THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

NEURAL INFORMATION PROCESSING SYSTEMS

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

x.shen3@uqconnect.edu.au

The University of Queensland, Brisbane, Australia

Contents

Background

Hearing Loss & Deaf

- **World**:

World Health Organization [1] (1st April 2021): Over **5%** of the world's population (or 430 million people) require rehabilitation to address their "disabling" hearing loss (432 million adults and 34 million children). By 2050 over 700 million people (or **one in every ten** people) will have disabling hearing loss.

- **Australia**:

Australian Federal Department of Health and Aged Care (DHAC) [2] (14th May 2024): **One in six** Australians suffers from hearing loss, which is expected to rise to **one in four** by 2050.

[1] Deafness and hearing loss https://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] The facts about Hearing Health in Australia https://www.health.gov.au/topics/ear-health/about

Background

Sign Language

- **World**:

  Sign language (SL) is the primary way for deaf or hearing loss people to express themselves. Sign languages are visual languages which convey information by signers' handshape, facial expressions, body movements, and so forth. Each sign language has its own unique vocabulary and grammar rules, much like spoken languages.

- **Australia**:

  Distinct sign languages develop in different regions, even among countries with the same spoken language, such as the American SL, British SL and Australia SL.

Motivation – Sign Language Processing

- Deaf and people with hearing lose will **face problems more easily than hearing people**.
  - Children: dropping out of school.
  - Adults: losing jobs.
  - They will **feel lonely due to social isolation**, since the **sign language** is what they can only use to communicate.

- To solve the above problems, there are two methods:
  - 1. Improve the hearing ability in medical level.
  - 2. With emerging **deep learning techniques** and **large-scale sign language datasets**, sign language processing achieves promising progress recently.

Motivation – Isolated Sign Language Recognition

Isolated Sign Language Recognition (ISLR) focuses on identifying individual sign language signs.

one hundred
(gloss)

Gloss is a unique label for a single sign. Each gloss is identified by a word or a phrase which is associated with the sign's semantic meaning.

Motivation – MM-WLAuslan

Auslan, as a sign language specific to Australia, still lacks a dedicated large-scale word-level dataset for the ISLR task. Moreover, most publicly available datasets have limitations in **gloss dictionary size**, **depth information**, and **recording perspectives**.

To fill this gap, we curate the first large-scale Multi-view Multi-modal Word-Level Australian Sign Language recognition dataset, dubbed **MM-WLAuslan**:

(1) the **largest** amount of data.

(2) the **most extensive** vocabulary.

(3) the **most diverse** of multi-modal camera views.

