Appendix for MM-WLAuslan:
Multi-View Multi-Modal Word-Level Australian Sign Language Recognition Dataset

Xin Shen   **Heming Du**   **Hongwei Sheng**   **Shuyun Wang**   **Hui Chen**[*]   **Huiqiang Chen**[*]
Zhuojie Wu   **Xiaobiao Du**[*]   **Jiaying Ying**   **Ruihan Lu**   **Qingzheng Xu**   **Xin Yu**[†]
The University of Queensland
x.shen3@uqconnect.edu.au

Figure 1: Schematic diagram of our recording studio.

A   Recording MM-WLAuslan

In this section, we introduce our recording environment, interactive recording interface and Auslan learning interface. Meanwhile, we detail the comparison between Kinect-V2 and RealSense.

A.1   Recording Studio

For recording a clean word-level Auslan dataset, as shown in Figure 1, our recording setup is situated in a studio environment with a green screen. In the studio, we position Kinect-V2 cameras at left-front, front, and right-front views, along with a centrally placed RealSense camera. The signer is positioned in the centre, surrounded by various RGB-D cameras. Directly in front of the signer, the "Front Kinect-V2" and "Front RealSense" are primarily used for recording the frontal aspects of the signing actions. From the left and right of the signer, the "Left-Front Kinect-V2" and "Right-Front Kinect-V2" are respectively placed to capture side movements and to enhance the depth of the signing actions. The background features a green screen, enabling easy post-processing to remove or alter the

---

[*]Work done while at the University of Queensland.
[†]Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.
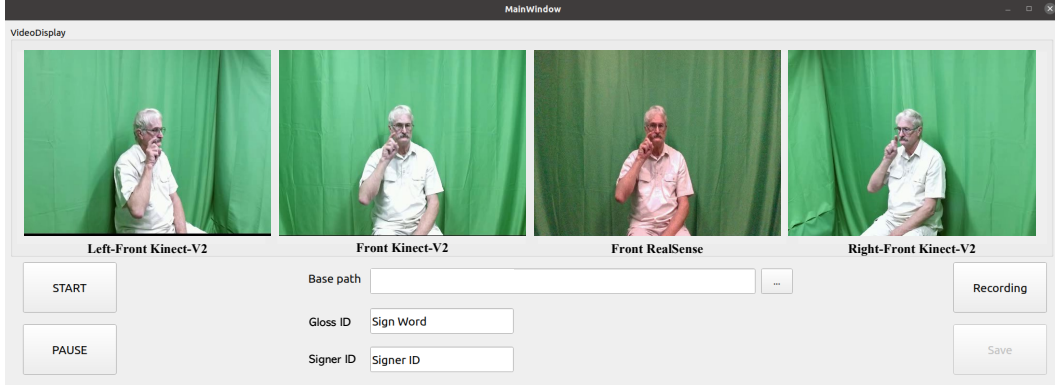
Figure 2: Recording interactive interface for MM-WLAuslan.
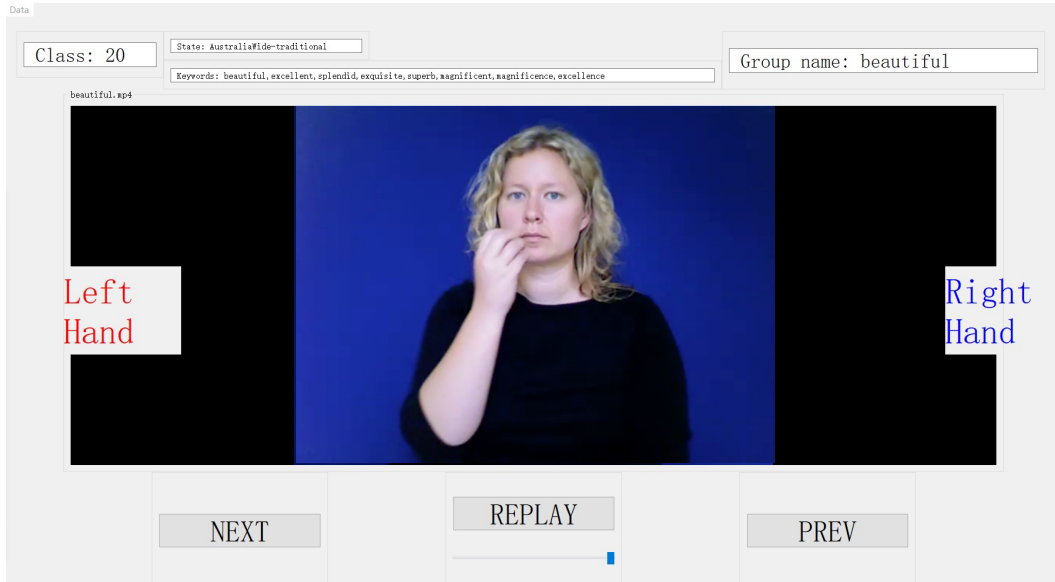
Figure 3: Auslan learning interactive interface.

background, focusing attention on the sign language. Our setup meticulously captures all necessary views and movements of the signer for accurate dataset collection.

