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
{guoqr22, yuehang22, zyj23, chuqi23, senliu24}@mails.jlu.edu.cn, jiangyu2011@jlu.edu.cn

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: (1) traditional graph convolutions for modeling physical joint connections, (2) dynamic geometric hypergraphs constructed via k-nearest neighbors to encode local spatial patterns, and (3) soft hypergraphs generated by learnable prototypes to reveal latent semantic associations. 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 (e.g., PHOENIX-2014, PHOENIX-2014T, and CSL-Daily) 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 state-of-the-art RGB-based methods across multiple evaluation metrics.

Introduction

Sign language is a primary means of communication within the deaf community, where semantics are conveyed through coordinated movements involving the hands, face, and body. These articulations exhibit strong spatial structures and intricate temporal dynamics. Within this context, Sign Language Recognition (SLR) aims to map sign language video sequences to continuous streams of gloss-level labels, while Sign Language Translation (SLT) extends this process by generating grammatically correct and semantically fluent sentences from the recognized glosses. Both tasks pose significant challenges due to the inherent complexity and structural richness of sign expressions.

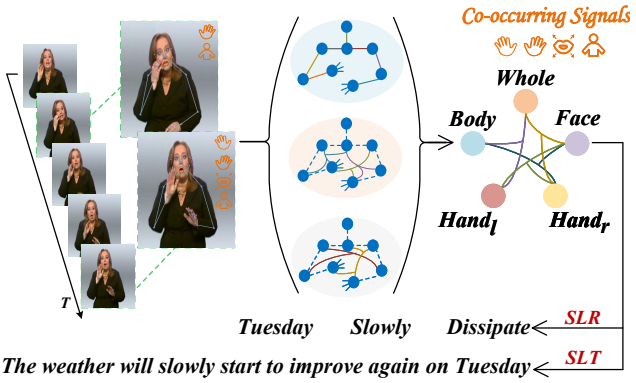

Figure 1: Hierarchical Hypergraph Modeling of Co-occurring Signals in Sign Language. The figure illustrates co-occurring signals in sign language, where expressive units emerge from interactions among the hands, face, and body. HyperSign models these high-order dependencies via hierarchical hypergraphs for accurate gloss prediction and translation.

In recent years, RGB-based approaches for SLR and SLT have achieved substantial progress. However, RGB inputs often suffer from visual redundancy and incur high computational costs. By contrast, skeleton data provide a compact and structured representation that filters out irrelevant visual content, allowing models to focus on the essential structural dynamics of sign gestures. Despite these advantages, skeleton-based methods generally fall short of their RGB counterparts. CoSign (Jiao et al. 2023) attributes this performance gap to the uniform treatment of all keypoints as a single input, which hampers the model's ability to learn structured semantic representations. To mitigate this issue, it proposes a part-based strategy that partitions keypoints into groups and processes them separately, thereby preserving localized structural information. Building upon this idea, MSKA (Guan et al. 2025) incorporates attention mechanisms to model the varying strengths of inter-joint re-

*Corresponding authors.

lationships, enabling the network to capture latent dependencies embedded in the skeletal topology. In addition, several skeleton-based methods for SLR and SLT (Pu, Lim, and Chong 2024; Lin et al. 2024; Li et al. 2025) have demonstrated encouraging progress. However, their core modeling paradigms largely remain confined to pairwise relational structures. Such modeling schemes exhibit clear limitations when tasked with capturing the higher-order relational dependencies that are prevalent in sign language. In practice, expressive units in sign language, such as a specific hand configuration or a compound posture that conveys interrogative intent, are typically formed through coordinated and synchronous movements involving multiple joints, as illustrated in Fig. 1. These functional entities emerge from complex co-occurrence patterns that span across different body regions and cannot be adequately represented by conventional pairwise graphs.

To address the limitations of conventional pairwise relational modeling in sign language understanding, we introduce **HyperSign**—a hierarchical hypergraph neural network framework tailored to capture the high-order relational structures intrinsic to sign language expressions. At the heart of HyperSign is a hierarchical disentanglement mechanism that systematically models multi-granular, heterogeneous, and semantically co-occurring dependencies. This is achieved through a bottom-up process that maps low-level keypoints to higher-level body components, enabling structured representation of complex sign semantics.

At the joint level, we introduce a **C**o-occurrence **G**raph **P**erception (**CGP**) Module that concurrently models three complementary types of structural dependencies. First, a physical graph encodes the topological connectivity among joints, capturing intrinsic anatomical constraints. Second, a dynamic geometric hypergraph captures local spatial correlations to model coordinated joint patterns. Third, a learnable prototype hypergraph captures semantic consistency and reveals latent co-occurrence structures. By structurally disentangling these components, the module enables unified and flexible modeling of heterogeneous graph representations, allowing the network to jointly learn multidimensional spatial dependencies and local co-occurrence patterns within a single framework.

At the body-part level, we propose the **M**eta-**P**art **H**ypergraph **F**usion (**MPHF**) Module, which abstracts the feature representations of key anatomical regions such as the hands, face, and body into nodes of a meta-level hypergraph. To capture high-order co-occurrence and semantic interactions across body parts, the module introduces prior-guided meta-hyperedges that explicitly encode inter-part relational structures. In addition, to account for the dynamic shift of information focus in sign language expressions, we propose an **U**ncertainty-**A**ware **C**ollaborative **D**istillation (**UACD**) mechanism. This method estimates the uncertainty of each body-part stream and dynamically assigns reliability weights to construct a weighted consensus teacher. By emphasizing more reliable and informative regions, the model is guided to attend more effectively to critical body parts during training.