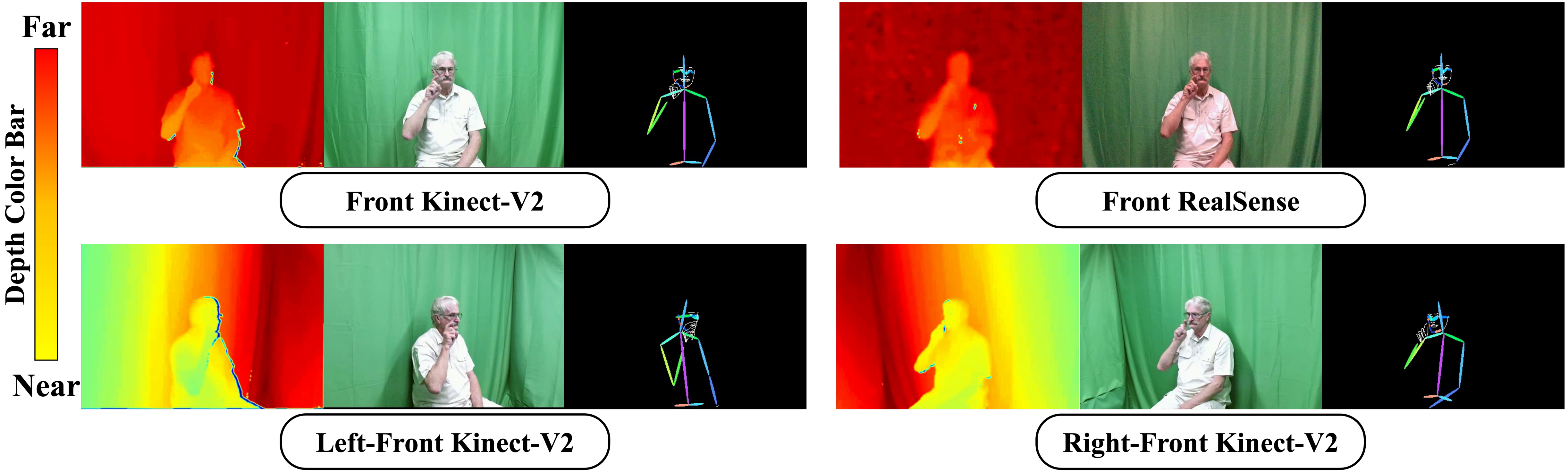
Contents

Recording Environment

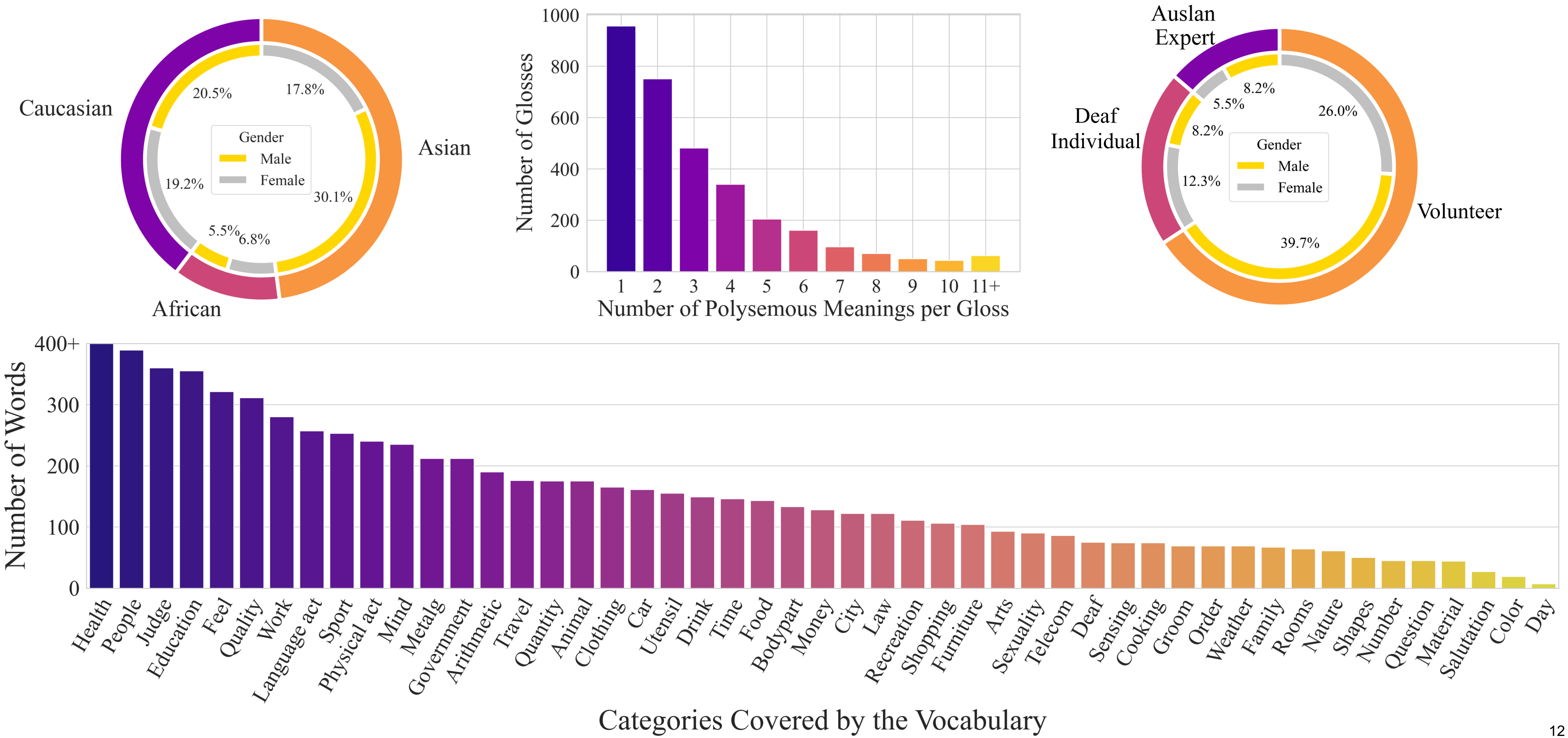
MM-WLAuslan Overview Silhouette

MM-WLAuslan includes three Kinect-V2 cameras and a RealSense camera arranged hemispherically around the front half of the signer to capture multi-view and multi-modal data.

Multi-Modal Data Sample



Statistics of Signers and Glosses



Comparison between MM-WLAuslan and Existing ISLR Datasets

Dataset	Country	Signs	Signers	Videos	Ave.Videos/Sign	Cross-Cam	Depth	Source
Purdue RVL-SLLL	USA	39	14	0.5K	14	✗	✓	Studio
RWTH-BOSTON 50	USA	50	3	0.5K	9.66	✓	✗	Studio
ASLLVD	USA	3,000	6	9.8K	3.27	✓	✗	Studio
WLASL	USA	2,000	119	21.1K	10.54	✗	✗	Web
MS-ASL	USA	1,000	222	25.5K	25.51	✗	✗	Web
ASL Citizen	USA	2,731	52	83.9K	30.73	✗	✗	Webcam
PopSign ASL v1.0	USA	250	47	214.3K	857.30	✗	✗	Smartphone
BSL-1K	GBR	1,064	40	273.0K	257	✗	✗	Web
DEVISIGN-L	CHN	2,000	8	24.0K	12.00	✗	✓	Studio
CSL 500	CHN	500	50	125.0K	250.00	✗	✓	Studio
DGS Kinect 40	DEU	40	14	2.8K	70.00	✗	✓	Studio
SMILE	DEU/CHE	100	30	-	-	✓	✓	Studio
