Transformer (DSTformer)[12] stands out as a bespoke solution for 2D-to-3D conversion tasks. By integrating this technique, we fortified our GLA-DenseNet architecture with depth nuances, a move that bore fruit in our experimental findings.

The sequence in which joints are presented holds significant information, reflecting the spatial layout of a skeleton. A fixed sequence can potentially result in significant data loss. Thus, this study emphasizes the importance of the Tree Structure Skeleton Image (TSSI), which aims to capture and

- *Chengyong Cui, Xiangjie Kong and Wenyi Zhang are with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China. E-mail: 202003151204@zjut.edu.cn, xjkong@ieee.org,wyzhang98@outlook.com.*

maintain these spatial relationships[8] by ordering joints in a depth-first manner.

- Introduction of the GLA-DenseNet network to address ISLR challenges. By integrating attention blocks, this network refines features and zooms in on the most informative regions in skeleton images.
- Incorporation of the DSTformer for 2D-to-3D lifting, enabling the extraction of depth information.
- A comprehensive evaluation of the proposed model across three language datasets (WLSAL, AUTSL, and LSM). Our experiments attest to the method's effectiveness and robustness.

The structure of this paper unfolds as such: Section 2 delves into relevant literature. In Section 3, we elucidate the intricacies of our proposed GLA-DenseNet network. Section 4 offers an exhaustive assessment of GLA-DenseNet across three distinct datasets. The paper culminates with a conclusion in Section 5.

2 Related Work

2.1 CNN based method

Human Action Recognition (HAR) has seen an influx of methods that represent skeleton sequences as images, which are then processed using convolutional Neural Networks.

The pioneering effort in this direction was evidenced in [13], where human body keypoints were segmented into five distinct groups and subsequently organized into a vector. This innovative approach visualized a skeleton sequence from multiple frames as an RGB image, where the (x,y,z) coordinates found representation in the (r,g,b) channels. Parallel research efforts have proposed alternative CNN-processing methods for these sequences. Examples include [14]'s approach of mapping skeleton joints to CNN channels, [15]'s method of incorporating angles and distances across frames, and [16]'s integration of skeletons with heatmaps. One noteworthy advancement came from [8], which introduced the TSSI method. Using depth-first-search (DFS) on a basic skeleton graph, this method determined an order to accurately represent relationships between keypoints.

The CNN-based processing approach was further enriched by [17]'s investigations into augmentation functions and the application of 3D heatmap volumes.

2.2 Attention

Attention mechanisms have found compelling applications in various fields, including image captioning[9][18], RGB-based action recognition[10][19], image classification[20][21], and sentiment analysis[22]. Many of these attention methodologies leverage image sequences as input[10], while others harness supplementary information from modalities such as text[9][18][19]. In skeleton-based recognition, where a single skeleton image embodies a spatio-temporal sequence, [8]proposed a unique frame-based visual attention model. This approach highlights how generic visual attention can produce 2D attention masks to accentuate spatial and temporal significance. In this setup, each row illustrates the spatial relevance, while each column conveys the temporal value.

2.3 3D human pose estimation

Mediapipe, though adept at extracting gesture features from videos, has been critiqued for its less-than-reliable z-axis estimation[23]. The model yields 2.5D normalized pixel coordinates for landmarks, referencing the top-left image corner. The depth ("0.5" in "2.5D") is gauged against the wrist point.

3D human pose estimation, which aspires to determine 3D human postures from RGB videos, has long been a challenge. Two primary strategies are discerned. One directly derives the 3D pose from images using CNN[24][25][26]. This approach, however, confronts the challenge of balancing 3D pose accuracy against appearance variability, an outcome of prevailing data collection methodologies. The second strategy extracts the 2D pose initially and then transitions this 2D estimate to 3D via a distinct neural network. Such a conversion or "lifting" can be effectuated through Fully Connected Networks[27][28], Temporal Convolutional Networks (TCN)[29][30], GCN[31][32][33], and Transformers[34][35][36][37].

3 Methodology

3.1 Order of joints

We represented a base human skeleton using a graph notation, $G = (E, V)$, where V denotes nodes corresponding to skeletal joints and E signifies edges connecting these nodes such that $e = (v_i, v_j)$, $e \in E$, and $v_i, v_j \in V$. This representation takes inspiration from the MediaPipe holistic model[23] which identifies 543 keypoints characterizing the joints of the human body and delineates their interconnections. Our study, however, utilized a subset of 68 keypoints, echoing the findings of [7]. As illustrated in Figure 1, this subset consists of 6 body keypoints, 20 face keypoints, 21 left-hand keypoints, and 21 right-hand keypoints. The face keypoints include four points for each eyebrow, eyes, and the mouth, which are emphasized due to their significance [38]. We converted the base skeleton graph into a tree structure, designating the midpoint of the shoulders as the root. Using DFS, we acquired a sequence of joints based on their visitation order, resulting in a 135-joint ordered list(as in Figure 1). This list facilitated the creation of a TSSI in subsequent stages.