A.2 MM-WLAuslan Recording Interface

The interactive interface of a software application designed for recording an Auslan (Australian Sign Language) dataset, as shown in Figure 2. The interface displays multiple video feeds of a signer from various views, captured by different RGB-D cameras: "Left-Front Kinect-V2", "Front Kinect-V2", "Front RealSense", and "Right-Front Kinect-V2". It ensures coverage of the movement of the signer from multiple perspectives to accurately capture the nuances of sign language. The interface also includes several control options: "START" and "PAUSE" buttons under the Left-Front Kinect-V2 playback window, and a "RECORDING" status indicator next to the Right-Front Kinect-V2 playback window, indicating that recording is in progress. Additionally, there are text fields for "Base path", "Gloss ID", "Sign Word", and "Signer ID", allowing users input specific data related to the recording session, such as the location of the saved data, identifiers for the signs, and the information of the signer. Overall, our recording interactive interface is tailored to facilitate the efficient capturing, labelling, and storage of sign language data. It is essential for creating a detailed and accessible word-level Auslan dataset.

2

Table 1: Comparison of Kinect-V2 and RealSense Cameras

Feature	Kinect-V2	RealSense
Manufacturer	Microsoft	Intel
Depth Sensing Technology	Time-of-Flight	Stereo Vision
Depth Range	0.5 to 4.5 meters	0.2 to 10 meters
Resolution	1920×1080	1280×720
Frame Rate	30 FPS	60 FPS
Field of View	Horizontal: 70°, Vertical: 60°	Horizontal: 87°, Vertical: 58°
SDK Support	Windows SDK, supports C# and C++	Intel RealSense SDK, supports multiple languages

A.3 Auslan Learning Interactive Interface

To facilitate the recording and learning process for volunteers without a background in Auslan, inspired by [1], we design an interactive learning interface. As shown in Figure 3, our interface displays the current sign being recorded, along with its primary meanings and a selection of synonyms to provide a comprehensive understanding of the usage of the sign. For instance, the sign for "beautiful" gloss is also related to words such as "excellent", "splendid", "magnificent" and "excellence".

The interface features a video display of the signer performing the sign, with labels "Left Hand" and "Right Hand" to guide viewers on the specific hand movements. Underneath the video, control buttons labelled "NEXT", "REPLAY" and "PREV" allow users to navigate through different signs easily. Our Auslan learning interactive interface not only supports the recording process but also acts as an educational tool. This promotes the learning and understanding of Auslan among the volunteers.

A.4 Comparison between Kinect-V2 and RealSense

In the MM-WLAuslan dataset, we utilize two distinct types of RGB-D cameras: the Kinect-V2 and the RealSense. As outlined in Table 1, these cameras differ primarily in their depth sensing technologies and other key specifications.

The Kinect-V2, developed by Microsoft, uses Time-of-Flight (ToF) technology to capture depth information. This method involves emitting infrared light pulses towards an object and then measuring the return time of these pulses to the sensor. The distance to each point on the object is calculated based on the time delay, using the constant speed of light. ToF technology is renowned for its ability to produce high-resolution and high-accuracy depth maps that perform robustly across various lighting conditions.

Conversely, the Intel RealSense cameras employ stereo vision, a technique that mimics human binocular vision. Stereo vision uses two cameras to capture images from slightly different angles. Depth is determined by identifying corresponding points between the two images and calculating the disparity in their positions. This approach allows the algorithm to estimate the distance of these points from the cameras, providing a versatile solution suitable for a broad range of applications.

Both Kinect-V2 and RealSense cameras offer substantial support through their respective SDKs, with Kinect supporting C# and C++ through the Windows SDK, and RealSense offering support for multiple languages through the Intel RealSense SDK.

In summary, while the Kinect-V2 excels in precision and is ideal for environments where accurate depth perception is crucial, RealSense cameras adapt more broadly due to their stereo vision capabilities. This makes each camera uniquely suited to different aspects of the MM-WLAuslan project, depending on the specific requirements for depth accuracy and environmental adaptability.

