2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Learning View-Invariant Sign Language Representations via Dual-Stream Contrastive Learning

Yuting Peng[1], Yuecong Min[1], Xilin Chen[1]

***Abstract*— Viewpoint shifts significantly change how gestures and facial expressions appear and frequently cause occlusions, posing a critical challenge for robust Sign Language Recognition (SLR). To address this challenge, we exploit the spatial flexibility and computational efficiency of skeleton data and propose ViSL, a dual-stream contrastive learning framework to learn View-invariant representations for Sign Language understanding. Specifically, the primary and lifting streams share a common visual feature extractor with different types of input: the primary stream (P-Stream) directly processes frontal-view skeleton data, and the lifting stream (L-Stream) synthesizes skeleton data from arbitrary viewpoints based on 3D estimations. We further propose a view-invariant contrastive loss to align representations across both viewpoints and streams. Experimental results on the challenging cross-view setting of MM-WLAuslan demonstrate that ViSL achieves substantial performance improvements, highlighting its potential for robust real-world SLR applications.**

I. INTRODUCTION

Sign language, expressed through hand gestures, body movements and facial expressions, is a primary mode of communication within the Deaf community. Vision-based Sign Language Recognition (SLR) aims to interpret the semantic meaning of these expressions in a non-intrusive manner and has developed rapidly in the last decade [1]. However, current SLR systems often struggle to generalize beyond controlled experimental settings [46]. Challenges such as viewpoint variability, complex backgrounds, and the demand for real-time performance continue to limit their effectiveness in everyday scenarios. Therefore, it is essential to design a SLR method that can robustly capture the spatiotemporal dynamics of sign language. Beyond recognition alone, integrating such methods with robotic technologies can overcome the limitations of static cameras by leveraging mobility, adaptive perception, and multimodal sensing, which opens new possibilities for natural and intuitive human-robot interaction. Such advancements not only enhance assistive technologies for the Deaf community but also broaden the scope of human-robot interaction in service, education, and social contexts.

Unlike spoken language, sign language conveys information through face-to-face visual communication, which indicates that SLR systems cannot always access data from ideal viewpoints. However, most widely used datasets are collected primarily from the frontal view, largely overlooking the impact of viewpoint variations encountered in real-world scenarios. Recent works [9], [46] have reported substantial performance drops for state-of-the-art methods under cross-view conditions, as they struggle with occlusions, motion blur, background clutter, and other forms of visual noise, thereby limiting their robustness and generalization beyond controlled environments. Some studies [14], [38], [46], [53], [55] attempt to improve the robustness to viewpoint variability by synthetic view augmentation and multi-modal fusion, but a substantial performance gap remains between frontal and non-frontal inferences, underscoring the pressing need for more robust, viewpoint-invariant approaches.

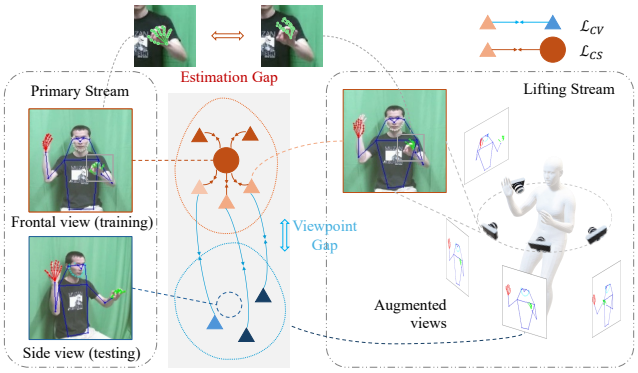

Fig. 1. Illustration of the proposed method. The dual-stream framework consists of a primary stream containing accurate 2D keypoints and a lifting stream containing noisy 3D keypoints with diverse augmented views. We employ a cross-view contrastive loss ($\mathcal{L}_{CV}$) and a cross-stream contrastive loss ($\mathcal{L}_{CS}$) to reduce both viewpoint and estimation gaps.

In this paper, we propose a dual-stream contrastive learning framework to achieve view-invariant representation for sign language recognition. As illustrated in Fig. 1, we adopt keypoint sequences as input, leveraging their spatial flexibility and computational efficiency, and further exploit different types of pose estimators. Specifically, the primary stream operates on 2D estimations, which provide higher accuracy but fail to capture depth cues in sign videos. The lifting stream employs 3D estimations to model viewpoint variations, effectively narrowing the gap between seen and unseen views while synthesizing inaccurate 2D samples.

To better exploit the complementary information in dual streams, we share the visual feature extracted between streams, and propose a view-invariant contrastive loss to

This work is partially supported by the National Natural Science Foundation of China (62506353, 62461160331, U24A20332), and the Postdoctoral Fellowship Program of CPSF under Grant Number GZB20240762.

[1]The authors are with State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China, and with University of Chinese Academy of Sciences, Beijing, 100049, China. { pengyuting24s, minyuecong, xlchen}@ict.ac.cn

align cross-view and cross-stream representations. Considering the noisy level of different pairs, we adopt asymmetric constraints for intra- and inter-stream pairs respectively. To achieve more stable optimization, we introduce learnable proxies to explicitly align cross-view, cross-stream pairs. Experimental results on the challenging cross-view setting of large-scale sign language dataset verify the effectiveness of the proposed framework, which achieves significant improvements while keeping inference lightweight by maintaining only the P-Stream. Besides, the consistent behavior under various viewpoints reveals the potential of the proposed method in real-world applications, which can be further combined with the mobility of robotic technology to achieve more flexible human-robot interaction. In summary, our contributions are summarized as follows:

1) Exploring the usage of different pose estimators in sign language recognition, and demonstrating the effectiveness of leveraging 3D estimations to enhance view-invariant representation.
2) Proposing a view-invariant contrastive loss to bridge the gap between different estimators and viewpoints in a non-fusion way, which can improve performance while preserving the lightweight efficiency of single-stream inference.
3) Designing a dual-stream contrastive learning framework that effectively integrates accurate 2D localization with complementary 3D depth cues to learn view-invariant representations, achieving significant performance improvements and highlighting its potential for real-world applications.