The main contributions are summarized as follows:

- We introduce HyperSign, a hierarchical hypergraph neural network that offers a new perspective on structured representation in sign language by jointly modeling high-order co-occurrence across both joint and body-part levels for the first time.
- We design the Co-occurrence Graph Perception and Meta-Part Hypergraph Fusion modules, integrating physical structural graphs, dynamic geometric hypergraphs, learnable semantic hypergraphs, and linguistically informed cross-part hyperedges to model high-order semantic synergy from local joints to the entire body.
- We introduce an Uncertainty-Aware Collaborative Distillation mechanism to improve the model's focus on critical body parts and enhance expression understanding.
- Extensive experiments on several standard SLR and SLT benchmarks demonstrate the superiority of our method in both speed and accuracy, outperforming existing skeleton-based models.

Related Works

Sign Language Recognition and Translation

Skeleton-based representations have gained increasing attention in both sign language recognition (SLR) and sign language translation (SLT) due to their robustness to viewpoint changes and occlusions. The goal of SLR is to transcribe continuous video or skeleton sequences into gloss-level word sequences, while SLT further maps these sequences into grammatically correct spoken-language sentences. Most existing methods employ CNNs, GCNs, or Transformer-based architectures to extract spatiotemporal features (Niu and Mak 2020; Hu et al. 2023b; Zheng et al. 2023; Jiang et al. 2024; Zhang et al. 2025; Guan et al. 2025). For instance, CoSign captures co-occurrence patterns between hands and torso via skeleton graphs (Jiao et al. 2023). The decoding typically involves connectionist temporal classification (CTC) (Graves et al. 2006), though its weak supervision can hinder convergence. To mitigate this, recent works leverage frame-level knowledge distillation (Chen et al. 2022b; Guo et al. 2023; Guan et al. 2025), improving feature discriminability. SLT is often modeled as a neural machine translation task, where visual encoders extract semantic features from video inputs, which are then decoded into natural language text. To enhance translation quality, many approaches incorporate semantic-level supervision, such as pretraining on SLR or adopting joint training strategies (Chen et al. 2022a; Zhou et al. 2021b; Guan et al. 2025). In our work, we follow the latter strategy to improve cross-task generalization through unified optimization of SLR and SLT.

Hypergraph Learning Methods

Unlike traditional pairwise graph structures, hypergraphs connect multiple nodes simultaneously via hyperedges, offering a more expressive and flexible mechanism for modeling complex high-order relationships. Hypergraph learning is first introduced by (Zhou, Huang, and Schölkopf

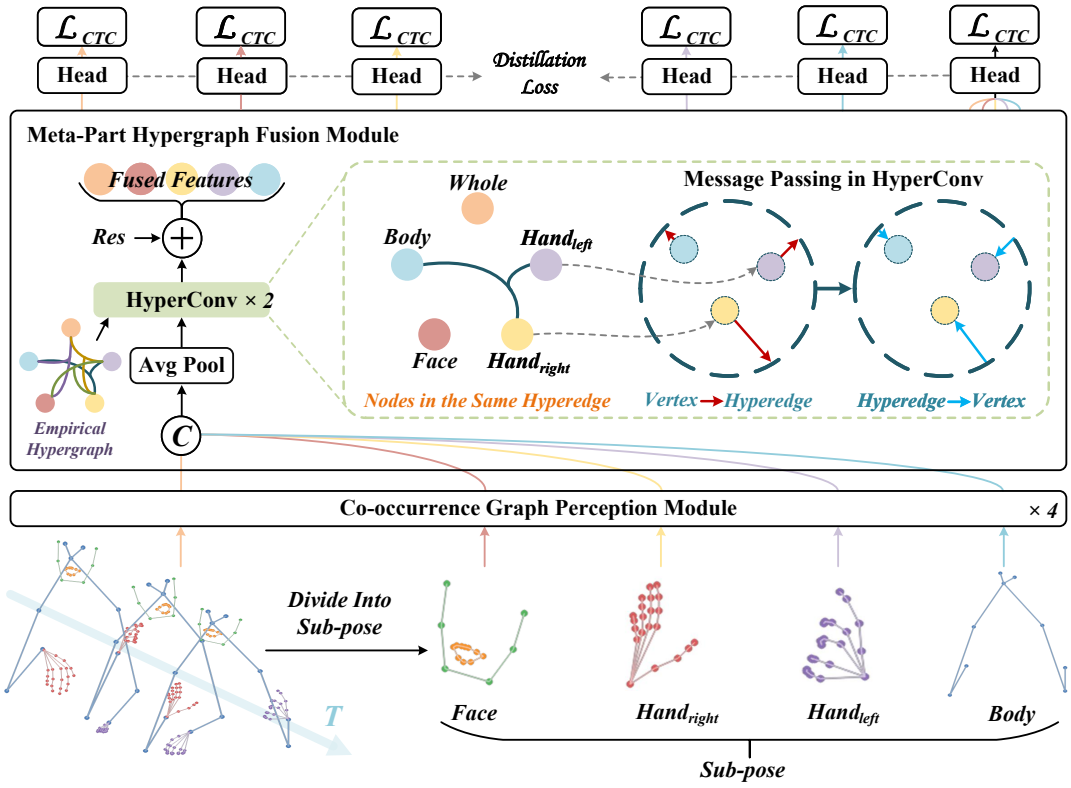

Figure 2: Detailed architecture of HyperSign, integrating sub-pose features to capture high-order body-part relations.