3.2 Tree Structure Skeleton Image(TSSI)

Conventional skeleton images face a drawback in their structure, with each row being determined by a fixed and often arbitrary order. CNN's inherent trait is its expanding receptive field at higher levels, implying that adjacent joints learned at initial stages share more spatial relations. However, this is not consistently true for the original skeleton structure.

For any given video, we extracted its skeleton frame data using MediaPipe, thereafter leveraging the DFS-derived joint sequence to produce a TSSI, depicted in Figure 2. We omitted frames where body pose estimation by MediaPipe was unsuccessful. In instances where hand or face coordinates were undetectable, we substituted them with wrist and nose coordinates, respectively. In the TSSI matrix, rows represent frame-specific skeleton data, columns symbolize

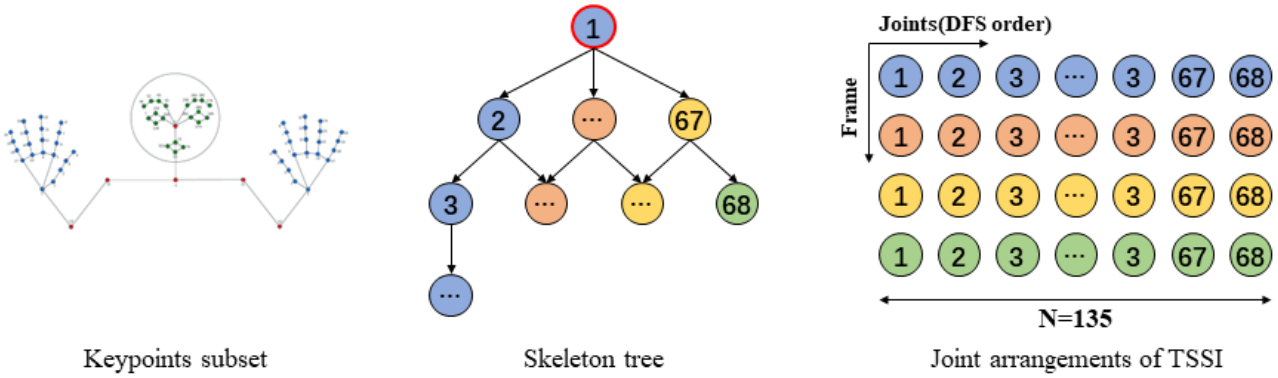

Fig. 1. A subset of keypoints (left) is transmuted into a skeleton tree (middle) with its inception at the root point (highlighted in red). Following the traversal pattern of DFS, the joints are methodically arranged into a list (right).

DFS-ordered joint data, and channels (r, g, b) encapsulate the (x, y, z) joint coordinates. Specifically, a TSSI I is represented as $I = [p_{1,1}, p_{i,j}, \ldots, p_{N,T}]$. Each $p_{i,j}$ corresponds to the pixel describing the (x, y, z) coordinates across the (r, g, b) channels for the joint at v_i in frame j.

3.3 GLA-DenseNet

3.3.1 DenseNet

Our approach leverages DenseNet as the foundational convolutional model, advocating that CNN-based methodologies excel at deciphering the spatial-temporal interrelationships among joints. Specifically, we utilized the DenseNet-121[39], a renowned deep learning blueprint optimized for image classification. This architecture boosts feature reuse and gradient flow via its unique composition of dense blocks and transitional layers that performs down-sampling via convolution and pooling. Our implementation is sourced from the Keras library[40].

3.3.2 Attention Networks

Base Attention Block: Recognizing the inherent representation of both spatial and temporal skeleton sequence data in skeleton images, we devised a dual-branch attention architecture to generate attention masks. This architecture comprises a "mask branch" dedicated to extracting attention masks, and a "residual branch" aimed at refining the features derived from previous CNN layers. As illustrated in Figure 3, these branches collaboratively produce a weighted CNN feature block. Our introductory model for this architecture is the base attention model, representing the foundational version of the two-branch attention structure. As depicted in Figure 3, the mask branch in this base model utilizes a single convolutional layer to achieve an expanded receptive field. Either Softmax or Sigmoid functions are employed for mask creation. The residual branch, on the other hand, ensures the preservation of input CNN features through a direct connection. This structure harmoniously fuses our bespoke attention blocks with the convolutional blocks of DenseNet.

Global Long-Sequence Attention Network (GLAN): Building on the foundational two-branch system described earlier, the GLAN, illustrated in Figure 3, incorporates an hourglass structure[21] into its mask branches. This facilitates faster feature size adjustments and enhances the receptive field. Within each hourglass mask branch, input CNN features undergo down-sampling to a minimum spatial resolution of 7×7, before being restored to their original size. Every down-sampling unit comprises a max pooling layer, followed by a residual unit, and a linked connection to the recovered feature of identical size. Conversely, each up-sampling unit encompasses a bilinear interpolation layer, a residual unit, and an element-wise sum with the linked connection. Echoing the findings of [8], the Convolution-Deconvolution structure effectively enlarges the receptive field, optimizing the learning of an attention mask. Regarding the residual branches, two additional residual units have been incorporated to further refine the CNN features. As depicted in Figure 4, the GLAN network was constructed

