Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

LSMA-Net: Local Spatiotemporal Multi-Head Attention Network for Skeleton-Based Continuous Sign Language Recognition

Yu Du
Beijing Union University
Wentao Liu
Beijing Union University
Dongyu Hu
Beijing Union University
Lei Fan
Beijing Union University
Ying Zheng
jqrzhengyin@buu.edu.cn
Beijing Union University

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11759752/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

LSMA-Net: Local Spatiotemporal Multi-Head Attention Network for Skeleton-Based Continuous Sign Language Recognition

Yu Du[1], Wentao Liu[1], Dongyu Hu[1], Lei Fan[1], Ying Zheng[1*]

[1]College of Robotics, Beijing Union University, Beijing, China.

*Corresponding author(s). E-mail(s): jqrzhengyin@buu.edu.cn;

Abstract

Continuous sign language recognition (CSLR) aims to predict an ordered sequence of glosses from unsegmented sign-language videos. Although skeleton-based approaches provide compact, structured motion representations, spatial feature aggregation and long-range temporal modeling may weaken local discriminative cues such as hand configuration, location, and motion trajectory. To coordinate local fine-grained information with global temporal context, we present LSMA-Net, a local spatiotemporal multi-head attention network for skeleton-based CSLR. A global–local dual-branch architecture combines a global context branch that captures long-range inter-gloss dependencies with a local visual feature reorganization branch that retains multi-stream features before global pooling over keypoints. The latter employs overlapping temporal windows, local spatiotemporal self-attention, and cross-stream reorganization to capture short-term motion details. Motion–confidence soft reweighting and progressive spatiotemporal DropBlock support feature learning, while multi-level CTC supervision and bidirectional consistency distillation coordinate the prediction branches. On PHOENIX-2014T and CSL-Daily, LSMA-Net achieves test word error rates of 19.21% and 25.90%, respectively. In extended sign language translation experiments, it obtains ROUGE-L/BLEU-4 scores of 54.77/29.64 and 55.41/26.69 on the two datasets, respectively. These results demonstrate competitive skeleton-based recognition performance and the applicability of the learned representations to downstream translation. The source code is publicly available at https://github.com/gysmhyt/LSMA-Net.

Keywords: Continuous sign language recognition, Skeleton keypoints, Global–local complementary representation, Local spatiotemporal attention, Collaborative learning

1 Introduction

Continuous sign language recognition (CSLR) converts unsegmented signing videos into ordered gloss sequences, supporting automatic sign-language understanding and accessible communication [1]. Unlike ordinary action recognition, CSLR must capture short-term handshape, hand location, motion trajectory, and facial non-manual cues together with inter-gloss temporal dependencies, despite the absence of explicit gloss boundaries. Coordinating local fine-grained actions and global temporal context is therefore an important recognition challenge.

Existing CSLR approaches use RGB video, combined RGB and skeleton information, or skeleton keypoints alone. RGB methods extract appearance features with convolutional networks and model temporal relationships using temporal

convolutions, recurrent networks, or Transformers [2–8]. RGB input provides rich texture and expression information but can be sensitive to background, lighting, clothing, and viewpoint changes [9]. TwoStream-SLR and related approaches combine RGB and skeleton representations [3, 10], at the cost of additional feature extraction and fusion complexity. Skeleton-based methods instead use structured keypoint coordinates and confidence values, reducing appearance interference and visual input size [11, 12]. CoSign models spatiotemporal relationships through group-specific graph convolutions and static–dynamic streams [13]; MSKA combines multi-stream keypoint attention with self-distillation [14]; and HyperSign employs hierarchical hypergraphs to represent higher-order joint relations [15]. These studies advance skeleton-based CSLR through topology, cross-region interaction, and hierarchical modeling.

Despite this progress, balancing local discriminative cues and long-range context remains difficult. Gloss discrimination depends on short-term handshape, position, movement, and facial cues, not merely aggregate motion. Spatial aggregation followed by long-range temporal encoding provides a compact sequence representation but may weaken keypoint-level details [16–19]. Unequal keypoint quality, occlusion, and inaccurate coordinates can further affect feature propagation [20, 21]. Moreover, sentence-level gloss annotations lack explicit frame boundaries, limiting direct temporal supervision of local features and semantic streams [22, 23]. Our central focus is thus to combine keypoint-level local discrimination with global temporal modeling, supported by quality adjustment and hierarchical supervision.

We propose LSMA-Net (Local Spatiotemporal Multi-Head Attention Network), a skeleton-based CSLR network with complementary global and local branches. The global branch models long-range context, while the local branch preserves unpooled multi-stream keypoint features and extracts fine-grained actions within temporal windows. Their fusion creates complementary representations, supported by progressive regularization and hierarchical supervision. Our contributions are as follows:

1. We develop a global–local complementary representation framework. The global branch captures long-range temporal dependencies; the local branch preserves keypoint-level features and combines overlapping windows, local spatiotemporal multi-head self-attention, cross-stream reorganization, and spatial–temporal–channel gating to enhance fine-grained motion representations.
2. We integrate progressive spatiotemporal regularization and hierarchical supervision. Motion–confidence soft reweighting and spatial/temporal DropBlock adjust feature contributions and structured perturbations; multi-level CTC objectives across four semantic streams and the main and fusion branches, together with bidirectional KL consistency distillation, support joint optimization.
3. We evaluate LSMA-Net on PHOENIX-2014T and CSL-Daily, obtaining test WERs of 19.21% and 25.90%, respectively. Sequential ablations identify the largest gain from local feature reorganization, with additional gains from regularization and supervision. The corresponding SLT ROUGE-L/BLEU-4 scores are 54.77/29.64 and 55.41/26.69, demonstrating applicability to downstream translation.

2 Related Work

2.1 Continuous and Isolated Sign Language Recognition

Sign language recognition includes isolated sign language recognition (ISLR) and CSLR [3, 24, 25]. As shown in Fig. 1, ISLR uses pre-segmented videos containing a single gloss and emphasizes local spatiotemporal discrimination. Hosain et al. [25] introduced hand-pose-guided 3D pooling to combine multilevel hand motion features. Since ISLR assumes known sign boundaries, it cannot directly process unsegmented signing.

CSLR predicts ordered gloss sequences from continuous video and typically has sentence-level labels without explicit frame boundaries. Alignment constraints, cross-module consistency, and pseudo-labels have been used to support implicit alignment and temporal feature learning [4–6, 26–29]. In contrast, sign language translation (SLT) generates natural-language text, mapping visual semantics to language. Joint recognition–translation frameworks combine gloss supervision and sequence generation [8].