B Processing MM-WLAuslan

In this section, we introduce the post-processing and storage of data collected during the project.

B.1 Pose Extraction and Clean

As mentioned in Section 3.2, followed by [2], we use Alphapose [3, 4, 5] to track people in each word-level sign video and obtain the whole body keypoints. For each frame, we save 136 keypoints for each person, including 26 pose landmarks from the body, 68 pose landmarks from the face and 21

3
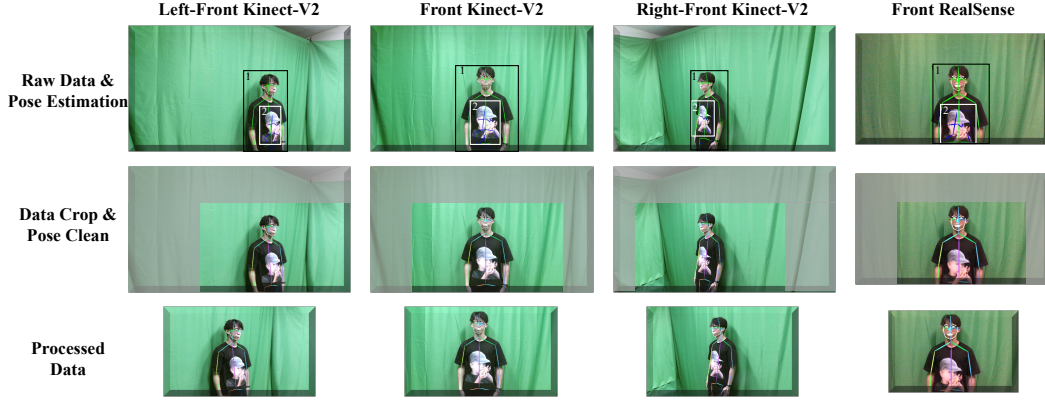
Figure 4: Data Post-Processing.

additional landmarks for each hand. Alphapose is an accurate multi-person pose estimator, which is the first open-source system. To efficiently match poses of the same person across frames, it provides an online pose tracker called Pose Flow. It is the first open-source online pose tracker.

We perform cleaning on the extracted pose sequences. The first row of Figure 4 shows the result of Alphapose in a special frame. Due to the signer wearing clothing with a character pattern, the pose estimation results mistakenly included keypoint sequences for two individuals. Thankfully, AlphaPose provides tracking IDs for each detected frame, enabling us to easily remove the unrelated pose sequence.

B.2 Sign Video Crop

After recording all the sign language videos, we notice that a significant portion of the footage consists of a green screen background, as shown in the first row of Figure 4. This not only provides unnecessary content but also poses a significant challenge in terms of video storage. Storing raw data would occupy more than 5TB of memory. Therefore, we crop the videos based on a fixed-size box that can cover every signer. We then proportionally resize the longest side to 512 pixels. In the last row of Figure 4, we display samples of the processed data.

B.3 Final Dataset Storage

The final datasets are stored in a folder on Google Drive G MM-WLAuslan. As shown in Figure 5, our **MM-WLAuslan** is organized into several main directories, each containing various types of data essential for training, validating, and testing in multimedia and sign language analysis projects. The detailed organization facilitates straightforward access to data types across different experimental settings, as described below:

- **Annotations**
  - *Pose* - The keypoint sequence of the signer in each sign language video.
  - *Split* - File delineating the division of data into training, validation, and test sets.
  - *Labels* - Gloss ID labels or annotations corresponding to the data samples.
- **Train**
  - *RGB* - Contains RGB videos.
  - *Depth* - Contains depth data, providing the distance of surfaces from a point of view.
- **Valid**
  - *RGB* - RGB data for validation.
  - *Depth* - Depth data for validation.
- **Test**
  - *Test-STU*: consistent scene settings with the training set.