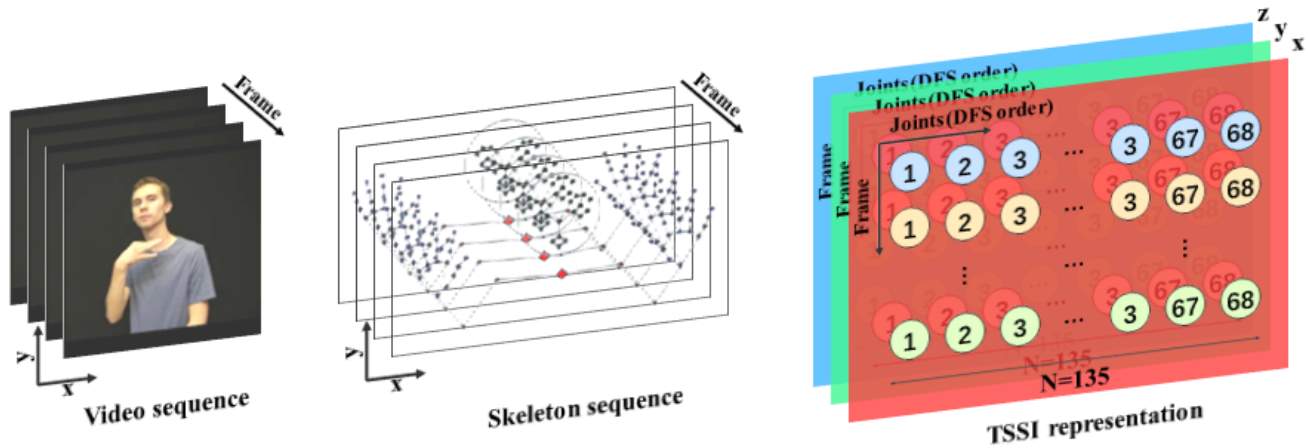

Fig. 2. Procedure for Generating TSSI Representation: For each frame of the video, skeleton data is extracted using MediaPipe. An image is then constructed where joints are arranged in the column dimension according to their DFS order. Sequential frames are aligned in the row dimension, while the coordinates (x, y, z) of the joints populate the channel dimension.

by inserting three GLAN attention blocks between the four dense blocks of DenseNet-121. Owing to varying input feature sizes, the depth of each GLAN block was adjusted accordingly. Additionally, to maintain an appropriate depth for the GLAN network, especially considering the increased depth of GLAN blocks compared to base attention blocks, the number of residual units in each block was reduced.

In essence, the GLAN model seamlessly integrates three attention blocks within the four dense blocks of DenseNet-121, creating a sophisticated, well-orchestrated network. Each GLAN block's depth varies due to different input feature sizes. Only a single convolutional layer was retained for the initial three dense blocks, while the final dense block preserved all convolutional layers as found in DenseNet-121.

Considering the current unreliability of the MediaPipe model's z-axis estimations, we adopted the DSTformer technique, endorsed by [12], to transform 2D skeletons into 3D.

3.4 3D human pose estimation

Although our GLA-DenseNet model processes tri-channel inputs for the x, y, and z axes, the z-axis estimates from MediaPipe are found to be erratic. To rectify this, we utilize the DSTformer model, aligning with the techniques highlighted in [12], to facilitate the 2D to 3D skeleton transition.

Our methodology also integrates supervision signals, a strategy that has gained traction in both linguistic[41][42][43] and visual domains[44][45]. Notably, a mask-based input technique helps recover depth information lost during 2D visual analyses. We initiate this process by obtaining 2D skeleton sequences, denoted as x, via orthogonal 3D motion projection. Following this, we introduce imperfections into x by random masking and noise addition, resulting in the corrupted 2D skeleton sequences. These sequences aptly emulate 2D detection outcomes, characterized by occlusions, detection inaccuracies, and other errors. We introduce masks at both the joint and frame levels, each with specific probability rates. The motion encoder cited in [32] aids in obtaining the motion representation E and reconstructing the 3D motion as $\hat{X}$. The joint loss, $\mathcal{L}_{3D}$, is ascertained by juxtaposing $\hat{X}$ with the 3D motion ground truth, X. Furthermore, in alignment with preceding studies[30][36], we amalgamate the velocity loss, $\mathcal{L}_O$. The equations for the 3D reconstruction losses are represented as:

$$\mathcal{L}_{3D} = \sum_{t=1}^{T} \sum_{j=1}^{J} ||\hat{X}_{t,j} - X_{t,j}||_2$$

and

$$\mathcal{L}_O = \sum_{t=2}^{T} \sum_{j=1}^{J} ||\hat{O}_{t,j} - O_{t,j}||_2$$

Here, $\hat{O}_t = \hat{X}_t - \hat{X}_{t-1}$, and $O_t = X_t - X_{t-1}$.