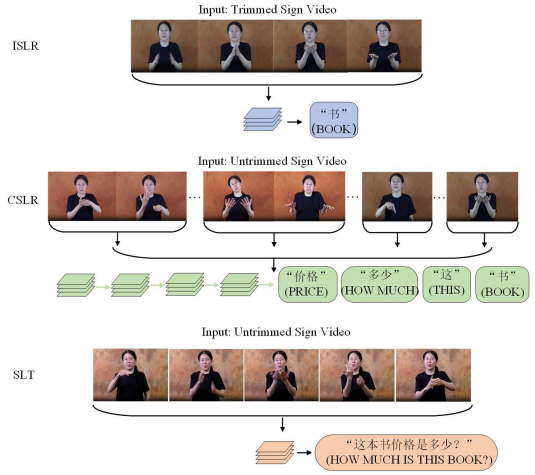


Fig. 1 Comparison of task paradigms for ISLR, CSLR, and SLT.

2.2 Skeleton-Based CSLR Methods

Skeleton sequences compactly represent handshape, location, motion, facial activity, and body pose, with explicit topological priors and reduced appearance interference. Related methods include graph-based temporal modeling, body-part-aware multi-stream fusion, and self-supervised pretraining [30]. Graph methods use spatial-temporal convolutions and adaptive topologies [31–33], while multi-stream modeling distinguishes body regions and masked pretraining learns motion priors from unlabeled sequences. Wang et al. [34] introduced multi-part skeleton learning; CoSign uses group-specific graph convolutions and static–dynamic streams [13]; and HyperSign models higher-order joint relations through hierarchical hypergraphs [15]. SignBERT+ learns hand-motion priors by multilevel masked pretraining [35], and MSKA combines multi-stream attention with self-distillation [14]. Further comparative analysis of skeleton representations and training is provided in [36]. Building on these approaches, LSMA-Net focuses on complementarity between unpooled keypoint-level local features and spatially aggregated global temporal representations, using dual branches and feature fusion.

2.3 Noise-Robust Modeling and Weakly Supervised Sequence Learning

Rapid motion, occlusion, and pose changes may cause missing or inaccurate keypoints. SignBERT+ applies joint-, frame-, and clip-level masking for self-supervised representation learning [35]; DropBlock masks contiguous regions to reduce reliance on local feature activations [37]. Our model applies motion–confidence soft reweighting and progressive spatiotemporal DropBlock to support dual-branch feature learning [38].

CTC learns sequence alignments without frame-level boundaries [29]. VAC combines auxiliary supervision and alignment constraints [26]; C2SLR introduces keypoint-guided spatial-attention and sentence-level consistency [27]; and SMKD uses self-mutual distillation between visual and contextual modules [39]. These works support temporal alignment and cross-branch consistency [40, 41]. LSMA-Net employs multi-level CTC across four streams and the global and fusion branches, complemented by bidirectional KL consistency.

3 Method

3.1 Overall Architecture of LSMA-Net

LSMA-Net addresses the weakening of local discriminative information, keypoint noise, and weak supervision in long-range skeleton modeling. As shown in Fig. 2, its multi-stream skeleton encoder feeds a global context branch and a local visual feature reorganization branch. Progressive spatiotemporal regularization and hierarchical supervision coordinate local and global feature learning.

HRNet extracts keypoints from the video, and 79 are selected: 42 hand, 11 upper-body, and 26 facial keypoints. Each contains normalized two-dimensional coordinates and detection confidence. The input is divided into left-hand, right-hand, facial, and whole-body streams. Each hand stream includes 21 hand joints and six reference body joints; the face and body streams contain 26 and 79 keypoints.

In multi-stream encoding, motion–confidence soft reweighting adjusts the keypoint features, yielding $\widetilde{X}^s$. Learnable global and input-dependent dynamic relations form adaptive adjacency matrices:

$$A_h^s = A_{0,h}^s + \gamma_h^s A_{d,h}^s, \tag{1}$$

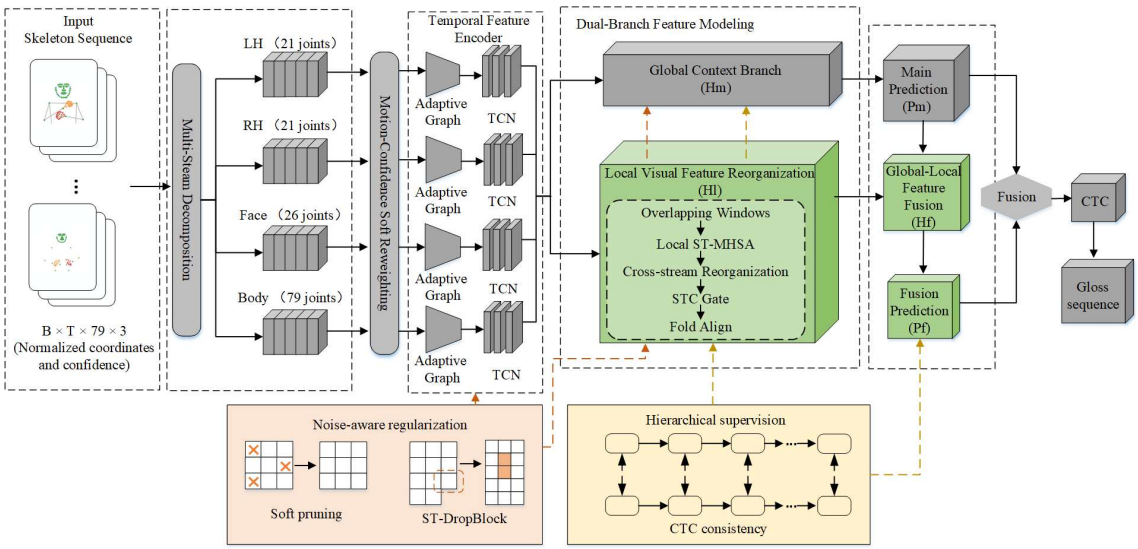


Fig. 2 Overall architecture of LSMA-Net.

where $s \in \mathcal{S} = \{lh, rh, fa, bo\}$ denotes the four semantic streams, h denotes the relation subspace, $A^s_{0,h}$ and $A^s_{d,h}$ denote global and dynamic relation matrices, and γ^s_h is a learnable combination coefficient. Spatial aggregation and temporal convolutions produce

$$F^s = E_s(\widetilde{X}^s, A^s) \in \mathbb{R}^{B\times D\times T'\times V_s}, \quad s \in \mathcal{S}, \quad (2)$$

where E_s is the stream encoder, B denotes batch size, D feature channels, T' encoded time steps, and V_s keypoints in stream s. Progressive spatial and temporal DropBlock yield regularized features $\widehat{F}^s$.