2006), whose core idea is to explicitly model high-order dependencies among samples through hyperedge construction and to enhance learning by propagating information via hypergraph convolution operations. Subsequent studies (Feng et al. 2019; Gao et al. 2022; Lei et al. 2025; Guo et al. 2026) further validate that well-designed hyperedges effectively capture non-local and multi-way interactions. Benefiting from their inherent capability to model complex structures, hypergraph methods have recently been applied to various vision tasks with promising results (Feng et al. 2024; Zhou et al. 2024).

Methods

Preliminaries

We propose a unified framework, HyperSign, that jointly models skeleton-based SLR and SLT. Inspired by prior works (Chen et al. 2022b; Guan et al. 2025), we use HRNet (Wang et al. 2020) to extract 133 skeletal keypoints. From these, we select 76 representative joints for downstream modeling, including 42 from the hands, 25 from the face, and 9 from the body. As illustrated in Fig. 2, 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.

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: a physical structural graph, a dynamic geometric hypergraph, and a soft semantic hypergraph. This unified design enables the capture of localized collaborative patterns from multiple perspectives, as illustrated in Fig. 3.

Given a skeleton sequence of length T, we represent it as a tensor $\mathbf{X} \in \mathbb{R}^{C\times T\times N}$, where $C = 3$ denotes the feature channels comprising two normalized spatial coordinates and one confidence score, T is the temporal length, and N is the number of selected keypoints. The input tensor is first processed by two independent 1×1 convolutional branches:

$$\mathbf{X}_a, \mathbf{X}_b = \mathcal{R}(\mathrm{Conv}_{\theta_a}(\mathbf{X})),\ \mathcal{R}(\mathrm{Conv}_{\theta_b}(\mathbf{X})) \in \mathbb{R}^{T\times N\times C/2}, \tag{1}$$

where $\mathcal{R}(\cdot)$ denotes a reshape operation.

Static Graph Path. GCN is applied using a fixed adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ encoding anatomical connections. The traditional graph convolution operation captures physical joint relationships as follows:

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

Dynamic Geometric Hypergraph Path. To model the local spatial proximity among joints, we dynamically construct a K-nearest neighbor (KNN) hypergraph based on the temporally pooled pose features $\mathbf{X}_b$. First, we apply temporal average pooling to obtain a compact representation:

$$\mathbf{F}_{\text{pool}} = \text{AvgPool}_T(\mathbf{X}_b) \in \mathbb{R}^{N \times C/2}. \tag{3}$$

We then compute the pairwise squared Euclidean distances between joints:

$$\mathcal{D}_{i,j} = \|\mathbf{F}_{\text{pool},i} - \mathbf{F}_{\text{pool},j}\|_2^2. \tag{4}$$

For each joint i, we select its $k = 8$ nearest neighbors to form a hyperedge, resulting in N hyperedges in total. Each hyperedge connects one center node and its corresponding k nearest nodes. The hypergraph is encoded using an incidence matrix $\mathbf{H}_{\text{knn}} \in \mathbb{R}^{N \times N}$, where $\mathbf{H}_{\text{knn}}[i, j] = 1$ if node j belongs to the i-th hyperedge (centered at node i), and 0 otherwise.

Soft Semantic Hypergraph Path. To capture high-order semantic co-occurrence relationships among joints, we construct a soft semantic hypergraph based on learnable prototype-based attention. Specifically, we introduce a set of P learnable semantic prototypes $\mathbf{E}_p \in \mathbb{R}^{P \times C/2}$ and compute a weighted incidence matrix as:

$$\mathbf{H}_{\text{soft}} = \text{Softmax}(\mathbf{F}_{\text{pool}} \cdot \mathbf{E}_p^\top) \in \mathbb{R}^{N \times P}, \tag{5}$$

each entry $\mathbf{H}_{\text{soft}}[i, j]$ represents the soft affiliation strength of joint i to semantic prototype j, thereby forming a set of soft hyperedges that encode cross-joint semantic affinities.

Hypergraph Computation and Fusion. To comprehensively capture localized co-occurrence patterns, we fuse outputs from three complementary paths. Specifically, we concatenate the semantic and geometric hypergraphs to form a unified incidence matrix:

$$\mathbf{H}_{\text{hyper}} = [\mathbf{H}_{\text{soft}} \mid \mathbf{H}_{\text{knn}}] \in \mathbb{R}^{N \times (P+N)}. \tag{6}$$

Subsequently, $\mathbf{X}_b$ is propagated through a hypergraph convolution operation to obtain enhanced representations:

$$\mathbf{F}_{\text{HYP}} = \text{HyperConv}(\mathbf{X}_b, \mathbf{H}_{\text{hyper}}) \in \mathbb{R}^{T \times N \times C/2}, \tag{7}$$

where the hypergraph convolution (Gao et al. 2020) is defined as:

$$\text{HyperConv}(\mathbf{F}, \mathbf{H}) = \mathbf{F} + \mathbf{D}_v^{-\frac{1}{2}} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^\top \mathbf{D}_v^{-\frac{1}{2}} \mathbf{F} \mathbf{\Theta}, \tag{8}$$

in which $\mathbf{D}_v$ and $\mathbf{D}_e$ are the degree matrices of the vertices and hyperedges, respectively, $\mathbf{W}$ is a diagonal matrix of hyperedge weights, $\mathbf{\Theta}$ is the learnable transformation matrix, and $\mathbf{F}$ is the input node feature matrix.

