HyperSign: Hierarchical Hypergraph-based Co-occurrence Modeling for Sign Language Recognition and Translation

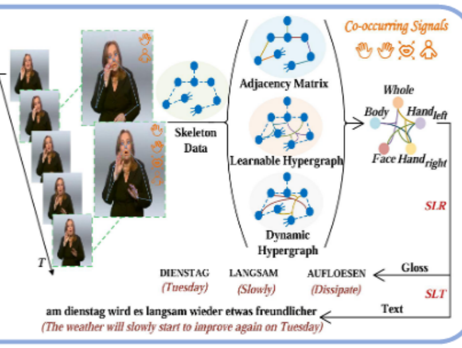
Qianren Guo[1], Yuehang Wang[1], Yongji Zhang[1], Qi Chu[2], Sen Liu[1], Yu Jiang[1,*]

[1]*College of Computer Science and Technology, Jilin University*

[2]*College of Software, Jilin University*

Abstract

Effectively capturing co-occurrence signals, such as hand shapes, facial expressions, and body postures, is critical for semantic understanding in sign language recognition (SLR) and translation (SLT). Although skeleton data offer greater efficiency and robustness than RGB inputs, existing methods typically rely on pairwise graph structures, limiting their ability to model complex high-order interactions across body regions. To address this limitation, we propose HyperSign, a hierarchical hypergraph neural network that systematically captures high-order co-occurrence patterns among diverse body parts. The Co-occurrence Graph Perception Module jointly learns relational structures via three complementary pathways. To further enhance structural modeling and semantic consistency, a Meta-Part Hypergraph Fusion Module abstracts feature streams from the hands, face, and body into unified hypergraph nodes, while leveraging empirically derived co-occurrence priors to model high-order cross-part dependencies. Moreover, an uncertainty-aware collaborative distillation mechanism guides the model to focus on critical body regions. Extensive experiments on standard SLR and SLT benchmarks demonstrate that HyperSign not only outperforms existing skeleton-based approaches in both speed and accuracy but also achieves competitive or superior results compared to several SOTA RGB-based methods across multiple evaluation metrics.



Method

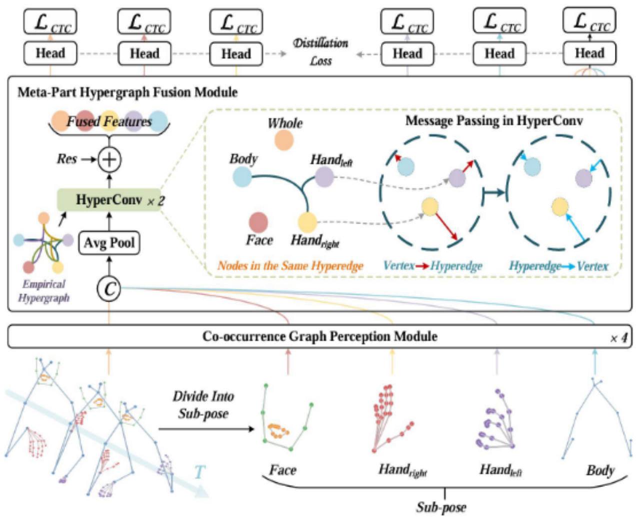


Figure 1: Detailed architecture of HyperSign, integrating sub-pose features to capture high-order body-part relations.

★ HyperSign Framework

As illustrated in Figure 1, HyperSign consists of two core modules. The CGP module takes multi-stream inputs from the hands, face, body, and whole skeleton to model high-order semantic co-occurrence patterns within local regions. The MPHF module then explicitly captures cross-part structural dependencies by leveraging linguistically inspired co-occurrence priors. To handle dynamic shifts in attention during sign expressions, we introduce an UACD mechanism.

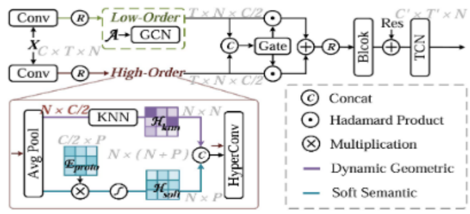


Figure 2: Structure of the Co-occurrence Graph Perception Module.