The four regularized streams enter the global and local branches, producing H_m and H_l. Feature fusion yields H_f, and the corresponding heads generate main and fusion posteriors P_m and P_f, respectively (Sec. 3.2). Training uses multi-level CTC and bidirectional KL consistency (Sec. 3.4). At inference, the main and fusion posteriors are averaged:

$$P = \tfrac{1}{2}(P_m + P_f), \quad (3)$$

then decoded by CTC. Motion–confidence reweighting uses the final training threshold, while random DropBlock and auxiliary training branches are disabled.

3.2 Local Visual Feature Reorganization

The two branches differ in the stage of spatial aggregation and the temporal range of their representations. The global branch captures long-range context from pooled regional features, whereas the local branch models interactions among unpooled keypoint features inside short temporal windows. The four regularized semantic streams $\{\widehat{F}^s\}_{s\in\mathcal{S}}$ are fed to both branches.

Global context branch. Average pooling over the keypoints in each stream produces a regional temporal representation,

$$\overline{F}^s = \mathrm{Pool}_{V_s}(\widehat{F}^s). \quad (4)$$

The four streams are concatenated along the channel axis and mapped into a fused representation:

$$F_m = \varphi_m\Big(\mathrm{Concat}_{s\in\mathcal{S}}(\overline{F}^s)\Big), \quad (5)$$

which is processed by the global temporal context encoder:

$$H_m = E_m(F_m). \quad (6)$$

Here φ_m is the stream-fusion projection and E_m is the global context encoder.

Local visual feature reorganization branch. The local branch retains keypoint-level features and partitions each stream into windows of length

w with regular stride δ. For $T' \geq w$, let K_0 denote the number of regular windows and $\mathcal{A}$ the ordered set of window starting indices:

$$K_0 = \left\lfloor \frac{T' - w}{\delta} \right\rfloor + 1,$$
$$\mathcal{A} = \{k\delta \mid 0 \leq k < K_0\} \cup \{T' - w\}, \quad K = |\mathcal{A}|. \tag{7}$$

The terminal start $T' - w$ is included only once; thus the last window covers the sequence end without tail padding. Write $a_0 < \cdots < a_{K-1}$ for the sorted elements of $\mathcal{A}$. The k-th window of semantic stream s is

$$W_k^s = \widehat{F}^s[:, :, a_k : a_k + w, :]. \tag{8}$$

When $0 < \delta < w$, adjacent windows overlap. Each window's temporal and keypoint dimensions are jointly unfolded into spatiotemporal nodes:

$$U_k^s = \mathrm{Reshape}(W_k^s) \in \mathbb{R}^{B \times (wV_s) \times D}. \tag{9}$$

Local spatiotemporal multi-head self-attention captures relationships among temporal positions and keypoints:

$$Z_k^s = \mathrm{Concat}_{h=1}^{H} \left[\mathrm{Softmax}\left(\frac{Q_{k,h}^s (K_{k,h}^s)^\top}{\sqrt{d_h}} \right) V_{k,h}^s \right] W^O. \tag{10}$$

Here H is the number of attention heads; $Q_{k,h}^s$, $K_{k,h}^s$, and $V_{k,h}^s$ are the query, key, and value matrices projected from local features, d_h is the key dimension, and W^O is the output projection. This operation jointly models keypoint relationships and short-term motion variations.

After local attention, features from the semantic streams are aligned according to their temporal positions, concatenated, and projected to form reorganized local features:

$$R_k = \mathcal{R}\left(\{Z_k^s\}_{s \in \mathcal{S}}\right), \quad k = 0, \ldots, K-1. \tag{11}$$

Here $\mathcal{R}$ is cross-stream local reorganization. An STC (spatial–temporal–channel) joint gating operation then reweights contributions along spatial, temporal, and channel dimensions, yielding enhanced windows $\widehat{R}_k$.

The enhanced windows are returned to their starts a_k. At each temporal position, overlapping contributions are averaged to recover a sequence of length T':

$$H_l = \mathrm{FoldAvg}\Big(\{(\widehat{R}_k, a_k)\}_{k=0}^{K-1}; T'\Big). \tag{12}$$

Here FoldAvg denotes position-aligned overlap averaging, including the terminal window when its start differs from the regular grid.

Global–local feature fusion. The two representations are first projected to the same temporal resolution and feature dimension:

$$\widetilde{H}_m = \varphi_g(H_m), \qquad \widetilde{H}_l = \varphi_l(H_l). \tag{13}$$

They are concatenated and mapped to a fused feature:

$$H_f = \varphi_f\Big(\mathrm{Concat}[\widetilde{H}_m; \widetilde{H}_l]\Big), \tag{14}$$

where φ_g, φ_l, and φ_f are learned projections. Separate prediction heads produce main-branch and fusion-branch gloss posteriors:

$$P_m = \mathrm{Softmax}(W_m H_m + b_m), \tag{15}$$
$$P_f = \mathrm{Softmax}(W_f H_f + b_f). \tag{16}$$

These posteriors, together with four stream-specific auxiliary predictions, are used for hierarchical supervision.

3.3 Progressive Spatiotemporal Regularization

Skeleton detection under occlusion, rapid motion, and complex poses may lead to low-confidence joints, shifted coordinates, or missing observations. The regularization module in Fig. 3 applies motion–confidence soft reweighting before spatiotemporal encoding and structured spatial/temporal DropBlock after encoding.