To adaptively combine structural and hypergraph-based representations, we employ a gating mechanism:

$$\mathbf{G} = \sigma\left(\text{MLP}\left([\mathbf{F}_{\text{GCN}} \mid \mathbf{F}_{\text{HYP}}]\right)\right), \tag{9}$$

$$\mathbf{F}_{\text{fused}} = \mathbf{G} \odot \mathbf{F}_{\text{GCN}} + (1 - \mathbf{G}) \odot \mathbf{F}_{\text{HYP}}, \tag{10}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function and $\odot$ represents the Hadamard product.

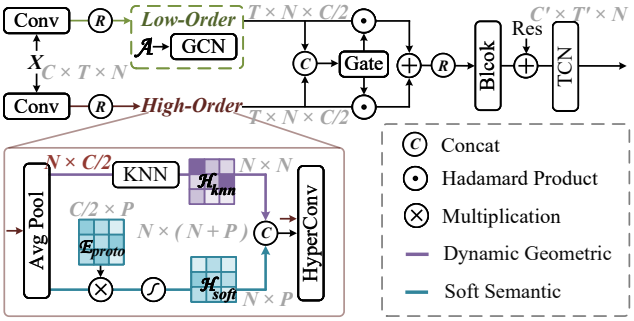

Figure 3: Structure of the Co-occurrence Graph Perception Module.

Finally, the fused features are reshaped and passed through a 1×1 convolutional layer to adjust the output dimensionality. A residual path is added for stability. The resulting features are fed into a Temporal Convolutional Network module, which effectively captures long-range temporal dependencies across frames:

$$\mathbf{Y} = \text{TCN}\Big(\text{Conv}_{\text{final}}\left(\mathcal{R}\left(\mathbf{F}_{\text{fused}}\right)\right) + \text{Conv}_{\text{residual}}(\mathbf{X})\Big) \in \mathbb{R}^{C' \times T' \times N}. \tag{11}$$

Meta-Part Hypergraph Fusion Module

As illustrated in Fig. 2, we propose the MPHF module to model high-level interactions among full-body regions. The part-level features are extracted from the output of the four-stage CGP module. Specifically, we define $\mathcal{P} = 5$ semantic regions: left hand, right hand, face, body, and whole, each serving as a node in the hypergraph. Let $\mathbf{X}^{\boldsymbol{\rho}} \in \mathbb{R}^{4C \times \frac{T}{4} \times N}$ denote the feature tensor for part $\boldsymbol{\rho}$. To obtain frame-level representations, we apply average pooling along the keypoint dimension: $\mathbf{F}^{\boldsymbol{\rho}} = \text{AvgPool}_N(\mathbf{X}^{\boldsymbol{\rho}}) \in \mathbb{R}^{4C \times \frac{T}{4}}$. The resulting feature is then transposed to a frame-major format, producing $\widetilde{\mathbf{F}}^{\boldsymbol{\rho}} \in \mathbb{R}^{\frac{T}{4} \times 4C}$. All part-wise features are then stacked along the part axis to form the unified meta-part representation: $\mathbf{X}_{\text{part}} \in \mathbb{R}^{\frac{T}{4} \times \mathcal{P} \times 4C}$.

Based on prior knowledge of co-occurrence and coordination among body parts, we define the hyperedge set as:

$$\mathcal{E} = \{\{0, 1, 2\}, \{2, 3, 4\}, \{1, 3, 4\}, \{0, 3, 4\}\},$$

where node indices represent: 0—whole, 1—body, 2—face, 3—right hand, 4—left hand. The corresponding incidence matrix is:

$$\mathbf{H}_{\text{meta}} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \in \{0, 1\}^{5 \times 4},$$

where each row corresponds to a semantic part node and each column to a hyperedge.

Hypergraph Reasoning and Fusion. To model the semantic dependencies among different parts, we apply two layers of hypergraph convolution over the input enhanced with positional embeddings. The input is defined as:

$$\mathbf{Z}^{(0)} = \mathbf{X}_{\text{part}} + \mathbf{E}_{\text{part}} \in \mathbb{R}^{\frac{T}{4} \times \mathcal{P} \times 4C}, \tag{12}$$

where $\mathbf{E}_{\text{part}} \in \mathbb{R}^{1 \times \mathcal{P} \times 4C}$ is a learnable part-level embedding. The hypergraph reasoning process is formulated as:

$$\mathbf{Z}^{(l+1)} = \text{HyperConv}^{(l+1)}(\mathbf{Z}^{(l)}, \mathbf{H}_{\text{meta}}), \quad l = 0, 1, \tag{13}$$

which allows semantic information to propagate across parts within each frame, enabling explicit modeling of high-order co-articulations. The final hypergraph-enhanced features $\mathbf{Z}^{(2)} \in \mathbb{R}^{\frac{T}{4} \times \mathcal{P} \times 4C}$ are split along the part dimension and fused with the original part-wise features through residual connections:

$$\mathbf{F}_{\text{out}}^{\boldsymbol{\rho}} = \widetilde{\mathbf{F}}^{\boldsymbol{\rho}} + \mathbf{Z}^{(2)}[:, i_{\boldsymbol{\rho}}, :] \in \mathbb{R}^{\frac{T}{4} \times 4C}, \quad \boldsymbol{\rho} = 1, \ldots, \mathcal{P}, \tag{14}$$

where $i_{\boldsymbol{\rho}}$ denotes the index of part $\boldsymbol{\rho}$ in the hypergraph node set.