In conclusion, the cumulative loss is determined using the equation:

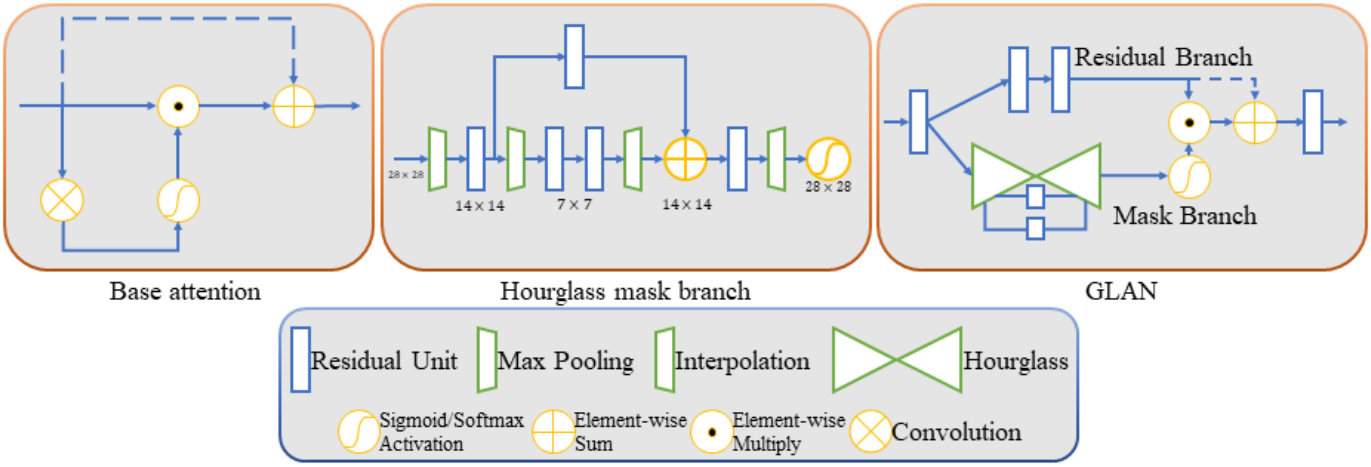

Fig. 3. A base module and a GLAN module: A base attention block(left), An expanded plot for the Hourglass mask branch in GLAN(middle), An attention block with GLAN structure, short for "GLAN block"(left).

$$\mathcal{L} = \mathcal{L}_{3D} + \lambda_O \mathcal{L}_O$$

Here, λ_O operates as a constant coefficient, ensuring a harmonious balance between the losses.

For the extraction of 2D skeletons from videos, we employ MediaPipe, and subsequently fine-tune the comprehensive network on the Human3.6M[28] training dataset[12].

4 Experiment

This section elucidates our experimental datasets, the setup, and provides both quantitative and qualitative analysis of the results.

4.1 Datasets

We utilized three distinct sign language datasets designed for ISLR, spanning American, Turkish, and Mexican sign languages. Additionally, Human3.6M was employed to train the DSTformer backbone and to appraise the impact of 3D human pose estimation.

WLASL(WLASL-100 subset)

The Word-Level American Sign Language (WLASL) dataset [46] encompasses a vast collection of isolated ASL videos, spanning 2,000 unique classes distributed over 21,083 videos from 119 distinct signers. Every video captures a signer executing a single sign, predominantly from a frontal perspective. These videos, sourced from 20 diverse platforms—including ASLU, ASL-LEX, and YouTube—exhibit a myriad of backgrounds and lighting conditions. The dataset bifurcates into three subsets: WLASL-100, WLASL-300, and WLASL-2000. Specifically, the WLASL-100 subset features 100 classes represented through 2,038 videos from 119 signers. It's subdivided into training (1,442 videos, 91 signers), validation (338 videos, 69 signers), and testing sets (258 videos, 56 signers). These videos, decoded at 25 fps(frame per second), have dimensions of 256×256 pixels. Our study employs the WLASL-100 subset for skeleton sequence estimation.

AUTSL(RGB data track)

The Ankara University Turkish Sign Language (AUTSL) dataset[47] stands as an expansive repository of isolated Turkish Sign Language (TSL) videos. It encapsulates 226 signs demonstrated by 43 unique signers, accumulating 36,302 video samples. The backdrop diversity spans 20 different settings and features a diverse pool of signers, including deaf individuals, coda, TSL educators, translators, students, and trained professionals. This dataset is segmented into training (28,142 videos), validation (4,418 videos), and testing (3,742 videos) subsets. Captured using the Microsoft Kinect v2, the dataset furnishes RGB video data alongside depth details. For our analysis, we exclusively tapped into the RGB data track to deduce skeleton sequences.

LSM Dataset

The LSM dataset[48] houses 3,000 distinct sign language samples, encapsulating 30 unique gestures from Mexican Sign Language (LSM). Every gesture has been executed 25 times by four individual signers. Using the OAK-D camera, each gesture was chronicled over 20 sequential frames.

