Appendix for Cross-View Isolated Sign Language Recognition via View Synthesis and Feature Disentanglement

Xin Shen[1] Xinyu Wang[3] Lei Shen[4] Kaihao Zhang[2] Xin Yu[1*]
[1]The University of Queensland
[2]Australian National University [3]Zhejiang University [4]Institute of Computing Technology, CAS
xin.shen@uq.edu.au

Appendix

This appendix is organized as follows:

- Limitation and Future Work (Section A).
- Broader Impacts (Section B).
- Data Collection of the *MTV-Test* Set (Section C).
- Details of the Transformation Function (Section D).
- Visualization of the View-Semantics Disentanglement (Section E).
- Further Discussions (Section F).
- Consent Form for the Data Recording (Section G).

A. Limitation and Future Work

Inaccurate pose estimation. CMVSR leverages the 3D pose information for data augmentation and further training. However, the accuracy of estimating 3D poses from RGB data remains limited. This shows a challenge in achieving high-quality pose estimation, which can affect the overall performance of the model. Thus, improving the accuracy of 3D pose estimation is a key point. We aim to explore advanced 3D pose estimation models [5, 13] for higher-quality 3D pose data, which will be crucial to achieve better performance in the sign language recognition task.

Focus on isolated sign language recognition. CMVSR aims to solve the cross-view isolated sign language recognition task, which limits its applicable range in real-world scenarios. Isolated signs are relatively easier to classify, while continuous sign language recognition (CSLR) [1, 14] and sign language translation (SLT) [6, 8] present greater challenges. In the future, we aim to expand our approach to these tasks, which requires the model to better handle temporal transitions and the gaps between adjacent signs. By addressing these challenges, we aim to improve the practicality and robustness of sign language systems in real-world applications.

Focus on Australian sign language. CMVSR is trained on the MM-WLAuslan [7] dataset and tested on it along with the expanded MTV-Test set, where all samples belong to Auslan. Due to the lack of cross-view data in other sign language datasets, the performance of our method has not been tested on other sign languages [2, 3]. In the future, we hope to further expand the test set based on existing sign language datasets and demonstrate the effectiveness of our method in multiple sign languages.

(a) Public Speaking

(b) Daily Communication

Figure 1. **Sign language presentation from different viewpoints.** (a) Various audience-visible views in public speeches or performances. (b) Side views due to seating and group interactions in daily communication.

B. Broader Impacts

Sign language is not only a communication tool for the deaf, but also a crucial part of cultural identity [9, 11]. It enables the deaf to express and communicate effectively through unique gestures and body language, overcoming language barriers and offering equal opportunities to engage with other social groups. As a sign language recognition model, *CMVSR* is conducive to the social integration and mental well-being of the deaf community. Cross-view sign language recognition [10] overcomes the limitations of traditional sign language recognition, which relies on a single frontal viewpoint. However, as shown in Figure 1, it is im-

*Corresponding author.

Table 1. **Overview of the MTV-Test set.**

Number of Signers	36
Proportion of Gender (M / F)	15 / 21
Number of Backgrounds	30
Number of Phone Types	12
Number of Webcam Types	18

possible to ensure that everyone is directly in front of the signer in situations such as public speaking. Similarly, in daily communication, it is common that people view signers from different viewpoints. By learning from multi-view data, CMVSR can solve the problem of misinterpretation caused by different capture perspectives and significantly improve the accuracy of recognition. This model is more adaptable to the dynamic perspectives of interaction in real life, especially in complex social environments, ensuring clearer sign language expression and thereby enhancing communication and quality of life among the deaf community.

C. Data Collection of the *MTV-Test* Set

To enhance the diversity of test set, we introduce the MTV-Test set based on the MM-WLAuslan dataset.

Data Source. The recorded videos of the multi-view test set (*MTV-Test*) are collected through custom recordings of phones and webcams. The recording process uses 12 types of phones and 18 types of webcams. As shown in Table 1, the dataset includes 36 signers, 3,215 glosses, and 30 real-world backgrounds, ensuring the diversity and wide range of sign language gestures.

Recording Viewpoint. The recorded video data in *MTV-Test* includes four distinct shooting viewpoints: ***up***, ***down***, ***left***, and ***right***. The specific parameters of each camera position are unknown, which makes the viewpoints random and diverse, and helps capture a wider range of sign language expressions.

Recording Background. The background of *MTV-Test* set is composed of randomly selected real-world scenes. In addition, the lighting conditions for the recordings vary, which means that each signer will experience different lighting changes. This adds diversity to the sign language dataset for real-world applications.