Loss Function

Uncertainty-Aware Collaborative Distillation. To further enhance the collaborative modeling of multi-view semantics, we introduce an UACD mechanism. This mechanism adaptively computes confidence weights based on the entropy discrepancies of prediction distributions from different visual heads and fuses their knowledge to construct a soft teacher distribution. During training, the outputs of all visual heads are aligned with this soft teacher via KL divergence, guiding the model to more stably attend to critical body parts.

Specifically, we concatenate the enhanced features of all body parts along the channel dimension to obtain a fused feature stream: $\mathbf{F}_{\text{ensemble}} = \text{Concat}\big(\mathbf{F}_{\text{out}}^{\boldsymbol{\rho}}\big)_{\boldsymbol{\rho}=1}^{5}$. After obtaining the enhanced part-level features and the fused representation, we feed them into six parallel Visual Head modules. Each visual head consists of a linear projection layer, a nonlinear activation function, a normalization layer, and a classification output layer, independently producing frame-level classification outputs denoted as $\{L_i \in \mathbb{R}^{\frac{T}{4} \times V} \mid i = 1, \ldots, 6\}$. We then compute the predicted probability distribution for each visual head's logits L_i as $\mathbf{Q}_i = \text{softmax}(L_i)$, and estimate the prediction entropy as an uncertainty measure:

$$H_i = -\sum_{v=1}^{V} \mathbf{Q}_i(v) \log(\mathbf{Q}_i(v) + \epsilon), \tag{15}$$

where ϵ is a small constant to ensure numerical stability, and V is the number of classes. The teacher logits are obtained by confidence-weighted summation $L_{\text{teacher}} = \sum_{i=1}^{6} w_i \cdot L_i$, where the confidence weights are computed as $w_i = \frac{\exp(-H_i/\tau)}{\sum_{j=1}^{6} \exp(-H_j/\tau)}$, with temperature hyperparameter τ controlling the smoothness of the weight distribution. The teacher logits are then softened by a temperature factor T_d to obtain the soft label distribution $\mathbf{Q}_{\text{teacher}} = \text{softmax}(L_{\text{teacher}}/T_d)$, and each student head prediction is similarly softened as $\mathbf{Q}_i(T_d) = \text{softmax}(L_i/T_d)$. Finally, the distillation loss is computed as the average Kullback-Leibler divergence between the teacher distribution and each student prediction:

$$\mathcal{L}_{\text{distill}} = \frac{1}{6} \sum_{i=1}^{6} \text{KL}(\mathbf{Q}_{\text{teacher}} \parallel \mathbf{Q}_i(T_d)). \tag{16}$$

HyperSign-based SLR. For the skeleton-based sign language recognition task, each semantic stream is independently fed into a recognition branch supervised by the CTC loss to predict the gloss sequence. To further enhance multi-view collaboration, we incorporate the UACD-based distillation loss $\mathcal{L}_{\text{distill}}$ described earlier. The overall SLR objective is formulated as:

$$\mathcal{L}_{\text{SLR}} = \frac{1}{6} \sum_{i=1}^{6} \text{CTC}(L_i, G) + \alpha \cdot \mathcal{L}_{\text{distill}}, \tag{17}$$

where $\alpha = 0.5$ is a hyperparameter that balances the contribution of the distillation loss.

HyperSign-based SLT. For sign language translation, the fused features are projected via a lightweight MLP into the input embedding space of a pretrained mBART (Liu et al. 2020) model. The entire translation branch is optimized end-to-end using a combination of the recognition loss $\mathcal{L}_{\text{SLR}}$ and the sequence-level cross-entropy loss $\mathcal{L}_{\text{CE}} = \text{CrossEntropy}(S, \hat{S})$, where S and $\hat{S}$ denote the ground-truth and predicted spoken language sentences, respectively. The final SLT objective is defined as

$$\mathcal{L}_{\text{SLT}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{SLR}}. \tag{18}$$

Experiments

Datasets and Evaluation Metrics

We evaluate our method on three public benchmark datasets for sign language recognition and translation: PHOENIX-2014 (Koller, Forster, and Ney 2015), PHOENIX-2014T (Camgoz et al. 2018), and CSL-Daily (Zhou et al. 2021a). PHOENIX-2014 is used for SLR, while PHOENIX-2014T and CSL-Daily provide annotations for both SLR and SLT. All ablation studies are conducted on the PHOENIX-2014T dataset.

- **PHOENIX-2014.** A German sign language dataset with 1,295 glosses, containing 5,672 training, 540 validation, and 629 test samples from 9 signers.
- **PHOENIX-2014T.** An extended version of PHOENIX-2014 with glosses and spoken sentence translations, including 1,085 glosses and 7,096 training, 519 validation, and 642 test samples.
- **CSL-Daily.** A large-scale Chinese sign language dataset with 2,000 common glosses, consisting of 18,401 training, 1,077 validation, and 1,176 test samples.