GSL 982	GRC	982	1	4.9K	5.00	✗	✗	Studio
INCLUDE	ISR	263	7	4.3K	16.30	✗	✗	Studio
KL-MV2DSL	ISR	200	-	5.0K	25	✓	✗	Studio
LSA64	ARG	64	10	3.2K	50.00	✗	✗	Studio
LSE-Sign	ESP	2,400	2	2.4K	1.00	✓	✗	Studio
LSFB-ISOL	FRA/BEL	395	100	47.6K	120.38	✗	✗	Studio
BosphorusSign22K	TUR	744	6	22.5K	30.30	✗	✓	Studio
AUTSL	TUR	226	43	38.3K	169.63	✗	✓	Studio
Auslan-Daily	AUS	600	21	3.0K	5.00	✗	✗	Web
MM-WLAuslan	**AUS**	**3,215**	**73**	**282.9K**	**88.00**	**✓**	**✓**	**Studio**

MM-Auslan Test Set

To evaluate the performance of ISLR systems under real-world scenarios, we provide a diverse test set with four distinct subsets, including:

- studio (STU) set
- in-the-wild (ITW) set
- synthetic background (SYN) set
- temporal disturbance (TED) set

Key Statistics of MM-WLAuslan Dataset Splits

Split	Train	Val	Test-STU	Test-ITW	Test-SYN	Test-TED
Num. Videos	154.3k	25.7k	25.7k	25.7k	25.7k	25.7k
Num. Signers	55	53	12	15	62	63
Num. OOS	-	-	10	2	15	10
BG Interference	✗	✗	✗	✓	✓	✗
TP Disturbance	✗	✗	✗	✗	✗	✓

"BG" and "TP" represent background and temporal, respectively.