II. RELATED WORK

A. View-invariant Sign Language Recognition

Sign language recognition, which aims to interpret the semantic meaning of sign language gestures, serves as a fundamental task in sign language understanding. Existing SLR approaches often leverage the human pose priors [19], [21], [65], [66], language priors [25], [56], [62], [67], and monotonous alignment between input and label sequences [4], [8], [15], [36] to guide the learning of discriminative representation from fine-grained video data. Although these methods achieve remarkable progress in both accuracy and efficiency, Shen et al. [46] demonstrate that state-of-the-art SLR models suffer severe performance degradation under varying viewpoints. Compared to face or action recognition [45], [50], [54], most current SLR datasets are recorded from the front view, and recognizing sign language from side views is particularly challenging due to the fine-grained nature of the gestures, which often leads to significant occlusion issues.

Recent works [14], [38], [46], [48], [53], [55] have explored the cross-view SLR setting, where models are evaluated on viewpoints not encountered during training. For example, [14] introduced multi-view training with a view-fusion strategy to enhance representation learning, while [38] leveraged an off-the-shelf 3D whole-body mesh recovery model to generate synthetic multi-view data from frontal-view input. However, these methods adopt simple augmentation or ensemble methods without considering the unique challenges met in sign language. To address these limitations, we propose ViSL, which not only learns more discriminative view-invariant features but also reduces the impact caused by inaccurate estimation.

B. Multi-view Contrastive Learning

Contrastive loss is a classical training objective [6] in deep metric learning that operates on pairs of inputs, minimizing the distance between positive pairs while maximizing the distance between negative pairs. Its extended variants have been widely used in both supervised [43], [51] and unsupervised learning [16], [37], which also show impressive improvement in visual representation learning. For instance, several successful attempts [5], [13], [17] in self-supervised learning demonstrate that simple contrastive frameworks can improve various downstream tasks. Building on this idea, [23] extends the self-supervised contrastive approach to the fully-supervised setting and outperforms cross-entropy-based methods. Additionally, some works [31], [58] explore proxy-based strategies to mitigate optimization challenges in contrastive learning, showing promising effectiveness in domain generalization.

Due to the success of contrastive learning, several works leverage it to learn view-invariant representation. [2], [28], [35], [44], [60] address multi-view action recognition by treating different viewpoints of the same action as positive pairs and pulling them closer together. [12] introduces additional contrast among temporally augmented features. Beyond action-level contrast, [49] employs a view-level contrastive objective to group different samples from the same viewpoint, thereby disentangling semantics from viewpoint and yielding more discriminative representations. Similar to this paradigm, [48] proposes to address view-invariant SLR by utilizing a contrastive multi-task learning paradigm that disentangles view from sign semantics. Different from these methods, we additionally notice the noise introduced by 3D estimation and propose a cross-stream contrastive loss to leverage the complementary strengths of different estimates in spatial localization and depth perception.

C. Robot-assisted Human-Centered Tasks

With the rapid development of robotic technology, diverse human-centered tasks have applied perceptive and adaptive robots, including intention prediction [22], [24], [33], trajectory prediction [41], [52], and target tracking [18], [29]. These works point to promising directions for deeper human-computer interaction, where robots dynamically adapt their actions to situational contexts to enhance both performance and user experience. To foster inclusive communication between the Deaf community and the wider society, previous research [11], [26], [30], [34], [39], [40] has explored robot-assisted sign language recognition (SLR). These systems typically rely on camera-equipped devices to capture sign