For stream s, let $X^s \in \mathbb{R}^{B \times T \times V_s \times 3}$ denote the input skeleton sequence, where each keypoint contains normalized 2D coordinates and a detection-confidence score. Motion saliency $m_{t,v}^s$ is calculated from the displacement of keypoint v between adjacent frames and normalized. It is combined with detection confidence $c_{t,v}^s$ into a joint score:

$$r_{t,v}^s = \alpha \widehat{m}_{t,v}^s + \beta c_{t,v}^s, \quad \alpha + \beta = 1, \quad \alpha, \beta \geq 0, \tag{17}$$

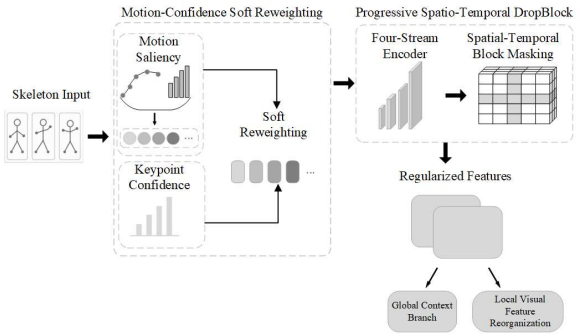


Fig. 3 Motion–confidence-guided progressive spatiotemporal regularization module.

where α and β are the corresponding weights. A smoothly increasing threshold avoids abruptly removing low-scoring joints:

$$\tau(e) = \tau_0 + \frac{(\tau_E - \tau_0)e}{E}. \tag{18}$$

Here e and E denote the current and total training epochs, and τ_0 and τ_E are the initial and final thresholds, with $\tau_E > \tau_0$. A continuous mask is calculated by

$$M^s_{t,v} = \sigma\left(\frac{r^s_{t,v} - \tau(e)}{\gamma}\right), \tag{19}$$

where $\gamma > 0$ is a temperature factor and σ is the sigmoid function. The input is softly reweighted:

$$\widetilde{X}^s_{t,v} = M^s_{t,v} \odot X^s_{t,v}, \tag{20}$$

with broadcast multiplication across the keypoint feature channels.

After encoding, spatial DropBlock masks connected joints or local body regions and temporal DropBlock masks contiguous temporal intervals to model partial observation. The masks are applied to the encoded features:

$$\widehat{F}^s = F^s \odot D^s_{sp} \odot D^s_{tm}, \tag{21}$$

where D^s_{sp} and D^s_{tm} denote structured spatial and temporal masks, respectively. To avoid overly strong regularization at the beginning of training, a progressive keep-probability schedule is used:

$$p(e) = p_0 - \frac{(p_0 - p_E)e}{E}, \qquad p_0 > p_E. \tag{22}$$

A relatively high initial keep probability allows the network to learn fundamental spatiotemporal relations; progressively stronger perturbations subsequently discourage reliance on fixed keypoints or intervals. These training operations support the common features supplied to the two recognition branches.

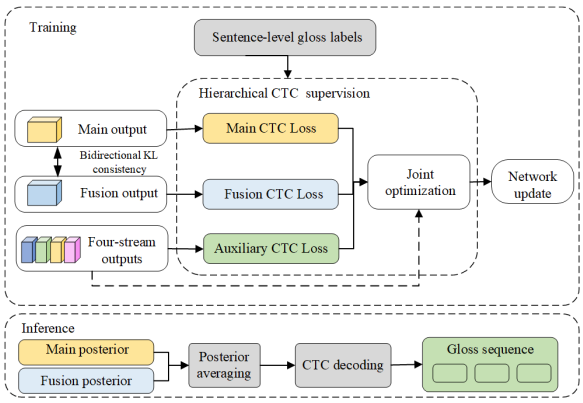


Fig. 4 Hierarchical CTC supervision and bidirectional branch-consistency learning.

3.4 Hierarchical Joint Supervision

CSLR normally uses sentence-level gloss supervision without explicit frame boundaries. LSMA-Net combines multi-level CTC objectives with bidirectional branch-consistency distillation, as illustrated in Fig. 4. For gloss distribution P and target sequence Y, the standard CTC loss is

$$\mathcal{L}_{\mathrm{CTC}}(P, Y) = -\log \sum_{\pi \in \mathcal{B}^{-1}(Y)} \prod_{t=1}^{T'} P_t(\pi_t), \tag{23}$$

where π is a framewise path that may contain blank labels, $\mathcal{B}^{-1}(Y)$ is the set of paths collapsed to Y, and T' is the length of the predicted sequence.

Both the main global branch and the global–local fusion branch are supervised by CTC:

$$\mathcal{L}_m = \mathcal{L}_{\mathrm{CTC}}(P_m, Y), \quad \mathcal{L}_f = \mathcal{L}_{\mathrm{CTC}}(P_f, Y). \tag{24}$$

An auxiliary CTC loss is applied to the four semantic streams:

$$\mathcal{L}_{\mathrm{aux}} = \sum_{s \in \mathcal{S}} \omega_s \mathcal{L}_{\mathrm{CTC}}(P_s, Y), \tag{25}$$

where P_s is the posterior for stream s, and ω_s is its loss weight. Inspired by deep mutual learning [42] and self-mutual distillation [39], two symmetric

KL terms encourage consistency without fixing one branch as a permanent teacher:

$$\mathcal{L}_{\mathrm{KD}} = D_{\mathrm{KL}}(P_m \| P_f) + D_{\mathrm{KL}}(P_f \| P_m). \quad (26)$$

The combined CSLR training objective is

$$\mathcal{L}_{\mathrm{SLR}} = \mathcal{L}_m + \lambda_f \mathcal{L}_f + \lambda_{\mathrm{aux}} \mathcal{L}_{\mathrm{aux}} + \lambda_{\mathrm{KD}} \mathcal{L}_{\mathrm{KD}}, \quad (27)$$

where the λ coefficients weight the fusion CTC, auxiliary CTC, and consistency losses, respectively. Together, these objectives coordinate features from different semantic streams and prediction levels under sentence-level supervision.

4 Experiments and Analysis

4.1 Datasets and Evaluation Metrics

We evaluate LSMA-Net on two public CSLR/SLT datasets covering distinct sign languages and application settings: PHOENIX-2014T [43] and CSL-Daily [44]. Both provide videos, gloss sequences, and natural-language text. PHOENIX-2014T consists of German weather broadcast recordings, with 7,096 training, 519 validation, and 642 test samples, and gloss/text vocabularies of 1,066 and 2,887 tokens. CSL-Daily contains 20,654 samples from daily Chinese communication, split into 18,401 training, 1,077 validation, and 1,176 test samples; its gloss and text vocabularies contain 2,000 and 2,343 entries. Topics include domestic life, social interactions, and healthcare.

CSLR is evaluated with word error rate (WER; lower is better). As an extension, SLT evaluates whether the learned visual representation can support downstream sentence generation, using BLEU [45] and ROUGE-L [46]. BLEU measures n-gram agreement with reference texts; ROUGE-L is based on the longest common subsequence. Higher translation scores indicate greater agreement with the reference sentences.