"OOS" indicates the signers only occur in the test set

Contents

- Background & Motivation
- MM-WLAuslan Dataset
- **MM-WLAuslan Benchmark**
- Conclusion

MM-WLAuslan ISLR Benchmark

Single-view RGB-based ISLR involves recognizing isolated sign language from video sequences captured from a single fixed camera.

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
ResNet2+1D	Pixel	58.71	77.03	13.83	18.37	26.14	39.58	51.14	69.97	37.45	51.24
TSN	Pixel	51.17	68.60	11.06	23.75	31.01	45.89	40.40	69.10	33.41	51.84
I3D	Pixel	63.97	84.93	14.18	26.52	36.17	57.22	60.96	80.63	43.82	62.33
S3D	Pixel	75.55	94.11	29.41	55.11	44.60	71.34	62.21	85.26	52.94	76.46
SlowFast	Pixel	80.68	96.08	32.22	64.81	53.17	78.30	66.21	82.18	58.07	80.34
Timesformer	Pixel	73.20	81.40	21.14	56.44	41.88	65.83	68.40	79.67	51.15	70.84
UMDR	Pixel	80.86	95.88	13.57	28.66	13.99	31.01	**82.69**	**95.67**	47.78	62.81
KVNet-V	Pixel	**84.51**	**97.57**	**39.88**	**68.00**	**56.56**	**82.18**	70.31	90.86	**62.82**	**84.65**
TGCN	2D pose	68.62	86.30	58.01	74.74	63.50	81.38	47.68	68.82	62.11	77.81
SL-GCN	2D pose	71.07	91.21	66.59	89.5	63.20	86.94	**69.98**	88.99	67.71	89.16
SPOTER	2D pose	72.81	92.69	64.12	86.36	66.81	88.11	69.42	**90.94**	68.29	89.53
KVNet-K	2D pose	**82.88**	**96.70**	**76.29**	**94.56**	**79.07**	**94.07**	69.05	89.80	**76.82**	**93.78**
SAM-SLR	2D pose + Pixel	83.98	97.12	74.30	91.65	80.73	94.93	71.21	86.56	77.55	83.91
NLA-SLR	2D pose + Pixel	**86.32**	**97.79**	**79.05**	**94.91**	**84.26**	**96.16**	**77.98**	**91.76**	**81.90**	**95.16**

MM-WLAuslan ISLR Benchmark

Single-view RGB-D-based ISLR aims to enhance the recognition of isolated signs by incorporating depth information along with RGB data.

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
I3D	Pixel + Depth	65.74	88.57	21.71	41.32	61.06	85.41	47.25	65.71	48.94	70.25
S3D	Pixel + Depth	79.70	95.93	64.97	89.16	76.38	92.67	66.11	88.62	71.79	91.60
KVNet-V	Pixel + Depth	82.22	96.75	38.79	66.11	57.88	82.92	66.94	88.58	61.46	83.59
UMDR	Pixel + Depth	**91.65**	**98.81**	**72.52**	**90.46**	**83.77**	**95.18**	**88.35**	**98.07**	**84.07**	**95.63**
TGCN	3D pose	70.19	89.78	59.52	76.59	66.35	84.06	51.48	71.17	61.88	80.40
SPOTER	3D pose	74.95	95.88	66.75	89.41	70.22	91.23	71.65	92.36	70.89	92.22
SL-GCN	3D pose	**77.76**	**96.98**	**72.26**	**91.49**	**74.91**	**92.57**	**72.27**	**94.88**	**74.30**	**93.98**
NLA-SLR	2D pose + Pixel + Depth	85.65	95.65	80.20	95.58	**83.36**	94.04	83.34	**94.63**	83.14	94.98
SAM-SLR	3D pose + Pixel + Depth	**87.05**	**98.93**	**81.29**	**96.92**	83.03	**95.86**	**85.07**	93.53	**84.11**	**96.31**

MM-WLAuslan ISLR Benchmark

Multi-view RGB-based ISLR employs multiple cameras to capture the sign language videos.