D. Details of the Transformation Function

To synthesize multi-view representations, we apply the operation $\mathcal{T}_\theta(\cdot)$, which consists of a rotation in 3D space followed by a perspective projection onto the 2D plane. Given an initial set of 3D keypoints $P_{3D} \in \mathbb{R}^{K\times 3}$, the rotated keypoints P'_{3D} are obtained through:

$$P'_{3D} = R(\theta_y)R(\theta_x)P_{3D}, \tag{1}$$

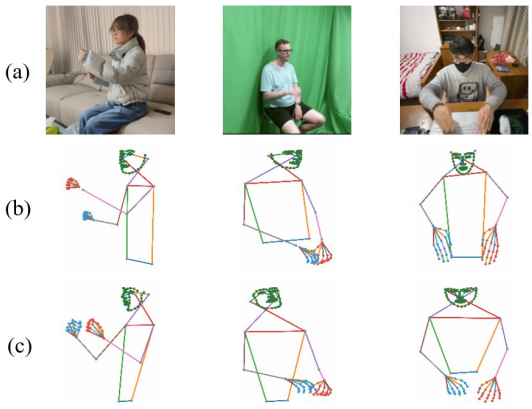


Figure 2. **Disentanglement visualization.** (a) RGB videos from the test set. (b) 2D poses corresponding to RGB videos. (c) Top-1 retrieval from the training set.

where $R(\theta_x)$ and $R(\theta_y)$ are the rotation matrices defined as:

$$R_x(\theta_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix}, \tag{2}$$

$$R_y(\theta_y) = \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix}. \tag{3}$$

Then, we apply the composite transformation:

$$R(\theta_y)R(\theta_x) = \begin{bmatrix} \cos\theta_y & \sin\theta_x\sin\theta_y & \cos\theta_x\sin\theta_y \\ 0 & \cos\theta_x & -\sin\theta_x \\ -\sin\theta_y & \sin\theta_x\cos\theta_y & \cos\theta_x\cos\theta_y \end{bmatrix}. \tag{4}$$

After rotation, the transformed keypoints P'_{3D} are projected onto the 2D plane using a standard pinhole camera model. The projection is defined as:

$$\tilde{P_{3D}} = \frac{P'_{3D}}{P'_{3D\_z}}, \tag{5}$$

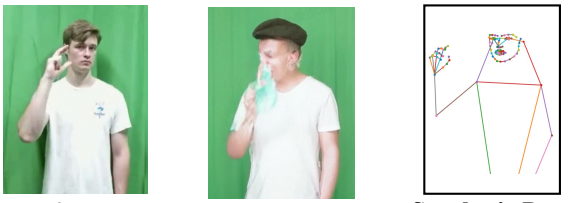
where P'_{3D_z} is the depth component. Finally, we apply the camera intrinsic matrix K and obtain the 2D coordinates P_{2D}:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \; P_{2D} = K\tilde{P_{3D}}. \tag{6}$$

E. Visualization of the View-Semantics Disentanglement

To evaluate the effectiveness of the proposed View-Semantics Disentanglement module, we visualize its disentanglement capability in Figure 2. The first row displays representative RGB video clips sampled from the test set,

Figure 3. **Cross-view synthesis comparison.** From left to right: the reference frame, the result of the RGB-based generator *CogVideoX-5B*, and that of our skeleton-based synthesis.



Reference Frame | CogVideoX -5B | **Synthetic Pose (Ours)**

Table 2. Training time and memory usage of different models.

Model	Time	Memory
KVNet-V/K [15]	48 h (A100)	23.5K MB
DSTA-SLR [4]	3.5 h (RTX 3090)	3.9K MB
CMVSR (Ours)	14.5 h (RTX 3090)	4.2K MB

and the second row shows the corresponding 2D skeletal poses. Each 2D pose is embedded into a shared latent space, from which our model disentangles a view feature T_v' and a semantic feature T_s'.

To assess disentanglement, we compute the cosine similarity between the view feature T_v' of a test sample and all view features in the training set. Then, for each test pose, we retrieve the most similar training samples based on either the view feature or the semantic feature. The retrieved results are presented in the third row of Figure 2, intuitively illustrating that T_v' primarily captures view information while T_s' captures semantic information.