★ Co-occurrence Graph Perception Module

To effectively model high-order semantic co-occurrence relationships among joints in skeleton sequences, we design the CGP Module, which integrates three complementary topological structures:

➤ ***Static Graph Path***

▸ The traditional graph convolution operation captures physical joint relationships as follows:

$$\mathbf{F}_{\mathrm{GCN}} = \mathrm{GCN}(\mathbf{X}_a, \mathbf{A}) \in \mathbb{R}^{T\times N\times C/2}.$$

➤ ***Dynamic Geometric Hypergraph Path***

▸ Apply temporal average pooling to obtain a compact representation:

$$\mathbf{F}_{\mathrm{pool}} = \mathrm{AvgPool}_T(\mathbf{X}_b) \in \mathbb{R}^{N\times C/2}.$$

▸ Compute the pairwise squared Euclidean distances between joints:

$$\mathcal{D}_{i,j} = \|\mathbf{F}_{\mathrm{pool},i} - \mathbf{F}_{\mathrm{pool},j}\|_2^2.$$

➤ ***Soft Semantic Hypergraph Path***

▸ Capture high-order semantic co-occurrence relationships among joints:

$$\mathbf{H}_{\mathrm{soft}} = \mathrm{Softmax}(\mathbf{F}_{\mathrm{pool}} \cdot \mathbf{E}_p^{\top}) \in \mathbb{R}^{N\times P}.$$

Experiment

Method	Pub	PHOENIX-2014T (DEV / TEST)		CSL-Daily (DEV / TEST)	
		ROUGE / BLEU4	ROUGE / BLEU4	ROUGE / BLEU4	ROUGE / BLEU4
RGB-based					
STMC (Zhou et al. 2021b)	TMM	48.24 / 24.08	46.65 / 23.65	– / –	– / –
MMTLB (Chen et al. 2022a)	CVPR	53.10 / 27.61	52.65 / 28.39	53.38 / 24.42	53.25 / 23.92
TwoStream-SLT (Chen et al. 2022b)	NeurIPS	54.08 / 28.66	53.48 / 28.95	55.10 / 25.76	55.72 / 25.79
SignBERT+ (Hu et al. 2023a)	TPAMI	51.12 / 24.95	50.63 / 25.70	– / –	– / –
GFSLT-VLP (Zhou et al. 2023)	ICCV	43.72 / 22.12	42.49 / 21.44	36.70 / 11.07	36.44 / 11.00
GASLT (Yin et al. 2023)	CVPR	– / –	39.07 / 15.74	– / –	20.35 / 4.07
IP-SLT (Yao et al. 2023)	ICCV	54.43 / 28.22	53.72 / 27.97	44.33 / 16.74	44.09 / 16.72
CV-SLT (Zhao et al. 2024)	AAAI	**55.05 / 29.55**	**54.54 / 29.52**	**56.36 / 28.24**	**57.06 / 28.94**
Skeleton-based					
TwoStream-SLT (Chen et al. 2022b)	NeurIPS	53.32 / 28.10	53.19 / 28.42	54.03 / 25.01	55.07 / 25.42
SignBERT+ (Hu et al. 2023a)	TPAMI	45.53 / 19.86	44.89 / 20.41	– / –	– / –
MSKA (Guan et al. 2025)	PR	52.67 / 27.63	53.54 / 29.03	53.54 / 25.16	54.04 / 25.52
HyperSign (Ours)	–	**54.36 / 28.81**	**54.26 / 29.35**	**55.21 / 26.39**	**56.14 / 26.87**
Improvement	–	+1.04 / +0.71	+0.72 / +0.32	+1.18 / +1.23	+1.07 / +1.35

Table 1: ROUGE and BLEU4 scores (higher is better) on PHOENIX-2014T and CSL-Daily. Best results are in bold.

For the SLT task, we adopt ROUGE and BLEU4 as evaluation metrics. As shown in Table 1, HyperSign consistently achieves the best performance across all datasets and splits. On the PHOENIX-2014T dataset, HyperSign obtains ROUGE/BLEU4 scores of 54.36/28.81 on the DEV set and 54.26/29.35 on the TEST set, surpassing the best skeleton-based SOTA, by +1.04/+0.71 and +0.72/+0.32, respectively.

Conclusion

This paper presents HyperSign, the first skeleton-based method that systematically models high-order semantic co-occurrence for sign language recognition and translation. HyperSign jointly constructs physical graphs, geometric hypergraphs, and semantic prototype hypergraphs to capture multidimensional co-occurrence patterns at the joint level, while a meta-part hypergraph further models cross-region semantic interactions. In addition, an Uncertainty-Aware Collaborative Distillation mechanism is introduced to enhance the model's focus on critical expressive regions.