Following standard practice in previous work, we use Word Error Rate (WER) as the evaluation metric for SLR, and adopt BLEU (Papineni et al. 2002) and ROUGE-L (Lin 2004) scores for evaluating SLT performance.

Method	PHOENIX14	PHOENIX14T	CSL-Daily
	DEV / TEST	DEV / TEST	DEV / TEST
RGB-based			
VAC (Min et al. 2021)	21.2 / 22.3	– / –	33.3 / 32.6
SMKD (Hao, Min, and Chen 2021)	20.8 / 21.0	20.8 / 22.4	28.4 / 27.5
CMA (Pu et al. 2020)	21.3 / 21.9	– / –	– / –
STMC (Zhou et al. 2021b)	21.1 / 20.7	19.6 / 21.0	– / –
MMTLB (Chen et al. 2022a)	– / –	21.9 / 22.5	– / –
C^2SLR (Zuo and Mak 2022)	20.5 / 20.4	20.2 / 20.4	– / –
TLP (Hu et al. 2022)	19.7 / 20.8	19.4 / 21.2	– / –
CorrNet (Hu et al. 2023b)	18.8 / 19.4	18.9 / 20.5	30.6 / 30.1
TwoStream-SLR (Chen et al. 2022b)	**18.4 / 18.8**	**17.7 / 19.3**	**25.4 / 25.3**
SignBERT+ (Hu et al. 2023a)	19.9 / 20.0	18.8 / 19.9	– / –
CTCA (Guo et al. 2023)	19.5 / 20.1	19.3 / 20.3	31.3 / 29.4
CVT-SLR (Zheng et al. 2023)	19.8 / 20.1	19.4 / 20.3	– / –
mLTSF+GFE (Xie et al. 2023)	22.9 / 23.0	– / –	– / –
AdaBrowse+ (Hu et al. 2023c)	19.6 / 20.7	19.5 / 20.6	31.2 / 30.7
Skeleton-based			
TwoStream-SLR (Chen et al. 2022b)	28.6 / 28.0	27.1 / 27.2	34.6 / 34.1
SignBERT+ (Hu et al. 2023a)	34.0 / 34.1	32.9 / 33.6	– / –
CoSign-1s (Jiao et al. 2023)	20.9 / 21.2	20.4 / 20.6	29.5 / 29.1
CoSign-2s (Jiao et al. 2023)	19.7 / 20.1	19.5 / 20.1	28.1 / 27.2
MSKA (Guan et al. 2025)	20.5 / 21.2	19.6 / 19.8	27.5 / 27.1
HyperSign (Ours)	**18.2 / 18.8**	**18.6 / 19.2**	**26.1 / 25.7**
Improvement	+1.5 / +1.3	+0.9 / +0.6	+1.4 / +1.4

Table 1: Word Error Rate (WER %, lower is better) on PHOENIX-2014, PHOENIX-2014T, and CSL-Daily. Best results are in bold.

Implementation Details

Our architecture comprises four sequential CGP modules with output channels of 64, 128, 128, and 256, followed by a single MPHF module. Temporal resolution is reduced to one-fourth of the input length for efficiency. For the SLR task, we train the model for 100 epochs using the Adam optimizer with an initial learning rate of 1×10^{-3}, a batch size of 4, and a beam width of 5 during decoding. For the SLT task, we initialize the translation module with the pretrained mBART-large-cc25 model and use a lightweight MLP to project skeletal features into its input embedding space. The MLP is trained with a learning rate of 1×10^{-3}, while the skeleton encoder and translation module are fine-tuned with 1×10^{-5}. Training runs for 100 epochs under the same settings as SLR. All experiments are conducted on a single NVIDIA RTX A4000 GPU using the PyTorch 2.6.0 framework under the Ubuntu 20.04.6 operating system.

Comparison with State-of-the-arts

For the SLR task, results in Table 1 show that HyperSign consistently performs strongly across all datasets and splits. On PHOENIX-2014, HyperSign achieves WERs of 18.2% and 18.8% on DEV and TEST, reducing WER by 0.2% compared to TwoStream-SLR on DEV, and by 1.5%/1.3% compared to CoSign-2s on DEV/TEST. On PHOENIX-2014T, it obtains 18.6% and 19.2% on DEV and TEST. While 0.9% higher than TwoStream-SLR on DEV, it reduces WER by 0.1% on TEST. Compared to CoSign-2s and MSKA, it reduces WER by 0.9%/0.9% and 1.0%/0.6% on DEV/TEST.

For the SLT task, we adopt ROUGE and BLEU4 as evaluation metrics. As shown in Table 2, 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. Compared with RGB-based models, HyperSign outperforms TwoStream-SLT in BLEU4 (29.35 vs. 28.95) on the TEST set and achieves a ROUGE score comparable to CV-SLT (54.26 vs. 54.54), demonstrating competitive translation quality without relying on RGB inputs. On the CSL-Daily dataset, HyperSign achieves ROUGE/BLEU4 scores of 55.21/26.39 on the DEV set and 56.14/26.87 on the TEST set, exceeding MSKA by +1.18/+1.23 and +1.07/+1.35.

Furthermore, as shown in Table 3, HyperSign exhibits significantly better efficiency than existing approaches, achieving the lowest FLOPs and parameter count while delivering the fastest inference speed.