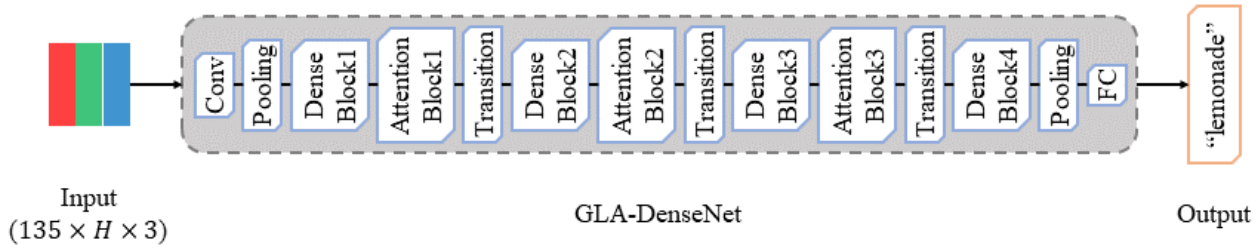

Fig. 4. Using the TSSI format, the $135 \times H \times 3$(width, height, channels) skeleton sequence is processed by the GLA-DenseNet network for classification. "H" is the average sequence length in the training set.

Within each frame, 543 keypoints—spanning the face, body, and hands—were delineated via MediaPipe. A subset of these, amounting to 67 keypoints, is distributed as: 20 for the face, 5 for the body, and 21 for each hand. These keypoints' coordinates underwent transformation into meters, utilizing the camera's focal length and depth data. To counterbalance the variances in distance between the camera and signer, these coordinates were normalized concerning the inner chest. A subtle alteration was made to the base graph showcased in Figure 1 to accommodate these 67 keypoints, excluding the nose.

Human3.6M

Human3.6M [49] is a commonly used indoor dataset for 3D human pose estimation which contains 3.6 million video frames of professional actors performing daily actions.

4.2 Experiment setup

Experiments were executed using an NVIDIA RTX 3090 GPU.

For DSTformer, geared towards 3D pose estimation, we sourced varied and realistic 3D human motion from both the Human3.6M [49] and AMASS[50] datasets. Following methodologies from previous studies [28][30], we segmented Human3.6M [49] into training and testing subsets. AMASS [50], a consolidation of numerous marker-based motion capture(Mocap) datasets, was aligned the body keypoint definitions with Human3.6M and calibrated in line with existing literature[12][51]. We incorporated random noise and missing joint simulations and employed a phased training strategy based on curriculum learning practices[52][53]. To be more specific, we randomly zero out 15% joints, and sample noises from a mixture of Gaussian and uniform distributions. We first train on 3D data only for 30 epochs, then train on both 3D data and 2D data for 60 epochs, following the curriculum learning practices.

To assess the efficacy of GLA-DenseNet, our procedure was tripartite: initial hyperparameter tuning using stratified 5-fold cross-validation, training with the optimal parameters, and final performance evaluation on the test set. The epochs varied based on the dataset being processed. While we used 120 epochs for the WLASL-100 datasets, we sued 30 epochs for the AUTSL and the LSM dataset. Optimization involved the cross-entropy loss combined with stochastic gradient descent, employing Nesterov momentum with a momentum coefficient of 0.97.

For hyperparameter tuning, we followed the procedure proposed by[7] that uses learning rate range tests to select the learning rate range and other hyperparameters such as weight decay, dropout, and batch size for a cyclical learning rate schedule. We performed a grid search of the following hyperparameter configurations: batch size $= [32, 64]$, dropout $= [0.1, 0.3, 0.5]$, weight decay $= [1e-5, 1e-6, 1e-7]$, learning rate range $= (0.001, 0.5)$. The final hyperparameters for each dataset are encapsulated in Table 1.

4.3 Quantitative results

We present the categorical top-1 accuracy obtained on test sets of WLASL-100, AUTSL, and LSM datasets, contrasting our model against both skeleton and RGB-based methods.

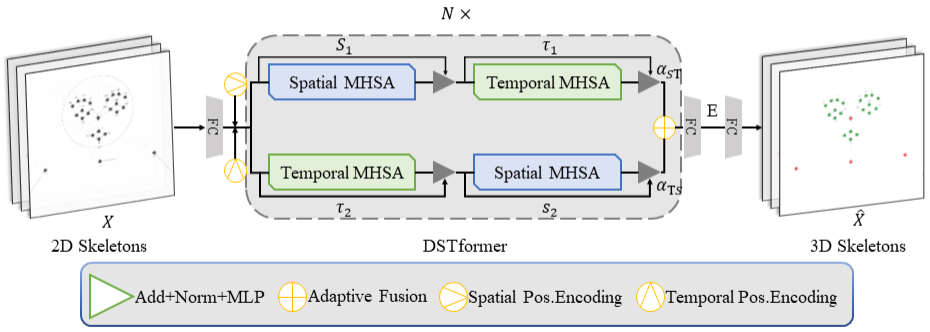

Fig. 5. 3D Pose Estimation: We use the Dual-stream Spatio-temporal Transformer (DSTformer) with N dual-stream fusion modules for 2D to 3D transitions. Each module has spatial and temporal MHSA branches, complemented by MLP. Spatial MHSA connects joints within a timestep, while Temporal MHSA captures joint movement.