4
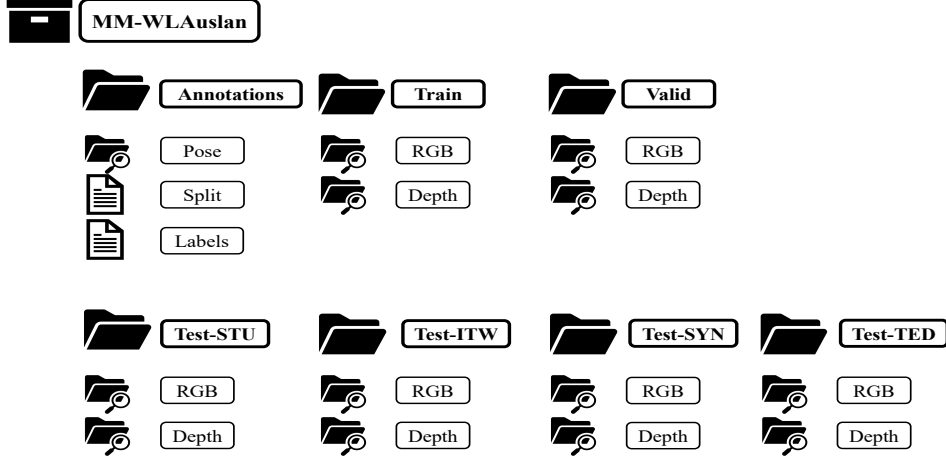
Figure 5: Dataset Storage.

- *Test-ITW*: green screens are removed and replaced with dynamic or static backgrounds.
- *Test-SYN*: synthesize indoor and outdoor backgrounds.
- *Test-TED*: randomly adjusting video segments.
    * *RGB* - RGB data for testing.
    * *Depth* - Depth data for testing.

C Experimental Settings

C.1 Baseline Models

We mention that all models used in this work are publicly available. We express profound gratitude to the aforementioned authors for their invaluable contributions. Each of the ISLR models we use is linked below:

- **RGB-based & RGB-D-based model:** ResNet2+1D [6] ⚙, TSN [7] ⚙, I3D [8] ⚙, S3D [9] ⚙, SlowFast [10] ⚙, Timesformer [11] ⚙, UMDR [12] ⚙ and KVNet-V [13] ⚙.
- **2D pose-based & 3D pose-based model:** TGCN [14] ⚙, SL-GCN [15] ⚙, STC-SLR [16] ⚙, DSTA-SLR [17] ⚙, SPTOTER [18] ⚙ and KVNet-K [13] ⚙.
- **Multi-modal-based model:** SAM-SLR [15] ⚙ and NLA-SLR [13] ⚙.

C.2 Training Hardware

All experiments are conducted on a machine equipped with four NVIDIA A100 80GB GPUs, which ensures robust computational capabilities for processing complex models and large datasets.

C.3 Hyperparameter Adjustment Formula

For different models, based on the available hardware configuration, we adjust the batch size and learning rate to optimize training performance. The learning rate (*lr*) adjustment is computed using the following formula:

$$lr[new] = lr[default] \times \frac{batch\_size[new] \times gpu\_number[new]}{batch\_size[default] \times gpu\_number[default]},$$

where:

- *lr[new]* is the adjusted learning rate for the new model.

5

Table 2: **The baseline of Single-view ISLR on MM-WLAuslan with Front RealSense camera.** "STU", "ITW", "SYN", "TED", and "AVG." represent the studio set, in-the-wild set, synthetic background set, temporal disturbance set and average performance across the four subsets, respectively. **Bold** indicates the highest value within the same data type.

Model	Data Type	STU		ITW		SYN		TED		AVG.	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
UMDR [12]	Pixel	**87.40**	**97.60**	15.06	34.75	**43.29**	65.69	**84.66**	**96.21**	**57.60**	**73.56**
KVNet-V [13]	Pixel	66.41	89.58	**26.82**	**52.05**	41.70	**68.52**	56.52	82.35	47.86	73.12
SPOTER [18]	2D Pose	53.34	75.12	49.89	73.25	51.21	72.13	32.16	64.29	46.65	71.20
DSTA-SLR [17]	2D Pose	**81.98**	**96.49**	**75.83**	**93.67**	**77.78**	**93.70**	**62.19**	**84.99**	**74.44**	**92.21**
STC-SLR [16]	2D Pose	79.14	95.44	72.24	92.32	75.29	92.71	58.43	82.26	71.27	90.68
KVNet-K [13]	2D Pose	66.55	89.33	58.08	85.22	62.93	85.89	36.79	65.68	56.09	81.53
KVNet-V [13]	Pixel + Depth	68.07	91.14	35.20	63.88	54.66	81.00	54.91	81.91	53.21	79.48
UMDR [12]	Pixel + Depth	**91.34**	**98.64**	**75.66**	**92.78**	**84.25**	**95.83**	**86.65**	**97.50**	**84.47**	**96.19**
SPOTER [18]	3D Pose	57.19	80.05	53.81	78.85	56.24	77.26	37.47	70.65	51.17	76.70
SL-GCN [15]	3D Pose	**73.02**	**88.69**	**59.28**	**81.72**	**68.15**	**84.50**	**46.81**	**73.68**	**61.82**	**82.15**
SAM-SLR [15]	Multi-Modal	78.02	94.69	64.28	87.72	73.15	91.50	53.81	80.68	67.31	88.65
NLA-SLR [13]	Multi-Modal	**83.65**	**96.96**	**68.30**	**90.40**	**76.70**	**93.15**	**67.15**	**89.25**	**73.95**	**92.44**