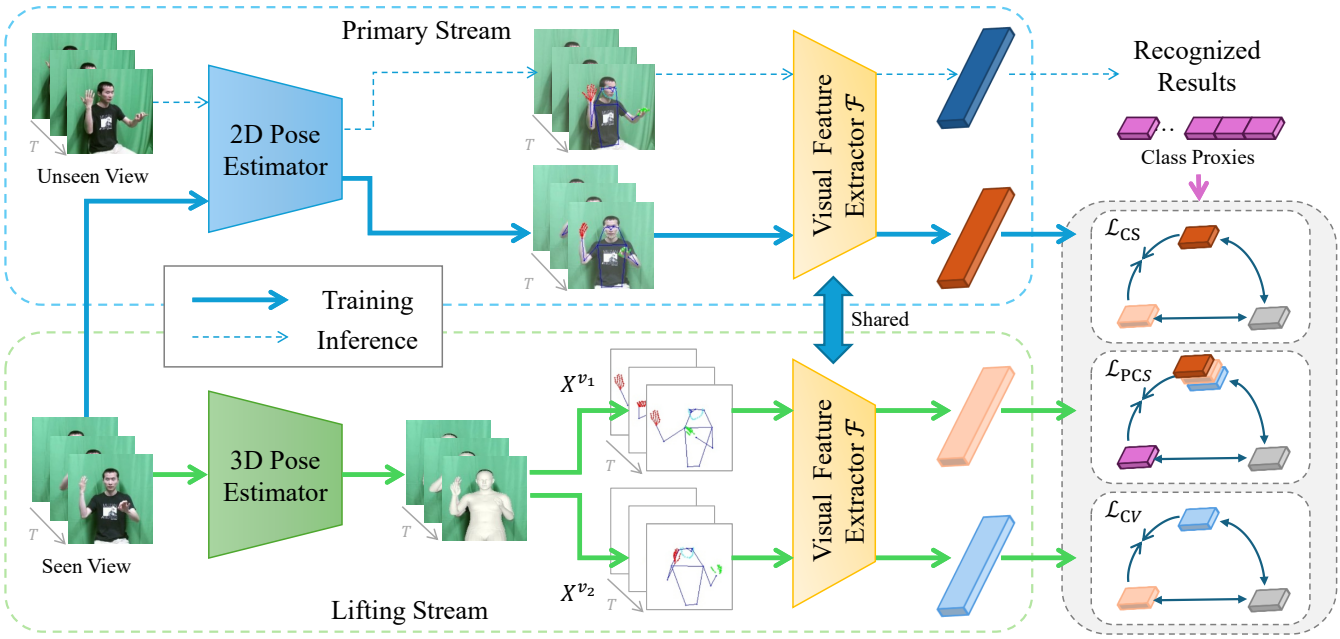

Fig. 2. During training, both 2D and 3D keypoints are estimated as the input for primary and lifting stream. The estimated 3D mesh is projected to diverse viewpoints to generate synthetic 2D keypoints. Keypoint sequences are processed by a shared visual feature extractor with cross-view $\mathcal{L}_{\text{CV}}$, cross-stream $\mathcal{L}_{\text{CS}}$ and proxy-based $\mathcal{L}_{\text{PCS}}$ contrastive learning to learn view-invariant representations. Notably, only the primary stream is used during inference.

language videos, coupled with algorithms for sign recognition. For instance, [40] deployed SLR techniques on the humanoid robot Pepper to recognize signs and display the results on its screen. Similarly, [26] integrated large language models to enable Pepper to generate co-speech gestures. While these approaches enhance convenience for people with hearing and speech impairments, current systems cannot adapt to the signer's position, requiring a direct-facing orientation. This limits the flexibility of the SLR system in dynamic environments and diminishes usability. By integrating our method with robotic mobility, we enable more flexible and seamless human-robot interaction.

III. METHOD

In this section, we first formulate the problem and demonstrate the baseline in III-A. Then we propose the cross-view contrastive learning, which helps model learning view-robust representations in III-B. After that, we discuss the necessity and challenge of leveraging both the depth information from 3D estimation and the accurate localization of 2D estimation, and present the proxy-assisted cross-stream contrastive learning in III-C. An overview of the proposed method is presented in Fig. 2.

A. Preliminary

As shown in Fig. 2, we adopt both 2D and 3D estimated keypoints as input, considering their complementary strengths in spatial localization and depth perception. For a given view-specific input data $\mathbf{X}^v$ recorded from viewpoint v, we employ a modified version of ST-GCN $\mathcal{F}$ [10] with a learnable adjacency matrix $\mathcal{A}$:

$$\mathbf{Z} = \mathcal{F}(\mathbf{X}^v; \mathcal{A}), \tag{1}$$

where the input keypoints can be either 2D or 3D ($\mathbf{X}^v \in \mathbb{R}^{\{2D,3D\}}$), The objective of cross-view sign language recognition is to learn a view-invariant representation that generalizes across a range of viewpoints $v \in [-\alpha, \alpha]$, including those not present in the training set. As illustrated in Fig. 2, the adopted baseline method synthesizes input data for unseen views based on 3D estimations as previous work does [38], and is supervised with a cross-entropy loss $\mathcal{L}_{\text{CE}}$.

B. Cross-view Contrastive Learning

Since most available sign language data is collected from the frontal view, we focus on a setting where training data is restricted to the frontal view, while evaluation data includes varying viewpoints. To improve generalization to unseen viewpoints, we first generate synthetic 2D keypoints with diverse viewpoints, and then leverage a contrastive loss to learn more discriminative yet view-invariant representation.

As shown in Fig. 2, we first reconstruct 3D human meshes from sign language videos, and subsequently project them onto multiple virtual cameras to obtain 2D keypoints with diverse viewpoints. Following the standard contrastive learning framework [5], we generate two different synthetic views $(X_i^{v_1}, X_i^{v_2})$ for each input sample X_i, and adopt a supervised contrastive loss [23] to provide supervision to enforce view-invariant representation:

$$\mathcal{L}_{\text{CV}}(i) = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(s_{i,p}/\tau\right)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp\left(s_{i,k}/\tau\right)}, \tag{2}$$

where N denotes the batch size, $P(i)$ is the set of synthetic samples sharing the same label as X_i within the batch, $s_{i,p}$ calculates the cosine similarity between $\mathcal{Z}_i$ and $\mathcal{Z}_p$, and τ

is the temperature hyperparameter controlling the sharpness of the distribution. The utilization of supervised contrastive loss enables multi-positive contrast, leading to more stable optimization and faster convergence.

C. Cross-stream Contrastive Learning

For skeleton-based sign language recognition, keypoint sequences can be obtained using either 2D or 3D estimators, such as DWPose [57], Mediapipe [32], and OSX [27]. While 2D estimators offer high localization accuracy, 3D estimators capture richer spatial relationships, which can be used to generate 2D keypoints from unseen viewpoints. However, due to the high annotation cost and the inherently ill-posed nature of 3D reconstruction, both estimated 3D keypoints and synthetic 2D keypoints are typically less accurate than direct 2D estimates. This discrepancy motivates an explicit alignment mechanism that can jointly leverage the richer depth context from 3D estimation and the higher localization fidelity of 2D estimation.

