CoSign: Exploring Co-occurrence Signals in Skeleton-based Continuous Sign Language Recognition Supplementary Material

Peiqi Jiao[1,2], Yuecong Min[1,2], Yanan Li[3], Xiaotao Wang[3], Lei Lei[3], Xilin Chen[1,2]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Xiaomi Inc., China
{peiqi.jiao,yuecong.min}@vipl.ict.ac.cn
{liyanan3, wangxiaotao, leilei1}@xiaomi.com, xlchen@ict.ac.cn

This supplementary material provides details that are not shown in the main paper. We first present additional implementation details including keypoints selection, group-wise centralization and network architecture in Sect. A. After that, we present ablations on the effects of auxiliary loss and the process of pre-training in CoSign-2s (§ B.1), dropout probability in group dropout (§ B.2) and the selection of α and β in complementary regularization (§ B.3). Then, we further compare the inference speed without taking CTC decoding time into account (§ B.4), and present ablation results of different signals (§ B.5) and keypoints selection of face (§ B.6). Meanwhile, we evaluate the influence of skeleton quality by conducting experiments under different spatial resolution (§ B.7). All ablation studies are conducted on PHOENIX14 and WER is adopted as the evaluation metric.

A. Additional Implementation Details

A.1. Keypoints Selection and Group-wise Centralization

For pose estimation, we choose HRNet [3] combined with DarkPose [4] trained on COCO-WholeBody [2] as the estimator. We adopt the implementation by MMPose[1] and generate 133 2D keypoints. As we mentioned in Sect. 3.1, we select 77 keypoints and divide them into five groups: 9 for body, 21 for left hand, 21 for right hand, 8 for mouth and 18 for face. We visualize these keypoints in Fig. A, and the coordinate of each group takes the root keypoint as the origin. For body group, we use the mid point of shoulders as the root keypoint.

[1]https://mmpose.readthedocs.io/en/latest/model_zoo/wholebody_2d_keypoint.html#topdown-heatmap-hrnet-dark-on-coco-wholebody

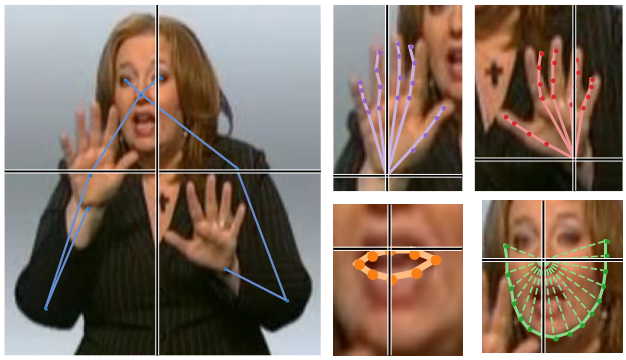

Figure A. Keypoints selected in our approach. The coordinate of each group takes the root keypoint as the origin.

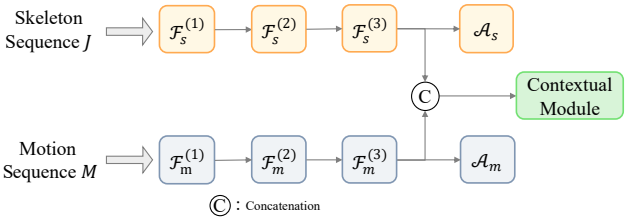

Figure B. The network architecture of late fusion.

A.2. Network Architecture

We detail the output dimension of different layers in both CoSign-1s and CoSign-2s in Table A, where ST-CGN layer i means the i-th ST-GCN layer in each module of the group-wise GCN. As for input, we use 2D coordinates and confidence scores generated by the estimator as the input to CoSign-1s and skeleton branch in CoSign-2s. For motion branch, the bidirectional movements and confidence scores are fed to it. The dropout probability p of both CoSign-1s and CoSign-2s are set to 0.2.

In Sect. 4.3, we compare two different fusion ap-

Table A. Output dimension of different layers in both CoSign-1s and CoSign-2s.

layers	CoSign-1s	CoSign-2s		
		skeleton branch	motion branch	fusion branch
Shared Linear	64	64	64	-
ST-GCN layer 1	64	64	64	128
ST-GCN layer 2	128	128	128	256
ST-GCN layer 3	256	256	256	512
Fusion MLP	1024	-	-	1024
1D CNN	1024	-	-	1024
BiLSTM	1024	-	-	1024

Table B. Inference speed comparison on PHOENIX14 without pose estimation stage taking into account.

	SMKD [1]	Baseline	CoSign-1s	CoSign-2s
Inference Speed w/ Decoding	11.2 seq/s	26.1 seq/s	18.8 seq/s	12.7seq/s
Inference Speed w/o Decoding	16.2 seq/s	82.8 seq/s	39.1 seq/s	19.2 seq/s

proaches, and the network architecture of late fusion is shown in Fig. B. Same as CoSign-2s, we attach two auxiliary classifiers on both skeleton and motion branches, and pre-train the skeleton and motion-based Cosign-1s independently for several epochs. The loss weight λ also keeps same as CoSign-2s.