Table 3: **The baseline of Single-view ISLR on MM-WLAuslan with Left-Front Kinect-V2.**

Model	Data Type	STU		ITW		SYN		TED		AVG.	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
UMDR [12]	Pixel	**85.67**	**97.15**	13.72	29.61	15.58	32.06	**83.05**	**95.75**	49.51	63.64
KVNet-V [13]	Pixel	80.59	95.74	**45.17**	**71.29**	**57.93**	**82.92**	64.73	86.86	**62.11**	**84.20**
SPOTER [18]	2D Pose	65.09	88.15	60.26	83.45	57.47	80.87	56.56	82.33	59.84	83.70
DSTA-SLR [17]	2D Pose	**78.80**	**95.38**	73.82	93.17	**73.75**	**91.88**	**62.55**	**86.33**	72.23	**91.69**
STC-SLR [16]	2D Pose	78.23	94.82	**76.08**	**93.83**	73.61	90.87	61.45	84.96	**72.34**	91.12
KVNet-K [13]	2D Pose	73.88	93.26	64.66	89.36	66.72	87.60	53.17	80.38	64.61	87.65
KVNet-V [13]	Pixel + Depth	85.46	97.36	**61.70**	**86.41**	75.47	92.53	69.95	90.03	73.14	91.58
UMDR [12]	Pixel + Depth	**91.16**	**98.71**	46.90	70.90	**79.29**	**92.93**	**76.74**	**97.23**	**76.02**	89.95
SPOTER [18]	3D Pose	66.10	87.46	62.43	83.60	61.21	84.07	55.24	79.62	61.24	83.69
SL-GCN [15]	3D Pose	**74.08**	**90.75**	**70.87**	**86.42**	**71.19**	**85.94**	**60.78**	**84.28**	**69.23**	**86.85**
SAM-SLR [15]	Multi-Modal	88.55	98.06	76.77	94.35	79.41	93.56	73.57	91.91	79.57	94.47
NLA-SLR [13]	Multi-Modal	**89.51**	**98.41**	**78.72**	**95.15**	**82.91**	**94.90**	**75.69**	**93.06**	**81.71**	**95.38**

Table 4: **The baseline of Single-view ISLR on MM-WLAuslan with Right-Front Kinect-V2.**

Model	Data Type	STU		ITW		SYN		TED		AVG.	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
UMDR [12]	Pixel	**88.27**	**97.77**	2.55	8.62	1.51	4.84	**83.22**	**95.54**	43.89	51.69
KVNet-V [13]	Pixel	80.82	95.68	**37.97**	**65.94**	**37.62**	**64.82**	62.80	85.85	**54.80**	**78.07**
SPOTER [18]	2D Pose	62.29	84.84	58.54	81.37	52.88	77.81	51.49	78.88	56.30	80.72
DSTA-SLR [17]	2D Pose	**80.72**	95.44	**73.61**	**92.53**	**73.64**	**91.40**	**61.73**	**85.01**	**72.42**	**91.09**
STC-SLR [16]	2D Pose	79.95	**95.54**	72.89	92.49	73.14	91.23	59.56	83.22	71.38	90.62
KVNet-K [13]	2D Pose	69.97	91.40	64.12	88.02	64.01	86.73	50.50	78.20	62.15	86.09
KVNet-V [13]	Pixel + Depth	84.09	96.87	**40.99**	**68.18**	67.54	88.88	66.33	88.55	64.74	**85.62**
UMDR [12]	Pixel + Depth	**90.95**	**98.56**	13.80	28.72	**73.92**	**90.74**	**85.81**	**96.87**	**66.12**	78.72
SPOTER [18]	3D Pose	64.98	**87.76**	60.53	83.37	55.99	79.11	52.49	80.44	58.50	82.67
SL-GCN [15]	3D Pose	**72.99**	85.40	**63.21**	**86.82**	**70.54**	**84.12**	**59.17**	**87.44**	**66.48**	**85.95**
SAM-SLR [15]	Multi-Modal	88.49	97.90	**76.65**	**94.50**	74.54	91.95	72.34	91.81	78.00	94.04
NLA-SLR [13]	Multi-Modal	**88.95**	**98.13**	70.49	91.86	**80.00**	**94.12**	**74.12**	**92.24**	**78.39**	**94.09**

- *lr[default]* is the default learning rate.