To address this challenge, we design a dual-stream framework that leverages both 2D and 3D inputs and introduce a cross-stream contrastive learning strategy to enhance view-invariance with the help of 3D information. As shown in Fig. 2, the primary stream takes 2D keypoints as input, providing precise localization but lacking depth information. In contrast, the lifting stream uses 3D human mesh to synthesize 2D keypoints as described in the previous section, offering auxiliary 3D context at the cost of lower accuracy. By jointly training the dual streams with a shared visual feature extractor, the proposed framework can enhance the 3D awareness of the primary stream while maintaining its localization accuracy with proper supervision.

Besides the noise inherent in 3D estimation, 2D and 3D estimators use heterogeneous keypoint formats [42], inducing a domain gap that makes implicit alignment challenging. To explicitly align the 2D and 3D feature spaces and narrow the gap between different annotation formats, we further propose a cross-stream contrastive loss. Since 3D lifting is noisy, the synthesized 2D keypoints may not coincide with directly estimated 2D keypoints, yielding false positive/negative pairs, necessitating a noise-robust contrastive loss. We therefore adopt RINCE [7] that performs robustly to noisy pairs to model the cross-stream relationships. We use its supervised variant similar to view-invariant loss, formulated as:

$$\mathcal{L}_{\mathrm{CS}}(i) = -\frac{1}{\eta_1 \cdot |Q(i)|} \sum_{q \in Q(i)} \Big(\exp\left(\tilde{s}_{i,q}/\tau\right) - \left(\lambda_1 \sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp\left(\tilde{s}_{i,k}/\tau\right)\right)^{\eta_1} \Big), \tag{3}$$

where $Q(i)$ is the set of samples from the other stream that share the same label as X_i within the batch, $\tilde{s}_{i,q}$ measures the cosine similarity between $\mathcal{Z}_i^{(2D)}$ and $\mathcal{Z}_q^{(3D \to 2D)}$, the hyperparameter λ controls the contribution of negative pairs, and $\eta \in (0, 1]$ balances the exploitation and exploration. A larger η places more weights on easy positive pairs and performs more robustly to noise [7]. To further bridge the estimation gap and enforce view-invariant representation, we additionally leverage a proxy-based contrastive loss and align both 2D and 3D samples to their corresponding learnable class proxies:

$$\mathcal{L}_{\mathrm{PCS}}(i) = -\frac{1}{\eta_2 \cdot |C(i)|} \sum_{c \in C(i)} \Big(\exp\left(\hat{s}_{y_i,c}/\tau\right) - \left(\lambda_2 \sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp\left(\hat{s}_{y_i,k}/\tau\right)\right)^{\eta_2} \Big), \tag{4}$$

where $C(i)$ is the set of samples that share the same label as X_i within the batch, $\hat{s}_{i,q}$ measures the cosine similarity between $\mathcal{Z}_i$ and its corresponding learnable proxy.

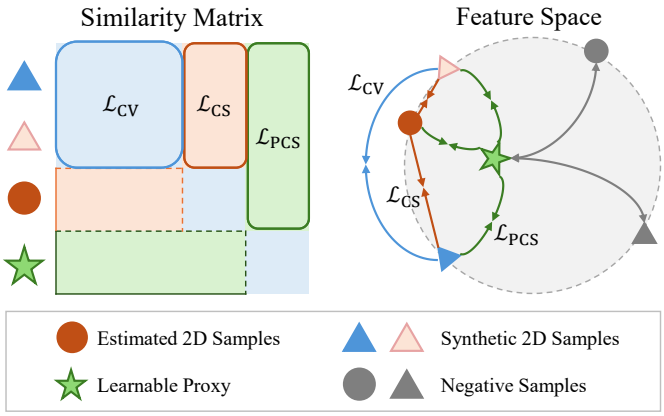

Fig. 3. Illustration of the proposed contrastive learning method. Synthetic, estimated, and proxy features are concatenated into a single batch, and contrastive losses are computed from the corresponding regions of the similarity matrix.

D. View-invariant Contrastive Learning

As mentioned above, we propose three kinds of contrastive losses to improve the view-invariant representation with different contrastive pairs. As shown in Fig. 3, the representation used in contrastive learning is composed of $2N$ synthetic 2D samples from the L-Stream, N estimated 2D samples from the P-Stream, and N learnable proxies, resulting in a $4N \times 4N$ similarity matrix. The final view-invariant contrastive objective is formulated as follows:

$$\mathcal{L}_{VI} = \mathcal{L}_{CV} + \mathcal{L}_{CS} + \mathcal{L}_{PCS}. \tag{5}$$

It is worth noting that both the lifting stream and the learnable proxies are employed to enforce view-invariant representations in a non-fusion manner, meaning that only the P-Stream is retained during inference. Consequently, the proposed method introduces no additional computational overhead at the inference stage.

IV. EXPERIMENT

A. Experimental setup

Datasets. Our experiments are conducted on MM-WLAuslan [46], a large-scale multi-view multi-modal Australian sign language (Auslan) recognition dataset, which

TABLE I

COMPARISON WITH SOTA METHODS (TOP-1, %) ON MM-WLAUSLAN. THE ENTRIES DENOTED BY '†' ARE IMPLEMENTED BY [46], THE ENTRIES DENOTED BY '‡' ARE IMPLEMENTED BY [48], '∗' DENOTES FUSION WITH ADDITION DEPTH MODALITY.