TABLE 1
Hyperparameters used for each dataset. "BS": batch size, "WD": weight decay, "DO": dropout, "LRR": learning rate range.

Dataset	BS	WD	DO	LRR
WLASL-100	64	1e-5	0.4	0.001-0.007
AUTSL	64	1e-5	0.3	0.01-0.46
LSM	64	1e-5	0.5	0.01-0.1

4.3.1 WLASL-100

Table 2 underscores the superior performance of our model when juxtaposed with other skeleton-based models and its commendable standing amidst RGB-based models. In this discussion, we consider the vanguard models in the SLR domain. While certain models perceive skeletons through the lens of graphs, others harness them via transformer architectures.

Key models that our model overtakes include GCN-BERT[54], SPOTER[55], Pose-TGCN[46], and SL-TSSI-DenseNet[7]:

- **GCN-BERT** visualizes a skeleton sequence as a graph, employing a graph convolutional neural network (GCN) to grasp spatial relationships and leveraging BERT[42] to capture temporal dynamics.
- **Pose-TGCN** integrates a temporal dimension with a GCN, visualizing the comprehensive skeleton sequence as an interconnected graph.
- **SPOTER** perceives a skeleton sequence as a vector, making nuanced modifications to the original Transformer[56] by feeding the decoder with a representation of the sample's class.
- **SL-TSSI-DenseNet** adopts the TSSI approach to sequence skeleton keypoints. The utilization of the DFS search in TSSI is believed to better preserve geometric information.

In the realm of models employing RGB-based input, our GLA-DenseNet stands tall against models like I3D, which is laden with a hefty 52M parameters, while our model is elegantly designed with just 8.5M parameters. However, when matched against the Full Transformer Network[57] and TK-3D ConvNet[58], our model trails a bit. This gap can be ascribed to the discrepancies in complexity denoted by the contrasting sizes of the models. To elucidate, the Full Transformer Network brandishes 20M parameters, and the TK-3D ConvNet comes equipped with an approximate 52M parameters.

A pivotal enhancement, the integration of data augmentation, ensures our GLA-DenseNet towers over all the referenced RGB-based and skeleton-based models.

TABLE 2
A comparative study of leading-edge models on the WLASL-100 dataset, inclusive of our model.
Legends are as follows: "Input" denotes input modality. "RGB" signifies raw RGB videos used as input. "Skeleton" indicates models using some variant of skeleton data. "DA" stands for "Data Augmentation

Method	Input	Accuracy (%)
I3D(baseline)[46]	RGB	65.89
TK-3D ConvNet[58]	RGB	77.55
Full Transformer Network[57]	RGB	80.72
GCN-BERT[54]	Skeleton	60.15
Pose-TGCN[46]	Skeleton	55.43
SPOTER[55]	Skeleton	63.18
SL-TSSI-DenseNet[7]	Skeleton	73.02
GLA-DenseNet(ours)	Skeleton	74.35
GLA-DenseNet+DA(ours)	Skeleton	82.45

4.3.2 AUTSL(RGB data track)

In Table 3, we present a comparison of results achieved on the AUTSL dataset. This comparison juxtaposes our model with other methods that are either skeleton-based

TABLE 3
Comparison of the state-of-the-art in the AUTSL dataset including our model.
"Input" : input modality. "RGB": Raw RGB videos as input. "Skeleton": it uses any form of skeleton data as input.

Method	Input	Accuracy(%)
CNN+FPM+BLSTM+ Attention(baseline)[47]	RGB	49.22
I3D+RGB-MHI[63]	RGB	93.53
ResNet2+1D[59]	RGB	95.00
SlowFast+Slow+TSM[60]	RGB	96.55
SSTCN[59]	Skeleton	93.37
SL-TSSI-DenseNet[7]	Skeleton	93.13
GLA-DenseNet(ours)	Skeleton	94.56

TABLE 4
Comparison of the state-of-the-art in the LSM dataset including our model.
"Input" : input modality. "Skeleton": it uses any form of skeleton data as input.

Method	Input	Accuracy (%)
RNN(baseline)[48]	Skeleton	92.44
LSTM[48]	Skeleton	96.66
GRU[48]	Skeleton	97.11
SL-TSSI-DenseNet[7]	Skeleton	98.0
GLA-DenseNet(ours)	Skeleton	98.87

or RGB-based. Notably, our model not only outperforms other skeleton-based techniques but also maintains a more compact architecture with approximately 8.5M parameters. To put this into perspective, the Multi-stream SL-GCN [59] method employs about 19.2M parameters. This technique harnesses spatio-temporal graph convolutional modules to analyze four graph representations of a skeleton sequence, which are derived from the joints and bone vectors.

Contrastingly, the RGB-based model introduced by the wenbinwuee team for the ChaLearn LAP Large Scale Signer Independent Isolated Sign Language Recognition Challenge [60] requires a significant 33M parameters at a minimum. Their model processes RGB data using various methods, such as SlowFast [61], SlowOnly [61], and TSM [62]. Following individual processing, the model consolidates class scores to yield a final prediction.