- *batch_size[new]* and *batch_size[default]* are the new and default batch sizes, respectively.

- *gpu_number[new]* and *gpu_number[default]* refer to the number of GPUs used in the new and default setups, respectively.

For other hyperparameters, we utilize the default values to train the models. This approach helps in maintaining consistency and reliability across different experimental runs.

6

Table 5: **The baseline of Cross-Camera ISLR on MM-WLAuslan.** *"K"*, *"RS"* and *"K+"* represent Front Kinect-v2, Front RealSence and Left-Front + Right-Front Kinect-v2, respectively. "STU", "ITW", "SYN", "TED", and "AVG." represent the studio set, in-the-wild set, synthetic background set, temporal disturbance set and average performance across the four subsets, respectively.

Model	Train	Test	Data Type	STU		ITW		SYN		TED		AVG.	
				Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
KVNet-K [13]	K	K	2D Pose	82.88	96.70	76.29	94.56	79.07	94.07	69.05	89.80	76.82	93.78
	RS	RS	2D Pose	66.55	89.33	58.08	85.22	62.93	85.89	36.79	65.68	56.09	81.53
	K	RS	2D Pose	58.23	85.42	56.89	84.46	54.42	81.46	46.60	71.43	54.04	80.69
	RS	K	2D Pose	55.36	82.72	36.10	65.90	48.95	76.00	40.61	68.41	45.26	73.26
	RS	K+	2D Pose	19.87	40.32	15.82	36.25	16.10	35.02	13.67	29.35	16.36	35.23
NLA-SLR [13]	K	K	Multi-Modal	86.32	97.79	79.05	94.91	84.26	96.16	77.98	91.76	81.90	95.16
	RS	RS	Multi-Modal	83.65	96.96	68.30	90.40	76.70	93.15	67.15	89.25	73.95	92.44
	K	RS	Multi-Modal	66.95	91.38	60.80	87.54	61.30	85.38	56.85	80.60	61.48	86.22
	RS	K	Multi-Modal	57.34	83.12	47.37	78.05	50.02	78.85	46.58	72.35	50.33	78.09
	RS	K+	Multi-Modal	15.19	32.99	9.18	24.06	11.57	27.01	10.03	23.86	11.49	26.98

Table 6: **The baseline of Cross-view ISLR on MM-WLAuslan.** *"L"*, *"F"* and *"R"* represent left-front, front and right-front Kinect-v2, respectively.

Model	Train	Test	Data Type	STU		ITW		SYN		TED		AVG.	
				Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
KVNet-K [13]	F	F	2D Pose	82.88	96.70	76.29	94.56	79.07	94.07	69.05	89.80	76.82	93.78
	L	L	2D Pose	73.88	93.26	64.66	89.36	66.72	87.60	53.17	80.38	64.61	87.65
	R	R	2D Pose	69.97	91.40	64.12	88.02	64.01	86.73	50.50	78.20	62.15	86.09
	F	L+R	2D Pose	27.33	52.06	28.97	54.53	31.76	57.71	24.08	47.92	28.04	53.06
	L	F+R	2D Pose	21.57	41.05	15.69	33.29	18.32	37.16	13.63	29.23	17.30	35.18
	R	F+L	2D Pose	16.31	32.79	18.70	35.72	15.58	31.91	12.53	26.96	15.78	31.84
NLA-SLR [13]	F	F	Multi-Modal	86.32	97.79	79.05	94.91	84.26	96.16	77.98	91.76	81.90	95.16
	L	L	Multi-Modal	89.51	98.41	78.72	95.15	82.91	94.90	75.69	93.06	81.71	95.38
	R	R	Multi-Modal	88.95	98.13	70.49	91.86	80.00	94.12	74.12	92.24	78.39	94.09
	F	L+R	Multi-Modal	38.91	66.31	29.37	54.98	35.78	62.53	28.84	53.33	33.23	59.29
	L	F+R	Multi-Modal	34.49	55.23	23.75	43.33	31.29	51.28	23.39	43.13	28.23	48.24
	R	F+L	Multi-Modal	31.05	52.42	32.42	54.18	30.67	52.32	23.94	43.80	29.52	50.68