A.3. Precision of Different Signals

In Fig. 4, we visualize six different sign examples and report the precision of different signals. Different to WER in Fig. 5, the precision of a signal for a sign is the percentage that this signal can make correct predictions for all examples of this sign in both PHOENIX14 dev and test sets.

B. Additional Results

B.1. Ablation on Auxiliary Loss and Pre-training

As we mentioned in Sect. 3.2, we attach two auxiliary CTC losses for both skeleton and motion branches in CoSign-2s, and pre-train them to ensure convergence. Experimental results in Table C show that both auxiliary loss and pre-training can reduce the WER. It illustrates that different branches have different convergence rates and auxiliary loss combined with pre-training can better help each branch converge.

Table C. Ablation results (WER, %) of auxiliary loss and pre-training in CoSign-2s. The best results are **bold**.

Auxiliary loss	Pre-training	Dev	Test
		20.9	21.5
✓		20.8	21.1
✓	✓	**20.7**	20.5

B.2. Ablation on Dropout Probability

In Sect. 3.1, we propose a group dropout mechanism where the dropout mask of each clip is independently sampled from a Bernoulli distribution $B(p)$. We evaluate different dropout probability p with a fixed clip length of 25 and present results in Table D. We choose 0.2 as the default setting as it can provide more diverse masks.

Table D. Ablation results (WER, %) of dropout probability p.

Dropout Probability	Dev	Test
0	21.8	21.9
0.1	21.4	21.4
0.2	**21.2**	21.4
0.3	21.4	22.5
0.4	21.8	21.6

Table E. Results (WER, %) of different α and β on PHOENIX14 Dev/Test sets using CoSign-1s.

α \ β	1	2	4
1	21.1/21.3	21.0/21.4	21.5/21.3
2	21.4/21.6	**20.9**/21.2	21.2/21.9
4	21.2/21.4	21.3/21.2	21.2/21.7

B.3. Ablation on different α and β

In Sect. 3.1, we propose complementary regularization and there are two hyper-parameters α and β control the loss weights of complementary regularization on auxiliary and primary predictions. We perform grid search for them and present the results in Table E. For CoSign-2s, we use the same α and β as CoSign-1s

B.4. Inference Speed

In Sect. 4.2, we report the training and inference efficiency on a NVIDIA GeForce RTX 3090 GPU with data cached. The inference speed is the average sequence per second on PHOENIX14 dev and test sets with a batch size of 1, which includes the time consumed by CTC decoding. Because the decoding time is method independent, we also report the inference speed without taking decoding time into account in Table B. Here, we also do not take

Table F. Ablation results (WER, %) of different signals.

Signals	Dev	Test
Body	40.0	38.7
Left Hand	55.2	54.6
Right Hand	35.7	35.4
Mouth	54.3	53.2
Face	52.2	50.7

Table G. Ablation results (WER, %) of combination of signals.

Combination of Signals	Dev	Test
Body	40.0	38.7
Body+Left Hand	36.6	35.9
Body+Left Hand+Right Hand	28.3	27.8
Body+Left Hand+Right Hand+Mouth	22.4	22.8
Body+Left Hand+Right Hand+Mouth+Face	**21.8**	21.9

Table H. Results (WER, %) on PHOENIX14 with additional face keypoints. Baseline method is CoSign-1s without complementary regularization.

Additional Face Keypoints	Dev	Test
Baseline	**21.8**	21.9
+ Keypoints of Eyebows	21.9	22.3
+ Keypoints of Eyes	**21.8**	22.4
+ Keypoints of Nose	**21.8**	22.0

pose estimation stage into account. For the inference efficiency, we make some attempts and provide the relationship between SLR performance and the spatial resolution in Sect. B.7. CoSign can achieve comparable results even when the spatial resolution is reduced to 96×96, which reveals the potential of CoSign in improving the inference efficiency. Meanwhile, we believe the speed of extracting skeleton from videos is not a bottleneck with the development of pose estimation (*e.g.*, MediaPipe Holistic can run in near real-time even on midtier devices like Samsung S9+ [2]).

B.5. Ablation on Different Signals

In Sect. 4.3, we evaluate the performance of different *trained* models on a specific group by masking keypoints of other groups and fine-tuning the model with *frozen* feature extractor. We further conduct experiments by directly training CoSign-1s without $\mathcal{L}_{CR}$ using a specific group as input and present the results in Table F. We can see using the signal from mouth or face only can achieve a WER around 50% which indicates these signals also play a critical role in CSLR. Meanwhile, as shown in Table G, the combined use of different signals could bring a WER decrease, which illustrates every signal is indispensable.