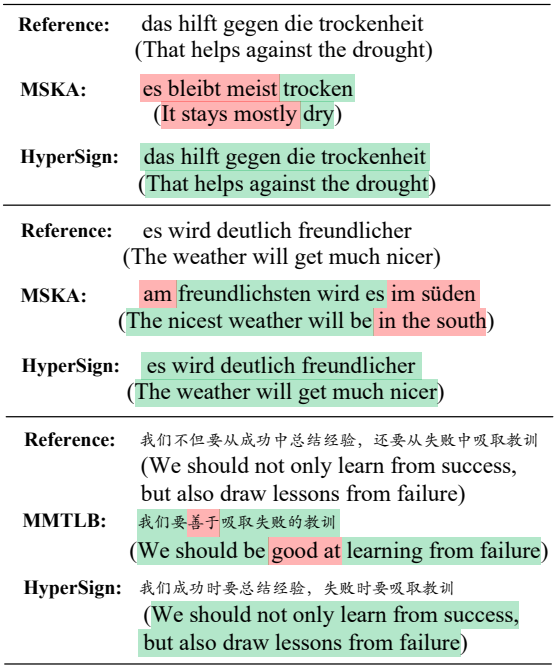

Figure 4: Qualitative translation results on PHOENIX-2014T and CSL-Daily datasets.

Ablation Study

As shown in Table 4, we conduct a module-wise ablation study on the PHOENIX-2014T dataset for the SLR task. The baseline model, composed of four spatiotemporal graph convolution layers, achieves a WER of 27.3%/28.1%. Introducing the CGP module significantly improves performance to 20.8%/21.1% by modeling joint-level co-occurrence.

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

Method	Params (M) ↓	FLOPs (G) ↓	FPS ↑
SMKD	31.6	183.2	913
TwoStream	104.8	312.7	318
CoSign-2s	30.1	28.2	–
MSKA	44.1	30.1	1224
Ours	**28.4**	**27.3**	**1397**

Table 3: Model complexity and inference speed on PHOENIX-2014T.

Adding the MPHF module further reduces the WER to 18.9%/19.7% through cross-part semantic fusion. Finally, incorporating the UACD module leads to the best result of 18.6%/19.2%. These results confirm the complementary benefits of each component and validate the overall design of our hierarchical semantic modeling strategy. We also conduct a joint ablation study of the hyperparameters K and P in the CGP module on the PHOENIX-2014T dataset, with results presented in Table 5. Here, K denotes the number of nearest neighbors in the dynamic geometric graph, and P represents the number of semantic part prototypes. The results reveal that performance is generally sensitive to both parameters. The best WER of 18.6%/19.2% is achieved when $K = 8$ and $P = 8$. When either K or P deviates from this optimal setting, performance tends to degrade slightly.

Baseline	CGP	MPHF	UACD	DEV/TEST
✓	✗	✗	✗	27.3 / 28.1
✓	✓	✗	✗	20.8 / 21.1
✓	✓	✓	✗	18.9 / 19.7
✓	✓	✓	✓	**18.6 / 19.2**

Table 4: Module-wise ablation study on PHOENIX-2014T for SLR (WER, %).

K \ P	7	8	9
7	18.7 / 19.4	18.7 / 19.3	18.8 / 19.5
8	**18.6** / 19.5	**18.6 / 19.2**	18.7 / 20.0
9	18.9 / 19.4	19.0 / 20.1	18.9 / 19.7

Table 5: Joint ablation of hyperparameters K and P in the CGP module on PHOENIX-2014T for SLR. Results are reported as WER (%) in the format of DEV/TEST.

Qualitative Analysis

Fig. 4 illustrates qualitative comparisons between model outputs and reference texts. Green highlights indicate correct translations, while red highlights mark incorrect ones. HyperSign consistently aligns well with the reference in both semantics and structure, while MSKA and MMTLB frequently exhibit semantic drift or omit critical information. For example, MSKA introduces irrelevant details (e.g., "in the south"), and MMTLB, in the Chinese example, omits the contrastive structure and shows semantic confusion.

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. Experimental results show that HyperSign outperforms existing skeleton-based approaches in both accuracy and efficiency across multiple benchmarks.

Acknowledgments

We sincerely thank all reviewers for their insightful feedback and for the valuable time and effort devoted to our work. This work was supported by the National Natural Science Foundation of China under Grant 62202201.

References

Camgoz, N. C.; Hadfield, S.; Koller, O.; Ney, H.; and Bowden, R. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7784–7793.

Chen, Y.; Wei, F.; Sun, X.; Wu, Z.; and Lin, S. 2022a. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5120–5130.

Chen, Y.; Zuo, R.; Wei, F.; Wu, Y.; Liu, S.; and Mak, B. 2022b. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35: 17043–17056.

Feng, Y.; Huang, J.; Du, S.; Ying, S.; Yong, J.-H.; Li, Y.; Ding, G.; Ji, R.; and Gao, Y. 2024. Hyper-yolo: When visual object detection meets hypergraph computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3558–3565.

Gao, Y.; Feng, Y.; Ji, S.; and Ji, R. 2022. HGNN+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3181–3199.

Gao, Y.; Zhang, Z.; Lin, H.; Zhao, X.; Du, S.; and Zou, C. 2020. Hypergraph learning: Methods and practices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2548–2566.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369–376.