D   Additional Experiments

D.1   More Baseline on MM-WLAuslan

In Table 3 of the main paper, we specifically highlight models that demonstrate superior performance on "Front Kinect-V2" data across various modalities. Table 2, Table 3, and Table 4 present the test results of these models when evaluated using "Front RealSense", "Left-Front Kinect-V2" and "Right-Front Kinect-V2" cameras, respectively. We observe that model performance improves with the increase in modalities. These experiments demonstrate that pixel-based models excel in controlled environments like STU, where conditions are stable and noise is minimal. Conversely, pose-based models perform better in challenging environments such as ITW and SYN, leveraging structural information over textural details. The NLA-SLR [13] model exemplifies the integration of these modalities. As the state-of-the-art model for ISLR, NLA-SLR combines the high-performance KVNet-V and KVNet-K models, which handle pixel and pose data, respectively. This model consistently achieves high accuracy across all test subsets, demonstrating its robustness in varied settings. We also include experiments on Cross-Camera and Cross-View ISLR with different modalities in Table 5 and Table 6, highlighting the complexity of this task. The NLA-SLR [13] model performs excellently within its training configuration. However, when test across different cameras and views, there is a noticeable drop in performance. This reveals the challenges in maintaining accuracy across cross-camera and cross-view ISLR settings and suggests that further research is necessary to enhance model robustness in these settings. It is crucial for developing reliable systems that operate effectively in varied environments.

D.2   Evaluation on More Noisy Settings

The high quality of the data in MM-WLAuslan allows for the effective simulation of low-quality datasets. For instance, we can add white noise to simulate noisy environments, compress videos to

7

Table 7: **Low-quality sign video test for KVnet-V model trained with MM-WLAuslan.** "CRF" represent constant rate factor. "STU", "ITW", "SYN", "TED", and "AVG." represent the studio set, in-the-wild set, synthetic background set, temporal disturbance set and average performance across the four subsets, respectively.

Noise Addition Metho	STU		ITW		SYN		TED		AVG.	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Original High Quality	84.51	97.57	39.88	68.00	56.56	82.18	70.31	90.86	62.82	84.65
Downsample (128x102)	81.54	96.92	31.87	57.82	52.88	79.92	65.83	89.05	58.03	80.93
Downsample (54x32)	50.53	76.14	4.84	13.89	16.02	35.79	35.27	60.81	26.67	46.66
White Noise (μ=0, σ=0.3)	71.10	89.74	12.21	27.66	26.16	45.43	53.41	77.22	40.72	60.01
White Noise (μ=0, σ=0.5)	13.75	30.22	3.19	8.99	7.50	16.77	13.16	27.47	9.40	20.86
Video Compression (CRF=38)	79.50	95.86	26.46	52.54	46.41	74.81	61.87	86.10	53.56	77.32
Video Compression (CRF=48)	36.35	66.30	2.75	10.17	10.17	24.84	27.59	52.63	19.21	38.48

Table 8: **Signer-independent Testing Results.**

Model	Data Type	Statistical Measures	STU		ITW		SYN		TED	
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
KVNet-V [13]	RGB	Maximum	100.00	100.00	64.66	91.72	100.00	100.00	100.00	100.00
		Minimum	56.37	83.33	25.06	49.50	0.00	12.60	23.81	67.74
		Mean	80.75	95.27	39.27	67.10	51.00	79.81	66.06	88.83
		Standard Deviation	2.21	0.35	1.10	1.41	4.40	2.92	2.01	0.84
DSTA-SLR [17]	2D Pose	Maximum	100.00	100.00	82.20	96.50	100.00	100.00	85.00	100.00
		Minimum	53.33	80.00	51.38	87.50	5.55	21.01	33.33	62.00
		Mean	80.49	94.06	73.17	93.68	74.11	91.39	63.01	86.44
		Standard Deviation	1.94	0.38	0.53	0.05	3.81	2.50	1.29	0.52

simulate compression artifacts and downsample data to simulate low-resolution footage. We apply these methods to all test sets of MM-WLAuslan and use the KVnet-V [13] model for testing.

As shown in Table 7, the model's performance remains unaffected under mild noise levels. However, as the noise level increases, the model's accuracy decreases. This indicates that recognising low-quality sign videos is a challenging task. In the future, we will simulate a wider variety of low-quality data, such as low bitrate, ill-illuminated conditions and motion blur, and enhance the model's robustness through fine-tuning methods.

Additionally, considering the complexity of real-world scenarios, we have synthesised complex backgrounds (SYN) and recorded in-the-wild (ITW) data for further testing. As noted in the Limitation Section, MM-WLAuslan still needs to collect more data captured in more complicated real-world scenarios. In our future work, we plan to address this limitation by expanding our dataset. This expansion will encompass videos captured by consumer-level cameras (e.g., mobile phones), varying noise levels, and different camera angles.