B.6. Ablation on keypoints selection of face

As mentioned in Sect. A, we only use keypoints of the cheek part during the keypoints selection of the face. A comparison study in Table H shows that adding keypoints out of cheek part can not bring performance gains in our preliminary experiments. We assume that this is because current datasets are collected under constrained conditions (PHOENIX14 is collected from TV weather forecasts) with limited samples, facial expressions play a limited role and adding more keypoints may lead to overfitting and affect the extraction of other useful information. Therefore, we only utilize keypoints of the cheek part for efficiency.

[2] https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html

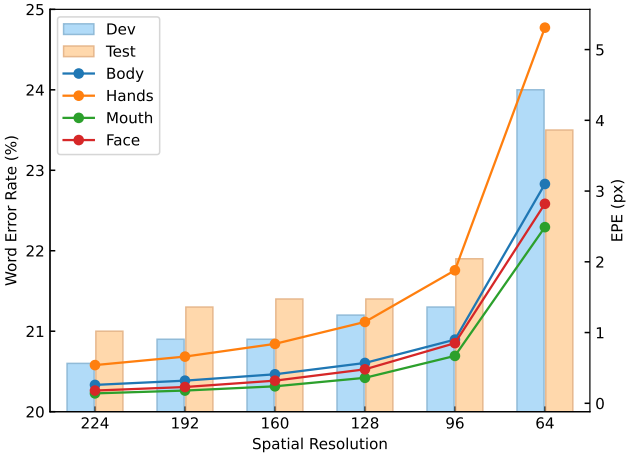

Figure C. WER (%) on PHOENIX14 dev and test sets, and EPE (px) of each group based on different spatial resolution. WER is labeled on the left axis and EPE is labeled on the right.

B.7. Influence of Skeleton Quality

As mentioned in Sect. 3.1, inaccurate estimations may affect the accuracy of CSLR models. To evaluate the influence of skeleton quality, we conduct experiments by reducing the spatial resolution for pose estimation. Based on PHOENIX14, we generate six degrees of estimation qualities of skeleton data by first resizing the frames in each sign video to six different spatial resolution: 224×224, 192×192, 160×160, 128×128, 96×96 and 64×64. Then we conduct pose estimation on these sets and train CoSign-1s using them.

As shown in Fig. C, the WER doesn't have a significant increase until spatial resolution is reduced to 64×64, where the WER increases by 3.4%/2.5% on the Dev/Test sets. Furthermore, we use mean end-point-error (EPE) to measure the skeleton quality of different degrees. Specifically, we use $\mathbf{J} = \{\mathbf{J}_1, \cdots, \mathbf{J}_T\}$ to represent the skeleton sequence estimated under resolution 256×256. Each skeleton frame contains K keypoints $\mathbf{J}_i = \{\mathbf{J}_{i,k} \in \mathbb{R}^2 | k = 1, \cdots, K\}$. We view the keypoints with a confidence score greater than 0.5 as ground truth. For each degree d, we divide all keypoints into four groups (body, hands, mouth and face) and

Table I. Percentage (%) of keypoints with a confidence score greater than 0.5 under different spatial resolution.

Group	Spatial Resolution					
	224	192	160	128	96	64
Body	99	99	99	99	99	98
Hands	88	88	88	88	87	83
Mouth	100	100	100	100	100	99
Face	100	100	100	100	100	99

the EPE of group α (denoted as $\mathcal{G}_\alpha$) is defined as:

$$\begin{aligned} \text{EPE}(d, \alpha) &= \frac{1}{T}\sum_{i=1}^{T}\frac{1}{|N_g|}\sum_{k\in N_g}||\mathbf{J}_{i,k} - \mathbf{J}_{i,k}^{d}||_2 \\ N_g &= \{k|k \in \mathcal{G}_\alpha, conf(\mathbf{J}_{i,k}) > 0.5\}, \end{aligned} \tag{1}$$

where $\mathbf{J}_{i,k}^{d}$ means the keypoint estimated under degree d and $conf(\mathbf{J}_{i,k})$ means the confidence score of $\mathbf{J}_{i,k}$.

A greater EPE represents a lower skeleton quality of this group. Fig. C visualizes the EPE of each group under different degrees. Although hands often occupy smaller regions than body, they have a much larger EPE than body (5.3 vs. 3.1 pixel under 64×64 resolution), which indicates keypoints of hands significantly suffer from inaccurate estimation. This phenomenon reminds us that estimating keypoints of hands needs a higher resolution and a lower resolution (128×128) is enough for keypoints of other groups which can further reduce the computation cost of pose estimation and make the proposed approach more practicable.

Finally, for each group, we calculate the percentage of keypoints with a confidence score greater than 0.5 and visualize the results in Table I. We can see even keypoints of hands have inaccurate estimation under resolution 64×64, the percentage of keypoints with high confidence scores doesn't reduce much. This phenomenon shows more efforts are needed to handle the estimation noise in skeleton-based CSLR.

References

[1] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11303–11312, 2021. 2

[2] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[3] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 1

[4] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2020. 1