4.2 Continuous Sign Language Recognition

Experiments were conducted with one NVIDIA RTX 5090 GPU. HRNet [48], pretrained on COCO-WholeBody [47], extracts 133 whole-body keypoints per frame; 79 keypoints are selected and their 2D coordinates normalized. The Adam optimizer [49] is configured with batch size eight, 100 training epochs, $\beta_1 = 0.9$, $\beta_2 = 0.998$, and weight decay 10^{-3}. Initial learning rates are 10^{-3} for PHOENIX-2014T and 7×10^{-4} for CSL-Daily. Both use cosine annealing [50], with $T_{\max} = 100$ and 140, respectively. CTC beam-search widths at test time are five and ten, respectively.

Figure 5 shows validation WER curves for LSMA-Net and the reproduced MSKA baseline. Both models converge as training progresses. On PHOENIX-2014T, LSMA-Net obtains validation/test WERs of 19.27%/19.21%, respectively, compared with 21.19%/21.08% for the reproduced baseline, corresponding to decreases of 1.92 and 1.87 percentage points. On CSL-Daily, the corresponding results are 26.67%/25.90% versus 32.34%/31.50%, decreases of 5.67 and 5.60 percentage points. The improvement is greater on CSL-Daily. Table 1 compares the recognition performance with existing studies. The *reported* MSKA results are taken from its publication, whereas the *reproduced* results were obtained under the experimental setting considered here.

As shown in Table 1, LSMA-Net reaches test WERs of 19.21% and 25.90% on PHOENIX-2014T and CSL-Daily, respectively, using skeleton keypoints only. Relative to the published MSKA values, the differences are 0.59 and 1.20 percentage points; relative to the reproduced baseline, they are 1.87 and 5.60 points. LSMA-Net is close to HyperSign on PHOENIX-2014T and achieves competitive accuracy on CSL-Daily, although some existing methods obtain slightly better scores. Unlike RGB or RGB–keypoint fusion architectures, LSMA-Net does not require an additional RGB feature-encoding branch. The incremental effects of its three components are examined in Section 4.4.

4.3 Sign Language Translation

To evaluate the downstream applicability of the learned representation, LSMA-Net is extended to SLT on PHOENIX-2014T and CSL-Daily. The translation system loads pretrained recognition weights for each dataset and employs a pretrained mBART language-generation model connected through a visual–language mapper. Adam and a cosine annealing schedule are used during translation training, while beam search with width five

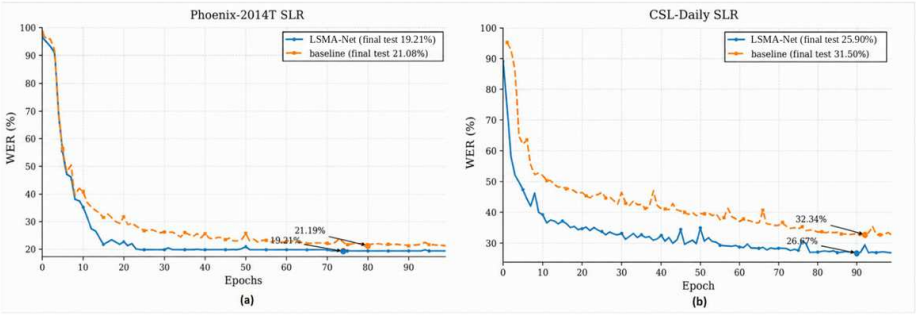


Fig. 5 Validation WER curves of LSMA-Net and the reproduced baseline on PHOENIX-2014T and CSL-Daily.

Table 1 Recognition results on PHOENIX-2014T and CSL-Daily (WER, %; lower is better). WER values are rounded to one decimal place for display.

Method	Modality	PHOENIX Dev	PHOENIX Test	CSL-Daily Dev	CSL-Daily Test
STMC [4]	RGB	19.6	21.0	–	–
CTCA [5]	RGB	19.3	20.3	31.3	29.4
CorrNet [6]	RGB	18.9	20.5	30.6	30.1
AdaBrowse+ [7]	RGB	19.5	20.6	31.2	30.7
SignBT [44]	RGB	22.7	23.9	33.2	33.2
MMTLB [10]	RGB	21.9	22.5	–	–
TLP [51]	RGB	19.4	21.2	–	–
SEN [52]	RGB	19.3	20.7	31.1	30.7
SlowFastSign [53]	RGB	17.7	18.7	25.5	24.9
TwoStream-SLR [3]	RGB+keypoint	17.7	19.3	25.4	25.3
SignBERT+ (+R) [35]	Keypoint	18.8	19.9	–	–
HyperSign [15]	Keypoint	18.6	19.2	26.1	25.7
CoSign-1s [13]	Keypoint	20.4	20.6	29.5	29.1
CoSign-2s [13]	Keypoint	19.5	20.1	28.1	27.2
MSKA (reported) [14]	Keypoint	19.6	19.8	27.5	27.1
MSKA (reproduced)	Keypoint	21.2	21.1	32.3	31.5
LSMA-Net (ours)	**Keypoint**	**19.3**	**19.2**	**26.7**	**25.9**

produces text at test time. Translation is assessed using BLEU-1 through BLEU-4 and ROUGE-L.

Figure 6 presents the SLT metric curves for both datasets. BLEU-1 and ROUGE-L generally stabilize toward later stages of training, and higher-order BLEU scores also show convergence. Higher-order BLEU scores on PHOENIX-2014T exhibit modest fluctuations, consistent with their sensitivity to local word substitutions and word order. The final translation comparisons are reported in Tables 2 and 3.

For PHOENIX-2014T, LSMA-Net achieves ROUGE-L/BLEU-4 scores of 54.77/29.64, improving over MSKA by 1.23/0.61 percentage points; all the metrics in Table 3 are highest among the listed methods. On CSL-Daily, the corresponding scores are 55.41/26.69, improvements of 1.37/1.17 points over MSKA, and remain competitive although some methods obtain higher scores. These results show that LSMA-Net representations can be used for downstream translation, supporting global–local feature learning as an input to sequence generation. The relatively lower CSL-Daily scores on some metrics may reflect the lack of hand texture, lip detail, or scene context in skeleton-only inputs.

4.4 Ablation Study

We conduct sequential incremental ablations on CSL-Daily by adding local visual feature reorganization, progressive spatiotemporal regularization, and hierarchical joint supervision in the order shown in Table 4. These results describe changes under this specific addition order.