Method	Modality			Test
	RGB	Skeleton	Fusion	
Methods on the CV-ISLR Challenge [47]				
gkdx2 [53]	✓			20.29
WANGXINYU1 [55]	✓			25.39
VIPL_SLP [38]	✓			36.15
VIPL_SLP [38]		✓		45.99
gkdx2* [53]			✓	24.53
WANGXINYU1* [55]			✓	33.97
Tonicemerald [63]			✓	40.30
VIPL_SLP [38]			✓	56.87
VIPL_SLP* [38]			✓	57.97
State-of-the-art Methods				
VKNet-V† [67]	✓			14.53
STC-SLR‡ [61]		✓		25.72
VKNet-K‡ [67]		✓		28.04
DSTA-SLR‡ [20]		✓		28.66
CMVSR [48]		✓		43.12
UMDR*† [64]			✓	22.82
NLA-SLR‡ [67]			✓	33.23
Baseline (P-Stream)		✓		45.53
Baseline (L-Stream)		✓		52.75
Ours		✓		**67.14**

TABLE II

ABLATION ON ESTIMATED AND SYNTHETIC 2D KEYPOINTS (TOP-1, %). VAL (F): FRONTAL-VIEW VALIDATION; VAL (S): AVERAGE OVER LEFT/RIGHT-FRONT VALIDATION; TEST: AVERAGE OVER ALL LEFT/RIGHT-FRONT TEST SETS.

Training set	Testing set	Val (F)	Val (S)	Test
Estimated	Estimated	92.58	50.26	45.53
Synthetic	Synthetic	86.17	24.11	22.60
Synthetic (aug)	Synthetic	86.49	57.78	52.75
Synthetic (aug)	Estimated	70.67	49.81	45.45

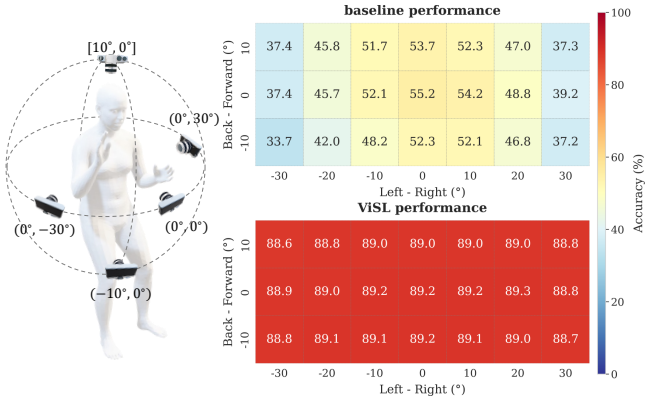

Fig. 4. Comparison between the baseline (P-Stream) and ViSL across different synthetic viewpoints. We estimate 3D keypoints from the frontal-view validation set and rotate them to the desired angles for evaluation.

contains 282K+ sign videos covering 3,215 glosses in Auslan. In the cross-view setting, where models are evaluated on the viewpoints unseen during training, only frontal-view videos are used in the training set. We use side-view validation sets (left-front and right-front) for model selection and additionally report performance on the frontal-view validation set for comparison. The test sets contain only side-view videos, and are divided into 4 subsets to manipulate real-world scenarios: in-the-wild (ITW) set, synthetic background (SYN) set, studio (STU) set and the temporal disturbance (TED) set. The split ratio of training, validation and testing set is 6:3:4.

Evaluation Metrics. We use Top-1 accuracy to evaluate the proposed method. The results on side-view validation sets are averaged for conciseness. The final test performance is computed as the average across all four subsets.

Implementation Details. We crop all sequences to a fixed length of 64 frames for both training and inference. For 3D viewpoint augmentation, elevation (up-down) and azimuth (left-right) rotation angles are sampled within $[-10°, 10°]$ and $[-35°, 35°]$ respectively. We train the model for 100 epochs on a single NVIDIA RTX 3090. AdamW optimizer is adopted, and the initial learning rate is set to 0.001, divided by 10 at the 70th and 95th epoch. The dimension of learnable proxy is 512. Both λ_1 and λ_2 are set to 0.01 in Eq. 3 and 4, and η_1 is set to 1, η_2 is set to 0.1. The P-Stream adopts 2D keypoints estimated by DWPose [57] as input, and the L-Stream adopts 3D keypoints obtained by OSX [27] as input with synthetic view augmentation.

B. Comparison with State-of-the-art Methods

We compare our method with SOTA approaches on the MM-WLAuslan dataset [46]. As shown in Tab. I, previous SOTA methods fail to maintain competitive results in the cross-view setting due to the large viewpoint variations. Moreover, methods in the CV-ISLR challenge [47] often rely on model ensembles to boost performance. The best-performing method [38] leverages both viewpoint augmentation and multi-modal fusion, achieving a Top-1 accuracy of 57.94%. Despite relying solely on the skeleton data, our method achieves 67.14%, outperforming all existing approaches by a large margin (+9.2%) and surpassing the best-performing skeleton-based method by 21.15%, demonstrating its effectiveness in handling viewpoint variations.

Compared to the state-of-the-art model trained solely on skeleton [38], our P-Stream baseline achieves comparable performance (45.53%) without viewpoint augmentation, attributable to DWPose's [57] precise keypoint localization and the strong framework. Although the L-Stream baseline utilizes noisy keypoints than the 2D counterpart, incorporating view augmentation boosts its accuracy to 52.75%. Finally, the proposed method further improves performance to 67.14% with the view-invariant contrastive loss, demonstrating the effectiveness of integrating precise localization with complementary depth information.

Fig. 4 further visualizes the effectiveness of view-invariant recognition. While the baseline model suffers a significant performance drop under large deviations from the frontal view, the proposed ViSL model consistently maintains robust performance across different viewpoints.

TABLE III

ABLATION ON THE DESIGNS OF CROSS-VIEW CONTRASTIVE LEARNING (TOP-1, %).