4.3.3 LSM dataset

In the LSM dataset, the results are showcased in Table 4. The baseline methods introduced by[48], which include RNN, LSTM, and GRU, utilize skeleton data in vector form. These methods incorporate recurrent dropout and culminate with a dense layer in the network's tail end. In contrast, SL-TSSI-DenseNet[7] does not feature attention blocks and a 3D pose estimator, elements present in our model. Notably, our model outperforms these alternatives, achieving a test accuracy of 98.87%. To date, no other models have benchmarked against this dataset.

4.4 Qualitative results

We conducted a qualitative analysis based on the results from the WLASL-100 dataset. From this, we derived the confusion matrix for the testing set, which allowed us to visualize the signs and identify which ones were most commonly misclassified.

Fig. 6. Incorrectly Predicted Similar Signs: "Bird" (Upper) vs "Drink" (Lower).

The results reveal significant improvements comparing to the SL-TSSI-DenseNet[7] network. For instance, our network exhibits a more accurate distinction between signs such as "thin" and "hot". This enhancement is likely attributed to the incorporation of the attention mechanism and the GLAN block in our model, coupled with the enriched geometric details provided by the 3D human pose estimation block.

However, there remain areas for improvement. Signs such as "bird" and "drink"(as in Figure 6) continue to be sources of confusion. This could be due to the absence of an attention mechanism tailored for temporal information. Although our current attention blocks excel in highlighting geometric differences, they need to be further refined to effectively capture temporal nuances.

4.5 Ablation study

Using the WLASL-100 dataset, we dissected the contributions of 3D pose estimation and data augmentation on our model's performance. We used 3 data augmentation techniques that transform the spatial and temporal characteristics of the skeleton motion: 1)Scale, scales the skeleton by a random factor between 0.5 and 1.0 to mimic different body sizes, 2)Flip, flips horizontally the skeleton with a random probability of 0.5, and 3)Speed, resizes vertically the TSSI to a random number of frames between 48 (25th percentile) and the 74 (75th percentile) of the training set video length using bilinear resizing.

As shown in Table 5, a CNN-based model (A) accepting 2D skeleton as input achieves 40.13% accuracy. By adding only data augmentation, model (B) obtains an increase of around 18% in accuracy. By adding only 3D human pose estimation, model (C) obtains an increase of around 21% in accuracy. Finally, by adding both 3D human pose estimation and data augmentation, the model (D) obtains an increase of around 32% in accuracy.

The results show that the accuracy in WLASL-100 increases with 3D human pose estimation. Table 6 shows the results of an ablation study to determine the effects of the data augmentation techniques in the best model obtained with 3D human pose estimation and data augmentation. The results show that the speed augmentation technique is

TABLE 5
Average top-1 accuracy on 5 runs of models generated with different configurations.

Model	3D human pose estimation	Augmentation	Accuracy(%)
A	✗	✗	40.13
B	✗	✓	58.45
C	✓	✗	62.02
D	✓	✓	72.43

TABLE 6
Average top-1 accuracy on 5-fold cross-validation after removing individually the data augmentation techniques "Flip", "Speed" and "Scale". The column "None" represents the result of not removing any data augmentation technique.

DA Technique	None	Flip	Speed	Scale
Accuracy	81.47	80.63	64.72	81.27

the most important as the accuracy drops down to 64.72% when it is removed. It also shows that the flip augmentation and the scale augmentation do not have a substantial impact when they are removed as the accuracy drops by only 1%.

5 Conclusion

Our research unveils the GLA-DenseNet—a state-of-the-art convolutional neural network that synergistically melds the DenseNet architecture with pioneering elements such as attention mechanisms, global long-sequence attention (GLAN), and 3D human pose estimation. Anchored by the dual-stream spatio-temporal transformer (DSTformer) and the tree-structured skeleton image (TSSI) encoding, our approach excels in discerning and accentuating vital spatial relationships—fundamental for consistent ISLR accuracy.

After exhaustive testing on three pivotal sign language datasets—WLASL, AUTSL, and LSM—GLA-DenseNet's supremacy over conventional skeleton-based methodologies became evident. Especially noteworthy was its ability, when complemented with data augmentation, to eclipse certain RGB-based models, all the while maintaining a streamlined architecture, underscoring its remarkable efficiency.

Our results go beyond mere numerical superiority; they emphasize the transformative capability of weaving diverse computational methodologies into CNN frameworks for ISLR. This evolution heralds the dawn of a more inclusive communication landscape bridging the hearing and the deaf or hard-of-hearing communities. As we chart the course ahead, we are keen on delving deeper into enhancements, including the development of specialized attention blocks for temporal dynamics and innovative pre-training paradigms.

Acknowledgments

References

[1] H. Haualand, "Sign language interpreting: A human rights issue," *International Journal of Interpreter Education*, vol. 1, no. 1, p. 7, 2009.

[2] A. A. Hosain, P. S. Santhalingam, P. Pathak, H. Rangwala, and J. Kosecka, "Hand pose guided 3d pooling for word-level sign language recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3429–3439.