As shown in Table 4, local visual feature reorganization reduces WER from 31.50% to 26.80%,

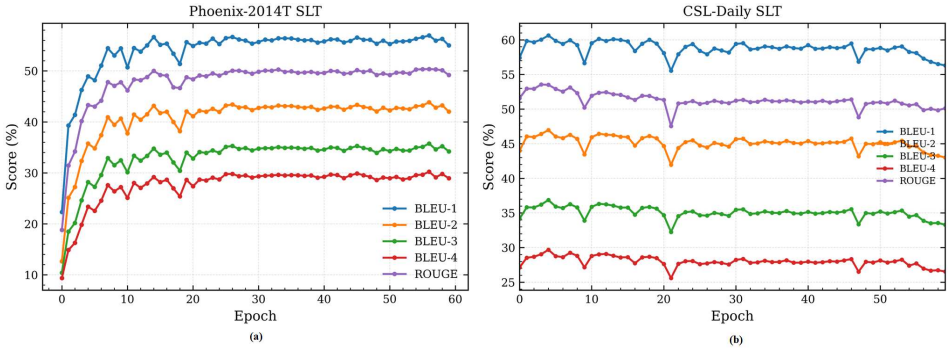


Fig. 6 SLT metric curves of LSMA-Net on PHOENIX-2014T and CSL-Daily.

Table 2 Translation results on CSL-Daily (%; higher is better).

Method	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
TwoStream-SLT (RGB) [3]	55.72	55.44	42.59	32.87	25.79
TwoStream-SLT (Pose) [3]	55.07	55.34	42.36	32.58	25.42
MMTLB [10]	53.25	53.31	40.41	30.87	23.92
GFSLT-VLP [54]	36.44	39.37	24.93	16.26	11.00
GASLT [55]	20.35	19.90	9.94	5.98	4.07
SLTUNET [56]	54.08	54.98	41.44	31.84	25.01
CV-SLT [57]	**57.06**	**58.29**	**45.15**	**35.77**	**28.94**
HyperSign [15]	56.14	–	–	–	26.87
XmDA [58]	49.34	–	–	–	21.58
IP-SLT [59]	44.09	44.85	31.50	22.66	16.72
MSKA [14]	54.04	56.37	42.80	32.78	25.52
LSMA-Net (ours)	55.41	56.57	43.39	33.59	26.69

Table 3 Translation results on PHOENIX-2014T (%; higher is better).

Method	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
TwoStream-SLT (RGB) [3]	53.48	54.90	42.43	34.46	28.95
TwoStream-SLT [3]	53.19	54.22	41.72	33.82	28.42
STMC-T [4]	46.65	46.98	36.09	28.70	23.65
MMTLB [10]	52.65	53.97	41.75	33.84	28.39
SignBERT+ (RGB) [35]	50.63	52.01	39.19	31.06	25.70
GFSLT-VLP [54]	42.49	43.71	33.18	26.11	21.44
SLTUNET [56]	52.11	52.92	41.76	33.99	28.47
CV-SLT [57]	54.33	54.88	42.68	34.79	29.27
GASLT [55]	39.86	39.07	26.74	21.86	15.74
HyperSign [15]	54.26	–	–	–	29.35
XmDA [58]	49.87	–	–	–	25.36
IP-SLT [59]	53.72	54.25	41.51	33.45	27.97
MSKA [14]	53.54	54.79	42.42	34.49	29.03
LSMA-Net (ours)	**54.77**	**55.58**	**43.11**	**35.15**	**29.64**

a decrease of 4.70 percentage points, indicating stronger discriminative representations with local spatiotemporal self-attention. Adding progressive spatiotemporal regularization further reduces WER to 26.30% (0.50 points), suggesting that motion–confidence reweighting and structured spatiotemporal perturbations help mitigate noisy or incomplete keypoint information. Hierarchical joint supervision then reduces WER to 25.90% (0.40 points), indicating benefits from multi-level CTC and bidirectional consistency distillation. Overall, LSMA-Net lowers WER by 5.60 points relative to the baseline. All three components provide incremental improvements in the specified order, supporting the effectiveness of the full framework for CSLR.

Table 4 Ablation on CSL-Daily (test WER, %). LR: local visual feature reorganization; PR: progressive spatiotemporal regularization; HS: hierarchical joint supervision.

Method	LR	PR	HS	WER
Model A (baseline)	×	×	×	31.50
Model B	✓	×	×	26.8
Model C	✓	✓	×	26.3
Model D (LSMA-Net)	✓	✓	✓	**25.9**

5 Conclusion

We introduced LSMA-Net to coordinate fine-grained local discriminative cues with global temporal context in skeleton-based CSLR. Its global branch models long-range dependencies, while the local branch retains keypoint-level features and employs overlapping windows, local spatiotemporal self-attention, and cross-stream reorganization to capture short-term motion details. Progressive spatiotemporal regularization and hierarchical joint supervision support multi-stream feature learning and branch cooperation. LSMA-Net achieves test WERs of 19.21% and 25.90% on PHOENIX-2014T and CSL-Daily, respectively. Sequential ablations show the largest incremental improvement from local feature reorganization, with additional gains from regularization and joint supervision. The extended SLT experiments further demonstrate the applicability of these representations to downstream translation. The results support global–local complementary learning as an effective approach to skeleton-based continuous sign language recognition.

Statements and Declarations

Funding. This work was supported by the Academic Research Projects of Beijing Union University (No. ZK30202406) and the R&D Program of Beijing Municipal Education Commission (No. KM202411417007).

Data availability. The experiments use the public PHOENIX-2014T and CSL-Daily datasets, subject to their respective access conditions.

Ethics approval. The study uses existing datasets and required no additional institutional ethics approval, as confirmed by the authors' institution.

Consent for publication. Permissions for the human images reproduced in the submitted figures have been verified by the authors.

Competing interests. The authors declare that they have no competing interests.

Author contributions. Yu Du: Conceptualization, Methodology, Supervision, Funding acquisition, Project administration, Writing–review & editing. Wentao Liu: Investigation, Data curation, Software, Formal analysis, Visualization, Writing–original draft. Dongyu Hu: Resources, Writing–review & editing. Lei Fan: Resources, Writing–review & editing. Ying Zheng: Resources, Writing–review & editing. All authors read and approved the final manuscript.

References

[1] S. Alyami, H. Luqman, M.A.A. Hammoudeh, Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects, Information Processing & Management 61 (5) (2024) 103774. https://doi.org/10.1016/j.ipm.2024.103774.

[2] R. Cui, H. Liu, C. Zhang, Recurrent convolutional neural networks for continuous sign language recognition by staged optimization, in: Proceedings of CVPR, 2017, pp. 7361-7369.