P-Stream	L-Stream	$\mathcal{L}_{CV}$	$\mathcal{L}_{CS}$	$\mathcal{L}_{PCS}$	Val(F)	Val(S)	Test
✓					92.58	50.26	45.53
	✓				86.49	57.78	52.75
	✓	✓			86.94	59.69	54.94
✓	✓	✓			91.63	70.62	65.31
✓	✓	✓	✓		92.02	71.44	66.13
✓	✓	✓	✓	✓	**92.64**	**71.71**	**67.14**

TABLE IV

ABLATION ON DIFFERENT VALUES OF η_1 AND η_2 ON THE VAL (S) SET (TOP-1, %).

Value	0.01	0.05	0.1	0.2	0.5	0.8	1
η_1	70.16	70.04	69.85	70.09	70.69	71.29	**71.44**
η_2	71.62	71.59	**71.71**	71.38	71.54	70.54	70.72

C. Ablation Results

Ablation on pose estimators. Tab. II shows that training and evaluating solely on synthetic 2D keypoints based on estimated mesh [27] yields substantially lower performance (24.11% on validation set) compared to using 2D keypoints estimated by DWPose (50.26%), highlighting the inferior localization accuracy of the 3D pose estimator. Interestingly, augmenting with synthetic keypoints from different viewpoints significantly boosts accuracy on the synthetic 2D keypoints (from 24.11% to 57.78%), confirming the strong benefit of injecting depth-derived rotational diversity during training. However, performance drops when testing on estimated 2D keypoints (from 57.78% to 49.81%) likely due to the mismatches in keypoint formats and distributions. Overall, these findings reveal a pronounced domain gap between 2D and 3D estimation results.

Ablation on contrastive loss design. As shown in Tab. III, incorporating $\mathcal{L}_{CV}$ can improve the performance of baseline (L-Stream) from 57.78% to 59.69%, which indicates that explicitly tightening intra-class distance and enlarging inter-class margins yields more discriminative, view-invariant features. Joint training of both streams results in a significant performance boost, demonstrating that co-learning effectively generalizes across heterogeneous keypoint formats. Further introducing $\mathcal{L}_{CS}$ and $\mathcal{L}_{PCS}$ raises performance to 71.71%, highlighting the benefit of explicitly enforcing compact class clusters and leveraging proxy centroids.

Another interesting observation from Tab. III is that frontal-view and side-view performance are largely uncorrelated. While the L-Stream substantially improves performance in the side-view setting, it suffers a drop on the frontal validation set (from 92.58% to 86.49%), likely due to noisy 3D keypoints. Combining the primary and lifting streams helps mitigate this estimation noise, and incorporating $\mathcal{L}_{CS}$ and $\mathcal{L}_{PCS}$ further enhances frontal-view performance, bringing it on par with using P-Stream only.

Ablation on the hyperparameter of loss design. We test a range of values for η_1, and fix the best η_1 to test the best η_2. As shown in Tab. IV, for η_1, the best performance is achieved when $\eta_1 = 1$. The model generally improves as the value of η_1 increases, suggesting that softer constraints are robust to noisy pairs. The optimal value for η_2 is 0.1, and a smaller η_2 performs better. Since proxy partially mitigates noisy pairs, stricter constraints encourage the model to explore by learning from hard pairs, therefore boosting the performance.

D. Qualitative Results

Impacts of viewpoint variation. Fig. 5 presents representative success and failure cases. Signs involving pronounced depth trajectories (e.g., back-and-forth arm motions) exhibit significant appearance shifts when observed from side views. The proposed ViSL model effectively infers implicit depth cues to handle these variations, enabling correct recognition even under such unseen viewpoints.

In failure cases, incorrect predictions are typically semantically related or kinematically similar to the ground-truth glosses under side views. These misclassifications suggest that ViSL relies primarily on coarse limb orientations, leading to ambiguity among signs with subtle differences. This highlights the necessity of incorporating more fine-grained features for robust cross-view SLR.

V. DISCUSSION

Recent advances in sign language recognition have largely benefited from research in computer vision and natural language processing [3], [59]. Yet, sign language is not only a rich, multi-modal form of human communication but also a structured and highly informative signal for interaction. Its spatiotemporal patterns, hand and body dynamics, and facial expressions provide dense cues that can support intention understanding, trajectory prediction, and adaptive behavior. Conversely, robotic platforms offer controlled and mobile environments for large-scale, multi-view, and multi-modal SLR data collection, enabling the study of complex interactions that are difficult to capture otherwise. However, the interaction between sign language and robotics capabilities remains largely unexplored. Integrating robust SLR into robots could enable more natural and inclusive human-robot communication. We hope that this perspective inspires more research that bridges sign language understanding and robotics, moving beyond conventional SLR benchmarks to real-world, dynamic, and interactive applications.

VI. CONCLUSIONS

Sign language recognition systems are prone to degradation under viewpoint changes, limiting their utility in assisting the Deaf community. In this study, we introduce a dual-stream contrastive learning framework that learns view-invariant representations. A cross-view contrastive learning loss promotes a more discriminative feature space, while a cross-stream contrastive learning loss fuses accurate 2D localization with complementary 3D depth cues. Leveraging class proxies can further improve cross-stream alignment. Experimental results demonstrate that ViSL can achieve state-of-the-art performance on unseen viewpoints using only