F. Further Discussions

Comparisons between RGB and skeleton-based synthesis: In order to synthesize cross-view data, CV-ISLR needs to perform out-of-plane rotations on the signer's video. Existing RGB-based generative models (*i.e.*, CogVideoX [12]) struggle to follow such geometric constraints and preserve image details (as shown in Figure 3). Therefore, direct augmentation in the RGB domain may introduce large distortion in synthesized videos, thus degrading CV-ISLR performance.
Training time and memory usage: We report the training time and memory usage of all models on a single GPU, with a batch size of 8 and epochs of 50 in Table 2.

G. Consent Form for the Data Recording

Due to the inclusion of facial information in our dataset, we obtain consent from volunteers and let them sign the consent form depicted in Figure 4 before recording data. **We do not release personal identification information** such as name, age, occupation, or indication of whether an individual is deaf or hard of hearing. It is worth noting that our dataset is strictly for academic purposes only and cannot be used for commercial purposes.

References

[1] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7784–7793. Computer Vision Foundation / IEEE Computer Society, 2018.

[2] 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 *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[3] Liqing Gao, Lei Zhu, Senhua Xue, Liang Wan, Ping Li, and Wei Feng. Multi-view fusion for sign language recognition through knowledge transfer learning. In *Proceedings of the 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 1–9, 2022.

[4] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Dynamic spatial-temporal aggregation for skeleton-aware sign language recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 5450–5460. ELRA and ICCL, 2024.

[5] Wenhao Shen, Wanqi Yin, Xiaofeng Yang, Cheng Chen, Chaoyue Song, Zhongang Cai, Lei Yang, Hao Wang, and Guosheng Lin. Adhmr: Aligning diffusion-based human mesh recovery via direct preference optimization. *arXiv preprint arXiv:2505.10250*, 2025.

[6] Xin Shen, Shaozu Yuan, Hongwei Sheng, Heming Du, and Xin Yu. Auslan-daily: Australian sign language translation for daily communication and news. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[7] Xin Shen, Heming Du, Hongwei Sheng, Shuyun Wang, Hui Chen, Huiqiang Chen, Zhuojie Wu, Xiaobiao Du, Jiaying Ying, Ruihan Lu, Qingzheng Xu, and Xin Yu. Mm-wlauslan: Multi-view multi-modal word-level australian sign language recognition dataset. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[8] Xin Shen, Lei Shen, Shaozu Yuan, Heming Du, Haiyang Sun, and Xin Yu. Diverse sign language translation. *arXiv preprint arXiv:2410.19586*, 2024.

[9] Xin Shen, Heming Du, Hongwei Sheng, Lincheng Li, and Kaihao Zhang. Auslanweb: A scalable web-based australian sign language communication system for deaf and hearing individuals. In *Proceedings of the ACM on Web Conference 2025*, pages 5212–5223, 2025.

Consent Form for Recording of the Australian Sign Language Dataset

Dear Participant,

Hello! We are a team dedicated to the research of sign language. We are conducting an academic project aimed at recording and analyzing Australian Sign Language (Auslan). We invite you to participate in this project. The purpose of this project is to facilitate the learning and dissemination of sign language and to enhance understanding and application of Auslan.

Mode of Participation:
You will be recorded while using Auslan for communication. These recordings may include your facial expressions and hand gestures.

Privacy and Data Use:
We commit to using the recorded data solely for academic research purposes and not for any commercial use. All data will be anonymized to ensure the security of your personal information. The video material may be presented at academic conferences, in research papers, or educational courses.

Consent Details:

1. I have read and understood the information about the research described above.
2. I agree to participate in the video recordings of Australian Sign Language.
3. I understand that my participation is voluntary, and I can withdraw at any time without any adverse consequences.
4. I agree that my facial expressions and hand gestures may be recorded and used for academic research.

Please fill out the following information and sign below to indicate your consent to participate:

- **Name:** _____
- **Email:** _____
- **Signature:** _____
- **Date:** _____

We greatly appreciate your participation and support!

Should you have any questions or require further information, please contact us at:

Contact Person: [Name of Coordinator]
Email: [Coordinator's Email]
Phone: [Coordinator's Phone]

Figure 4. **Consent form for the data recording.**

[10] 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.

[11] Hongwei Sheng, Xin Shen, Heming Du, Hu Zhang, Zi Huang, and Xin Yu. Ai empowered auslan learning for parents of deaf children and children of deaf adults. *AI and Ethics*, pages 1–11, 2024.

[12] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. 2024.

[13] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, et al. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *arXiv preprint arXiv:2501.09782*, 2025.

[14] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1316–1325. Computer Vision Foundation / IEEE, 2021.

[15] Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14890–14900. IEEE, 2023.