[3] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, B. Mak, Two-stream network for sign language recognition and translation, Adv. Neural Inf. Process. Syst. 35 (NeurIPS 2022) (2022) 17043–17056. https://doi.org/10.52202/068431-1240.

[4] H. Zhou, W. Zhou, Y. Zhou, H. Li, Spatial-temporal multi-cue network for sign language recognition and translation, IEEE Trans. Multimedia 24 (2022) 768-779. https://doi.org/10.1109/TMM.2021.3059098.

[5] L. Guo, W. Xue, Q. Guo, B. Liu, K. Zhang, T. Yuan, S. Chen, Distilling cross-temporal contexts for continuous sign language recognition, in: Proceedings of CVPR, 2023, pp. 10771-10780. https://doi.org/10.1109/CVPR52729.2023.01037.

[6] L. Hu, L. Gao, Z. Liu, W. Feng, Continuous sign language recognition with correlation network, in: Proceedings of CVPR, 2023, pp. 2529-2539. https://doi.org/10.1109/CVPR52729.2023.00249.

[7] L. Hu, L. Gao, Z. Liu, C.-M. Pun, W. Feng, AdaBrowse: Adaptive video browser for efficient continuous sign language recognition, in: Proceedings of ACM Multimedia, 2023, pp. 709-718. https://doi.org/10.1145/3581783.3611745.

[8] N.C. Camgoz, O. Koller, S. Hadfield, R. Bowden, Sign language transformers: Joint end-to-end sign language recognition and translation, in: Proceedings of CVPR, 2020, pp. 10023-10033.

[9] S. Xue, L. Gao, L. Wan, W. Feng, Multi-scale context-aware network for continuous sign language recognition, Virtual Reality & Intelligent Hardware 6 (4) (2024) 323-337. https://doi.org/10.1016/j.vrih.2023.06.011.

[10] Y. Chen, F. Wei, X. Sun, Z. Wu, S. Lin, A simple multi-modality transfer learning baseline for sign language translation, in: Proceedings of CVPR, 2022, pp. 5110-5120. https://doi.org/10.1109/CVPR52688.2022.00506.

[11] D. Wei, H. Hu, G.-F. Ma, Part-wise graph Fourier learning for skeleton-based continuous sign language recognition, Journal of Imaging 11 (8) (2025) 286. https://doi.org/10.3390/jimaging11080286.

[12] W. Zhao, W. Zhou, H. Hu, M. Wang, H. Li, Self-supervised representation learning with spatial-temporal consistency for sign language recognition, IEEE Transactions on Image Processing 33 (2024) 4188-4201. https://doi.org/10.1109/TIP.2024.3416881.

[13] P. Jiao, Y. Min, Y. Li, X. Wang, L. Lei, X. Chen, CoSign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20676–20686. https://doi.org/10.1109/ICCV51070.2023.01890.

[14] M. Guan, Y. Wang, G. Ma, et al., MSKA: Multi-stream keypoint attention network for sign language recognition and translation, Pattern Recognit. 165 (2025) 111602. https://doi.org/10.1016/j.patcog.2025.111602.

[15] Q. Guo, Y. Wang, Y. Zhang, Q. Chu, S. Liu, Y. Jiang, HyperSign: Hierarchical hypergraph-based co-occurrence modeling for sign language recognition and translation, Proc. AAAI Conf. Artif. Intell. 40 (6) (2026) 4421-4429. https://doi.org/10.1609/aaai.v40i6.42440.

[16] Q. Zhu, J. Li, F. Yuan, Q. Gan, Multiscale temporal network for continuous sign language recognition, Journal of Electronic Imaging 33 (2) (2024) 023059. https://doi.org/10.1117/1.jei.33.2.023059.

[17] Z. Huang, W. Xue, Y. Zhou, J. Sun, Y. Wu, T. Yuan, et al., Dual-stage temporal perception network for continuous sign language recognition, The Visual Computer 41 (3) (2025) 1971-1986. https://doi.org/10.1007/s00371-024-03516-x.

[18] L. Hu, L. Gao, Z. Liu, W. Feng, Spatial temporal aggregation for efficient continuous sign language recognition, IEEE Transactions on Emerging Topics in Computational Intelligence 8 (6) (2024) 3925-3935. https://doi.org/10.1109/TETCI.2024.3378649.

[19] C. Liu, L. Hu, Rethinking the temporal downsampling paradigm for continuous sign language recognition, Multimedia Systems 31 (2) (2025) 134. https://doi.org/10.1007/s00530-025-01695-8.

[20] H. Hu, J. Peng, Z. Xiao, L. Guo, Y. Hu, D. Wu, Motion-temporal calibration network for continuous sign language recognition, Complex & Intelligent Systems 12 (1) (2026) 35. https://doi.org/10.1007/s40747-025-02156-5.

[21] Y. Jiang, Y. Wang, S. Li, Y. Zhang, Q. Guo, Q. Chu, et al., EvCSLR: Event-guided continuous sign language recognition and benchmark, IEEE Transactions on Multimedia 27 (2025) 1349-1361. https://doi.org/10.1109/TMM.2024.3521750.

[22] L. Gao, P. Shi, L. Hu, J. Feng, L. Zhu, L. Wan, et al., Cross-modal knowledge distillation for continuous sign language recognition, Neural Networks 179 (2024) 106587. https://doi.org/10.1016/j.neunet.2024.106587.

[23] R. Zuo, B.K.-W. Mak, Improving continuous sign language recognition with consistency constraints and signer removal, ACM Transactions on Multimedia Computing, Communications, and Applications 20 (6) (2024) Article 163. https://doi.org/10.1145/3640815.

[24] D. Guo, S. Tang, R. Hong, M. Wang, Review of sign language recognition, translation and generation, Comput. Sci. 48 (3) (2021) 60–70 (in Chinese). https://doi.org/10.11896/jsjkx.210100227.

[25] A.A. Hosain, P.S. Santhalingam, P.H. Pathak, H. Rangwala, J. Kosecka, Hand pose guided 3D pooling for word-level sign language recognition, in: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3428–3438. https://doi.org/10.1109/WACV48630.2021.00347.

[26] Y. Min, A. Hao, X. Chai, X. Chen, Visual alignment constraint for continuous sign language recognition, in: Proceedings of ICCV, 2021, pp. 11542-11551.

[27] R. Zuo, B. Mak, C2SLR: Consistency-enhanced continuous sign language recognition, in: Proceedings of CVPR, 2022, pp. 5131-5140.