D.3 Signer-independent Testing

We conduct signer-independent testing by evaluating each signer individually to measure Top-1 and Top-5 accuracy. We evaluate using KVnet-V [13] and DSTA-SLR [17]. Since MM-Auslan contains a large number of signers, we only provide the minimum, maximum, mean and standard deviation across different signers in Table 8. The results indicate that DSTA-SLR outperforms KVnet-V, as it not only achieves a higher mean accuracy but also exhibits a lower standard deviation. Based on the signer-independent test, we can assess how well the models generalize to different signers, especially those unseen during training.

E   The usefulness of ISLR task and our MM-WLAuslan dataset

The ISLR task serves as a critical foundation for the development of practical applications and sign language-related research. For practical applications, MM-WLAuslan offers higher-resolution videos and multiple viewing angles. Compared to the official Auslan Bank (https://auslan.org.au/dictionary/), it would be an ideal resource for an Auslan dictionary. Additionally, ISLR models have been used in sign language teaching applications, similar to the [1]. From a research perspective, ISLR also influences the development of sign language-related tasks. For instance, ISLR models can be used to

8

extract video features for gloss-free sign language translation [19], recognise short segments in video streams to achieve real-time continuous sign language recognition [20], and generate continuous sign language with 3D Avatars using isolated signs [21]. It can even be used in ways similar to spoken language conversations [22, 23, 24, 25, 26]. Moreover, MM-WLAuslan can be used to research multi-view 3D sign language generation and cross-view ISLR. We believe that exploring how MM-WLAuslan can be leveraged to improve the performance of other sign language-related tasks is a valuable direction for future research. In conclusion, while our dataset currently focuses on ISLR, it is designed with both practical applications and future research in mind. We are working towards expanding its utility to support more complex tasks, such as continuous recognition and translation, thereby enhancing its relevance and impact in sign language research and application.

F   Consent Form for MM-Auslan Recording

Due to the inclusion of facial information in our dataset, we obtain consent from volunteers and have them sign the consent form depicted in Figure 6 before recording data. **We do not release personally identifiable information** such as names, ages, occupations, or indications of whether individuals are deaf or hard of hearing. It is important to note that our dataset is strictly for academic use and can not be used for commercial purposes.

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 6: Consent Form for Recording.

9

References

[1] 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.

[2] Xin Shen, Shaozu Yuan, Hongwei Sheng, Heming Du, and Xin Yu. Auslan-daily: Australian sign language translation for daily communication and news. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *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.

[3] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[5] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[6] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6450–6459. Computer Vision Foundation / IEEE Computer Society, 2018.

[7] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[8] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017.

[9] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, pages 318–335. Springer, 2018.

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6201–6210. IEEE, 2019.

[11] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[12] 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 Trans. Pattern Anal. Mach. Intell.*, 45(10):11428–11442, 2023.

[13] 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.

[14] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020.

[15] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

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

10

[17] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Dynamic spatial-temporal aggregation for skeleton-aware sign language recognition. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *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.

[18] Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 182–191, January 2022.

[19] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[20] Ronglai Zuo, Fangyun Wei, and Brian Mak. Towards online sign language recognition and translation. *arXiv preprint arXiv:2401.05336*, 2024.

[21] Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. A simple baseline for spoken language to sign language translation with 3d avatars. *arXiv preprint arXiv:2401.04730*, 2024.

[22] Lei Shen, Haolan Zhan, Xin Shen, Yonghao Song, and Xiaofang Zhao. Text is NOT enough: Integrating visual impressions into open-domain dialogue generation. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4287–4296. ACM, 2021.

[23] Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 1598–1608. ACM, 2021.

[24] Lei Shen, Haolan Zhan, Xin Shen, and Yang Feng. Learning to select context in a hierarchical and global perspective for open-domain dialogue generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 7438–7442. IEEE, 2021.

[25] Yuqi Liu, Wenqian Zhang, Sihan Ren, Chengyu Huang, Jingyi Yu, and Lan Xu. Scope: Sign language contextual processing with embedding from llms. *arXiv preprint arXiv:2409.01073*, 2024.

[26] Yiwei Wei, Shaozu Yuan, Meng Chen, Xin Shen, Longbiao Wang, Lei Shen, and Zhiling Yan. Mpp-net: Multi-perspective perception network for dense video captioning. *Neurocomputing*, 552:126523, 2023.

11