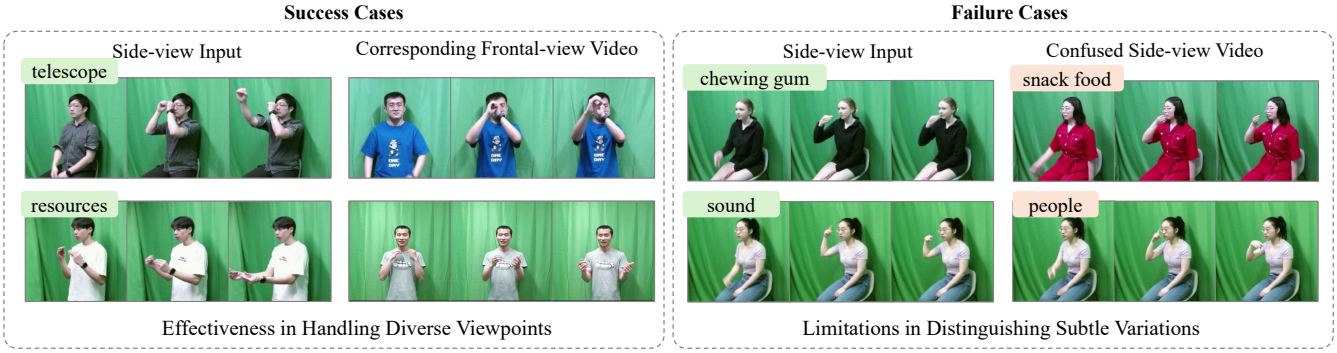

Fig. 5. Qualitative success and failure cases. Given side-view input, we show the corresponding frontal-view videos to highlight appearance differences caused by viewpoint variations for success cases. For failure cases, we show the side-view clip of the predicted gloss to illustrate the potential confusion.

skeleton modality, and the method remains lightweight at inference. Based on the proposed view-invariant framework, we further advocate introducing robotic agents into SLR systems to enable view-adaptive SLR, thereby extending the accessibility and inclusivity of advanced human-computer interactive technologies to the Deaf community.

REFERENCES

[1] Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE transactions on multimedia*, 24:1750–1762, 2021.

[2] Cunling Bian, Wei Feng, Fanbo Meng, and Song Wang. Global–local contrastive multiview representation learning for skeleton-based action recognition. *CVIU*, 229:103655, 2023.

[3] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31, 2019.

[4] Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. Aligning subtitles in sign language videos. In *ICCV*, pages 11552–11561, 2021.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PmLR, 2020.

[6] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, pages 539–546. IEEE, 2005.

[7] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *CVPR*, pages 16670–16681, 2022.

[8] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *CVPR*, pages 7361–7369, 2017.

[9] Nguyen Son Dinh, Tuan Dung Nguyen, Duc Tri Tran, Nguyen Dang Huy Pham, Thuan Hieu Tran, Ngoc Anh Tong, Quang Huy Hoang, and Phi Le Nguyen. Sign language recognition: A large-scale multi-view dataset and comprehensive evaluation. In *WACV*, pages 7887–7897. IEEE, 2025.

[10] Jeonghyeok Do and Munchurl Kim. Skateformer: skeletal-temporal transformer for human action recognition. In *ECCV*, pages 401–420. Springer, 2024.

[11] Gregory Dudek, Junaed Sattar, and Anqi Xu. A visual language for robot control and programming: A human-interface study. In *ICRA*, pages 2507–2513. IEEE, 2007.

[12] Xuehao Gao, Yang Yang, Yimeng Zhang, Maosen Li, Jin-Gang Yu, and Shaoyi Du. Efficient spatio-temporal contrastive learning for skeleton-based 3-d action recognition. *IEEE Transactions on Multimedia*, 25:405–417, 2021.

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020.

[14] Zhong Guan, Yongli Hu, Huajie Jiang, Yanfeng Sun, and Baocai Yin. Multi-view isolated sign language recognition based on cross-view and multi-level transformer. *Multimedia Systems*, 31(3):1–15, 2025.

[15] Leming Guo, Wanli Xue, Qing Guo, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Shengyong Chen. Distilling cross-temporal contexts for continuous sign language recognition. In *CVPR*, pages 10771–10780, 2023.

[16] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[18] Benjamin Hepp, Tobias Nägeli, and Otmar Hilliges. Omni-directional person tracking on a flying robot using occlusion-robust ultra-wideband signals. In *IROS*, pages 189–194. IEEE, 2016.

[19] Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE TPAMI*, 45(9):11221–11239, 2023.

[20] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Dynamic spatial-temporal aggregation for skeleton-aware sign language recognition. *arXiv preprint arXiv:2403.12519*, 2024.

[21] Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *ICCV*, pages 20676–20686, 2023.

[22] Kushal Kedia, Atiksh Bhardwaj, Prithwish Dan, and Sanjiban Choudhury. Interact: Transformer models for human intent prediction conditioned on robot actions. In *ICRA*, pages 621–628. IEEE, 2024.

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 33:18661–18673, 2020.

[24] Sang Uk Lee. Commonsense spatial knowledge-aware 3-d human motion and object interaction prediction. In *ICRA*, pages 3057–3063. IEEE, 2024.

[25] Yuhao Li, Xinyue Chen, Hongkai Li, Xiaorong Pu, Peng Jin, and Yazhou Ren. Vsnet: Focusing on the linguistic characteristics of sign language. In *CVPR*, pages 24320–24330, 2025.

[26] JongYoon Lim, Inkyu Sa, Bruce MacDonald, and Ho Seok Ahn. A sign language recognition system with pepper, lightweight-transformer, and llm. *arXiv preprint arXiv:2309.16898*, 2023.

[27] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, pages 21159–21168, 2023.

[28] Lilang Lin, Jiahang Zhang, and Jiaying Liu. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *CVPR*, pages 2363–2372, 2023.

[29] Yue Lin, Yang Liu, Pingping Zhang, Xin Chen, Dong Wang, and Huchuan Lu. Safety-first tracker: A trajectory planning framework for omnidirectional robot tracking. In *IROS*, pages 5416–5423. IEEE, 2024.