[28] K. Renz, N.C. Stache, N. Fox, G. Varol, S. Albanie, Sign segmentation with changepoint-modulated pseudo-labelling, in: Proceedings of CVPR Workshops, 2021, pp. 3403-3412.

[29] A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber, Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of ICML, 2006, pp. 369-376. https://doi.org/10.1145/1143844.1143891.

[30] Z. Wang, D. Li, R. Jiang, M. Okumura, Continuous sign language recognition with multi-scale spatial-temporal feature enhancement, IEEE Access 13 (2025) 5491-5506. https://doi.org/10.1109/ACCESS.2025.3526330.

[31] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, Proc. AAAI Conf. Artif. Intell. 32 (1) (2018) 7444-7452. https://doi.org/10.1609/aaai.v32i1.12328.

[32] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: Proceedings of CVPR, 2019, pp. 12026-12035.

[33] C. Li, Z. Cui, W. Zheng, C. Xu, J. Yang, Spatio-temporal graph convolution for skeleton based action recognition, Proc. AAAI Conf. Artif. Intell. 32 (1) (2018). https://doi.org/10.1609/aaai.v32i1.11776.

[34] Z. Wang, J. Zhang, Continuous sign language recognition based on multi-part skeleton data, in: 2021 International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1–8. https://doi.org/10.1109/IJCNN52387.2021.9534003.

[35] H. Hu, W. Zhao, W. Zhou, H. Li, SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 11221-11239. https://doi.org/10.1109/TPAMI.2023.3269220.

[36] Y. Min, Y. Yang, P. Jiao, Z. Nan, X. Chen, A closer look at skeleton-based continuous sign language recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2025, pp. 4968-4974. https://doi.org/10.1109/ICCVW69036.2025.00515.

[37] G. Ghiasi, T.-Y. Lin, Q.V. Le, DropBlock: A regularization method for convolutional networks, in: Advances in Neural Information Processing Systems 31 (NeurIPS 2018), 2018, pp. 10727–10737.

[38] Q. Zhu, J. Li, F. Yuan, Q. Gan, Continuous sign language recognition based on motor attention mechanism and frame-level self-distillation, Machine Vision and Applications 36 (1) (2025) 7. https://doi.org/10.1007/s00138-024-01633-0.

[39] A. Hao, Y. Min, X. Chen, Self-mutual distillation learning for continuous sign language recognition, in: Proceedings of ICCV,

2021, pp. 11303-11312. https://doi.org/10.1109/ICCV48922.2021.01111.

[40] L. Guo, W. Xue, B. Liu, K. Zhang, T. Yuan, D. Metaxas, Gloss prior guided visual feature learning for continuous sign language recognition, IEEE Transactions on Image Processing 33 (2024) 3486-3495. https://doi.org/10.1109/TIP.2024.3404869.

[41] Z. Ke, S. Liu, Y. Feng, Fine-grained cross-modality consistency mining for continuous sign language recognition, Pattern Recognition Letters 191 (2025) 23-30. https://doi.org/10.1016/j.patrec.2025.02.017.

[42] Y. Zhang, T. Xiang, T.M. Hospedales, H. Lu, Deep mutual learning, in: Proceedings of CVPR, 2018, pp. 4320-4328.

[43] N.C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural sign language translation, in: Proceedings of CVPR, 2018, pp. 7784-7793. https://doi.org/10.1109/CVPR.2018.00812.

[44] H. Zhou, W. Zhou, W. Qi, J. Pu, H. Li, Improving sign language translation with monolingual data by sign back-translation, in: Proceedings of CVPR, 2021, pp. 1316-1325.

[45] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proceedings of ACL, 2002, pp. 311-318. https://doi.org/10.3115/1073083.1073135.

[46] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74-81.

[47] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, P. Luo, Whole-body human pose estimation in the wild, in: Proceedings of ECCV, 2020, pp. 196-214. https://doi.org/10.1007/978-3-030-58545-7_12.

[48] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of CVPR, 2019, pp. 5693-5703.

[49] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: International Conference on Learning Representations (ICLR), 2015.

[50] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: International Conference on Learning Representations, 2017.

[51] L. Hu, L. Gao, Z. Liu, W. Feng, Temporal lift pooling for continuous sign language recognition, in: Proceedings of ECCV, 2022, pp. 511-527. https://doi.org/10.1007/978-3-031-19833-5_30.

[52] L. Hu, L. Gao, Z. Liu, W. Feng, Self-emphasizing network for continuous sign language recognition, Proc. AAAI Conf. Artif. Intell. 37 (1) (2023) 854-862. https://doi.org/10.1609/aaai.v37i1.25164.

[53] J. Ahn, Y. Jang, J.S. Chung, Slowfast network for continuous sign language recognition, in: Proceedings of ICASSP, 2024, pp. 3920-3924. https://doi.org/10.1109/ICASSP48485.2024.10445841.

[54] B. Zhou, Z. Chen, A. Clapés, J. Wan, Y. Liang, S. Escalera, Z. Lei, D. Zhang, Gloss-free sign language translation: Improving from visual-language pretraining, in: Proceedings of ICCV, 2023, pp. 20814-20824. https://doi.org/10.1109/ICCV51070.2023.01908.

[55] A. Yin, T. Zhong, L. Tang, W. Jin, T. Jin, Z. Zhao, Gloss attention for gloss-free sign language translation, in: Proceedings of CVPR, 2023, pp. 2551-2562.

[56] B. Zhang, M. Müller, R. Sennrich, SLTUNET: A simple unified model for sign language translation, in: International Conference on Learning Representations, 2023.

[57] R. Zhao, L. Zhang, B. Fu, C. Hu, J. Su, Y. Chen, Conditional variational autoencoder for sign language translation with cross-modal alignment, Proc. AAAI Conf. Artif. Intell. 38 (17) (2024) 19643-19651. https://doi.org/10.1609/aaai.v38i17.29937.

[58] J. Ye, W. Jiao, X. Wang, Z. Tu, H. Xiong, Cross-modality data augmentation for end-to-end sign language translation, in: Findings of EMNLP, 2023, pp. 13558-13571. https://doi.org/10.18653/v1/2023.findings-emnlp.904.

[59] H. Yao, W. Zhou, H. Feng, H. Hu, H. Zhou, H. Li, Sign language translation with iterative prototype, in: Proceedings of ICCV, 2023, pp. 15546-15555. https://doi.org/10.1109/ICCV51070.2023.01429.