Guan, M.; Wang, Y.; Ma, G.; Liu, J.; and Sun, M. 2025. MSKA: Multi-stream keypoint attention network for sign language recognition and translation. *Pattern Recognition*, 165: 111602.

Guo, L.; Xue, W.; Guo, Q.; Liu, B.; Zhang, K.; Yuan, T.; and Chen, S. 2023. Distilling cross-temporal contexts for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10771–10780.

Guo, Q.; Wang, Y.; Zhang, Y.; Qi, H.; Hu, Y.; and Jiang, Y. 2026. Hyper-BTS: Brain tumor segmentation based on hypergraph guidance. *Pattern Recognition*, 169: 111926.

Hao, A.; Min, Y.; and Chen, X. 2021. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11303–11312.

Hu, H.; Zhao, W.; Zhou, W.; and Li, H. 2023a. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 11221–11239.

Hu, L.; Gao, L.; Liu, Z.; and Feng, W. 2022. Temporal lift pooling for continuous sign language recognition. In *European conference on computer vision*, 511–527. Springer.

Hu, L.; Gao, L.; Liu, Z.; and Feng, W. 2023b. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2529–2539.

Hu, L.; Gao, L.; Liu, Z.; Pun, C.-M.; and Feng, W. 2023c. Adabrowse: Adaptive video browser for efficient continuous sign language recognition. In *Proceedings of the 31st ACM international conference on multimedia*, 709–718.

Jiang, Y.; Wang, Y.; Li, S.; Zhang, Y.; Guo, Q.; Chu, Q.; and Gao, Y. 2024. Evcslr: Event-guided continuous sign language recognition and benchmark. *IEEE Transactions on Multimedia*.

Jiao, P.; Min, Y.; Li, Y.; Wang, X.; Lei, L.; and Chen, X. 2023. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 20676–20686.

Koller, O.; Forster, J.; and Ney, H. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141: 108–125.

Lei, M.; Wu, Y.; Li, S.; Zheng, X.; Wang, J.; Gao, Y.; and Du, S. 2025. Softhgnn: Soft hypergraph neural networks for general visual recognition. *arXiv preprint arXiv:2505.15325*.

Li, Z.; Zhou, W.; Zhao, W.; Wu, K.; Hu, H.; and Li, H. 2025. Uni-Sign: Toward Unified Sign Language Understanding at Scale. *arXiv preprint arXiv:2501.15187*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, K.; Wang, X.; Zhu, L.; Zhang, B.; and Yang, Y. 2024. SKIM: Skeleton-based isolated sign language recognition with part mixing. *IEEE Transactions on Multimedia*, 26: 4271–4280.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742.

Min, Y.; Hao, A.; Chai, X.; and Chen, X. 2021. Visual alignment constraint for continuous sign language recognition. In *proceedings of the IEEE/CVF international conference on computer vision*, 11542–11551.

Niu, Z.; and Mak, B. 2020. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *European conference on computer vision*, 172–186. Springer.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pu, J.; Zhou, W.; Hu, H.; and Li, H. 2020. Boosting continuous sign language recognition via cross modality augmentation. In *Proceedings of the 28th ACM international conference on multimedia*, 1497–1505.

Pu, M.; Lim, M. K.; and Chong, C. Y. 2024. Siformer: Feature-isolated Transformer for Efficient Skeleton-based Sign Language Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9387–9396.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Xie, P.; Cui, Z.; Du, Y.; Zhao, M.; Cui, J.; Wang, B.; and Hu, X. 2023. Multi-scale local-temporal similarity fusion for continuous sign language recognition. *Pattern Recognition*, 136: 109233.

Yao, H.; Zhou, W.; Feng, H.; Hu, H.; Zhou, H.; and Li, H. 2023. Sign language translation with iterative prototype. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15592–15601.

Yin, A.; Zhong, T.; Tang, L.; Jin, W.; Jin, T.; and Zhao, Z. 2023. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2551–2562.

Zhang, T.; Liu, P.; Lu, Y.; Cai, M.; Zhang, Z.; Zhang, Z.; and Zhou, Q. 2025. Cwnet: Causal wavelet network for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8789–8799.

Zhao, R.; Zhang, L.; Fu, B.; Hu, C.; Su, J.; and Chen, Y. 2024. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *Proceedings of the aaai conference on artificial intelligence*, volume 38, 19643–19651.

Zheng, J.; Wang, Y.; Tan, C.; Li, S.; Wang, G.; Xia, J.; Chen, Y.; and Li, S. Z. 2023. Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23141–23150.

Zhou, B.; Chen, Z.; Clapés, A.; Wan, J.; Liang, Y.; Escalera, S.; Lei, Z.; and Zhang, D. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20871–20881.

Zhou, D.; Huang, J.; and Schölkopf, B. 2006. Learning with hypergraphs: Clustering, classification, and embedding. *Advances in neural information processing systems*, 19: 1601–1608.

Zhou, H.; Tian, Z.; Han, X.; Du, S.; and Gao, Y. 2024. ccRCC metastasis prediction via exploring high-order correlations on multiple WSIs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 145–154. Springer.

Zhou, H.; Zhou, W.; Qi, W.; Pu, J.; and Li, H. 2021a. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1316–1325.

Zhou, H.; Zhou, W.; Zhou, Y.; and Li, H. 2021b. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24: 768–779.

Zuo, R.; and Mak, B. 2022. C2slr: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5131–5140.