[30] Edmond Liu, Jong Yoon Lim, Vineeth Johnson, Bruce MacDonald, and Ho Seok Ahn. Signpepper: Multimodal social robot for sign language teaching. In *HRI*, pages 1791–1793. IEEE, 2025.

[31] Yimin Liu, Meibin Qi, Yongle Zhang, Qiang Wu, Jingjing Wu, and Shuo Zhuang. Improving consistency of proxy-level contrastive learning for unsupervised person re-identification. *TIFS*, 2024.

[32] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[33] Ren C Luo and Licong Mai. Human intention inference and online human hand motion prediction for human-robot collaboration. In *IROS*, pages 5958–5964. IEEE, 2019.

[34] Alex Meade. Dexter–a finger-spelling hand for the deaf-blind. In *ICRA*, volume 4, pages 1192–1195. IEEE, 1987.

[35] Qianhui Men, Edmond SL Ho, Hubert PH Shum, and Howard Leung. Focalized contrastive view-invariant learning for self-supervised skeleton-based action recognition. *Neurocomputing*, 537:198–209, 2023.

[36] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *ICCV*, pages 11542–11551, 2021.

[37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[38] Yuting Peng, Peiqi Jiao, Honggang Zou, Yuecong Min, and Xilin Chen. Synthetic view augmentation for sign language recognition. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2448–2452, 2025.

[39] Marc Peral, Alberto Sanfeliu, and Anaís Garrell. Efficient hand gesture recognition for human-robot interaction. *RA-L*, 7(4):10272–10279, 2022.

[40] Arman Sabyrov, Medet Mukushev, and Vadim Kimmelman. Towards real-time sign language interpreting robot: Evaluation of non-manual components on recognition accuracy. In *CVPRW*, 2019.

[41] Tim Salzmann, Hao-Tien Lewis Chiang, Markus Ryll, Dorsa Sadigh, Carolina Parada, and Alex Bewley. Robots that can see: Leveraging human pose for trajectory prediction. *RA-L*, 8(11):7090–7097, 2023.

[42] István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *WACV*, pages 2956–2966, 2023.

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[44] Ketul Shah, Anshul Shah, Chun Pong Lau, Celso M de Melo, and Rama Chellappa. Multi-view action recognition using contrastive learning. In *WACV*, pages 3381–3391, 2023.

[45] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016.

[46] Xin Shen, Heming Du, Hongwei Sheng, Shuyun Wang, Hui Chen, Huiqiang Chen, Zhuojie Wu, Xiaobiao Du, Jiaying Ying, Ruihan Lu, et al. Mm-wlauslan: Multi-view multi-modal word-level australian sign language recognition dataset. In *NeurIPs D&B*.

[47] Xin Shen, Heming Du, Miao Xu, Miaomiao Liu, and Xin Yu. Cross-view isolated sign language recognition challenge: Design, results and future research. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2444–2447, 2025.

[48] Xin Shen, Xinyu Wang, Lei Shen, Kaihao Zhang, and Xin Yu. Cross-view isolated sign language recognition via view synthesis and feature disentanglement. In *ICCV*, 2025.

[49] Nyle Siddiqui, Praveen Tirupattur, and Mubarak Shah. Dvanet: Disentangling view and action features for multi-view action recognition. In *AAAI*, volume 38, pages 4873–4881, 2024.

[50] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *FG*, pages 53–58. IEEE, 2002.

[51] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *NeurIPS*, 29, 2016.

[52] Li Sun, Zhi Yan, Sergi Molina Mellado, Marc Hanheide, and Tom Duckett. 3dof pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data. In *ICRA*, pages 5942–5948. IEEE, 2018.

[53] Fei Wang, Kun Li, Yiqi Nie, Zhangling Duan, Peng Zou, Zhiliang Wu, Yuwei Wang, and Yanyan Wei. Exploiting ensemble learning for cross-view isolated sign language recognition. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2453–2457, 2025.

[54] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *CVPR*, pages 2649–2656, 2014.

[55] Xinyu Wang. Cross-view isolated sign language recognition with graph. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2458–2462, 2025.

[56] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Learnt contrastive concept embeddings for sign recognition. In *ICCV*, pages 1945–1954, 2023.

[57] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, pages 4210–4220, 2023.

[58] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *CVPR*, pages 7097–7107, 2022.

[59] Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing. In *ACL*, pages 7347–7360, 2021.

[60] Long Zhao, Yuxiao Wang, Jiaping Zhao, Liangzhe Yuan, Jennifer J Sun, Florian Schroff, Hartwig Adam, Xi Peng, Dimitris Metaxas, and Ting Liu. Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization. In *CVPR*, pages 12793–12802, 2021.

[61] Weichao Zhao, Wengang Zhou, Hezhen Hu, Min Wang, and Houqiang Li. Self-supervised representation learning with spatial-temporal consistency for sign language recognition. *IEEE TIP*, 33:4188–4201, 2024.

[62] Jiangbin Zheng, Yile Wang, Cheng Tan, Siyuan Li, Ge Wang, Jun Xia, Yidong Chen, and Stan Z Li. Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. In *CVPR*, pages 23141–23150, 2023.

[63] Zhongtian Zheng. Zero-shot multi-view australian sign language recognition. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2463–2467, 2025.

[64] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, and Fan Wang. A unified multimodal de-and re-coupling framework for rgb-d motion recognition. *IEEE TPAMI*, 45(10):11428–11442, 2023.

[65] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *AAAI*, volume 34, pages 13009–13016, 2020.

[66] Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *CVPR*, pages 5131–5140, 2022.

[67] Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *CVPR*, pages 14890–14900, 2023.