[3] R. Rastgoo, K. Kiani, and S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimedia Tools and Applications*, vol. 79, pp. 22 965–22 987, 2020.

[4] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[5] T. Soo Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 20–28.

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[7] D. Laines, M. Gonzalez-Mendoza, G. Ochoa-Ruiz, and G. Bejarano, "Isolated sign language recognition based on tree structure skeleton images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 276–284.

[8] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio–temporal visual attention on skeleton image sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2405–2415, 2018.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[10] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[11] R. Yue, Z. Tian, and S. Du, "Action recognition based on rgb and skeleton data sets: A survey," *Neurocomputing*, 2022.

[12] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: Unified pretraining for human motion analysis," *arXiv preprint arXiv:2210.06551*, 2022.

[13] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*. IEEE, 2015, pp. 579–583.

[14] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," *arXiv preprint arXiv:1804.06055*, 2018.

[15] A. Banerjee, P. K. Singh, and R. Sarkar, "Fuzzy integral-based cnn classifier fusion for 3d skeleton action recognition," *IEEE transactions on circuits and systems for video technology*, vol. 31, no. 6, pp. 2206–2216, 2020.

[16] H. Wang, B. Yu, K. Xia, J. Li, and X. Zuo, "Skeleton edge motion networks for human action recognition," *Neurocomputing*, vol. 423, pp. 1–12, 2021.

[17] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2969–2978.

[18] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[19] W. Du, Y. Wang, and Y. Qiao, "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3725–3734.

[20] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5209–5217.

[21] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[22] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[23] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines. arxiv 2019," *arXiv preprint arXiv:1906.08172*, vol. 5, 2019.

[24] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 133–10 142.

[25] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 529–545.

[26] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu, "Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2344–2353.

[27] H. Ci, M. Wu, W. Zhu, X. Ma, H. Dong, F. Zhong, and Y. Wang, "Gfpose: Learning 3d human pose prior with gradient fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4800–4810.

[28] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2640–2649.

[29] Y. Cheng, B. Yang, B. Wang, and R. T. Tan, "3d human pose estimation using spatio-temporal networks with explicit occlusion training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 631–10 638.

[30] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7753–7762.

[31] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2272–2281.

[32] H. Ci, C. Wang, X. Ma, and Y. Wang, "Optimizing network structure for 3d human pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2262–2271.

[33] J. Wang, S. Yan, Y. Xiong, and D. Lin, "Motion guided 3d pose estimation from videos," in *European Conference on Computer Vision*. Springer, 2020, pp. 764–780.

[34] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool, "Mhformer: Multi-hypothesis transformer for 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 147–13 156.

[35] W. Shan, Z. Liu, X. Zhang, S. Wang, S. Ma, and W. Gao, "P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 461–478.

[36] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 232–13 242.

[37] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 656–11 665.

[38] U. Von Agris, M. Knorr, and K.-F. Kraiss, "The significance of facial features for automatic sign language recognition," in *2008 8th IEEE international conference on automatic face & gesture recognition*. IEEE, 2008, pp. 1–6.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[40] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://keras.io

[41] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[43] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[44] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[45] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[46] D. Li, C. Rodriguez, X. Yu, and H. Li, "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*, 2020, pp. 1459–1469.

[47] O. M. Sincan and H. Y. Keles, "Autsl: A large scale multi-modal turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181 340–181 355, 2020.

[48] K. Mejía-Peréz, D.-M. Córdova-Esparza, J. Terven, A.-M. Herrera-Navarro, T. García-Ramírez, and A. Ramírez-Pedraza, "Automatic recognition of mexican sign language using a depth camera and recurrent neural networks," *Applied Sciences*, vol. 12, no. 11, p. 5523, 2022.

[49] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[50] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[51] H. Ci, X. Ma, C. Wang, and Y. Wang, "Locally connected network for monocular 3d human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1429–1442, 2020.

[52] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[53] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4555–4576, 2021.

[54] A. Tunga, S. V. Nuthalapati, and J. Wachs, "Pose-based sign language recognition using gcn and bert," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 31–40.

[55] M. Boháček and M. Hrúz, "Sign pose-based transformer for word-level sign language recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 182–191.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[57] Y. Du, P. Xie, M. Wang, X. Hu, Z. Zhao, and J. Liu, "Full transformer network with masking future for word-level sign language recognition," *Neurocomputing*, vol. 500, pp. 115–123, 2022.

[58] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, "Transferring cross-domain knowledge for video sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6205–6214.

[59] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3413–3423.

[60] O. M. Sincan, J. Junior, C. Jacques, S. Escalera, and H. Y. Keles, "Chalearn lap large scale signer independent isolated sign language recognition challenge: Design, results and future research," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3472–3481.

[61] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[62] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7083–7093.

[63] O. M. Sincan and H. Y. Keles, "Using motion history images with 3d convolutional networks in isolated sign language recognition," *IEEE Access*, vol. 10, pp. 18 608–18 618, 2022.