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	Pixel	**92.56**	**99.09**	23.78	44.22	22.12	42.61	**90.13**	**98.23**	57.15	71.04
KVNet-V	Pixel	91.57	99.00	**62.25**	**86.19**	**70.90**	**90.97**	79.78	94.68	**76.13**	**92.71**
SPOTER	2D pose	76.92	95.55	67.79	89.98	69.21	92.16	74.34	**94.14**	72.06	92.96
KVNet-K	2D pose	**90.45**	**98.56**	**86.23**	**97.77**	**85.73**	**95.47**	**77.26**	93.93	**84.92**	**96.43**
SAM-SLR	2D pose + Pixel	85.85	97.68	77.36	92.88	84.26	95.69	79.92	88.10	81.85	93.59
NLA-SLR	2D pose + Pixel	**94.62**	**99.31**	**89.75**	**98.60**	**88.94**	**96.98**	**85.19**	**96.69**	**89.63**	**97.90**

Multi-view RGB-D-based ISLR incorporates depth data in a multi-view setup.

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	Pixel + Depth	**93.25**	**99.11**	**74.98**	**92.19**	**86.14**	**96.24**	**90.42**	**97.39**	**86.20**	**96.36**
KVNet-V	Pixel + Depth	87.67	98.22	66.01	88.80	83.06	95.27	74.23	92.28	77.74	93.64
SPOTER	3D pose	79.91	**96.91**	73.44	91.29	**76.41**	**93.58**	76.87	94.45	76.66	94.06
ST-GCN	3D pose	**81.77**	95.07	**77.34**	**93.13**	76.38	92.83	**79.36**	**96.73**	**78.71**	**94.44**
SAM-SLR	3D pose + Pixel + Depth	89.21	98.83	80.51	94.18	83.76	96.67	85.68	93.78	84.79	95.87
NLA-SLR	2D pose + Pixel + Depth	**94.43**	**99.37**	**88.95**	**98.49**	**89.52**	**97.14**	**85.13**	**96.46**	**89.51**	**97.87**

MM-WLAuslan ISLR Benchmark

Cross-Camera ISLR aims to test the robustness of the model against variations in camera specifications and settings. Training and testing data are captured from different cameras. It is challenging for the model to generalize across hardware-induced discrepancies.

Model	Train	Test	Data Type	STU Top-1	STU Top-5	ITW Top-1	ITW Top-5	SYN Top-1	SYN Top-5	TED Top-1	TED Top-5	AVG. Top-1	AVG. Top-5
KVNet-V	*K*	*K*	Pixel	84.51	97.57	39.88	68.00	56.56	82.18	70.31	90.86	62.82	84.65
	RS	*RS*	Pixel	66.41	89.58	26.82	52.05	41.70	68.52	56.52	82.35	47.86	73.12
	K	*RS*	Pixel	53.33	81.06	18.88	41.58	32.32	60.09	46.05	71.03	37.65	63.44
	RS	*K*	Pixel	31.28	55.3	5.85	15.73	14.35	30.39	25.35	46.55	19.21	36.99
	RS	*K+*	Pixel	5.36	14.45	1.97	6.36	1.97	6.39	3.84	11.03	3.28	9.56
UMDR	*K*	*K*	Pixel + Depth	91.65	98.81	72.52	90.46	83.77	95.18	88.35	98.07	84.07	95.63
	RS	*RS*	Pixel + Depth	91.34	98.64	75.66	92.78	84.25	95.83	86.65	97.50	84.47	96.19
	K	*RS*	Pixel + Depth	79.09	94.67	44.00	67.81	0.64	2.33	71.47	90.91	48.80	63.93
	RS	*K*	Pixel + Depth	71.20	89.87	35.08	59.93	46.11	68.40	61.05	83.88	53.36	75.52
	RS	*K+*	Pixel + Depth	11.25	26.67	2.45	8.03	3.84	11.37	7.88	19.00	6.36	16.27

K, ***RS*** and ***K+*** represent Front Kinect-v2, Front RealSence and Left-Front + Right-Front Kinect-v2, respectively.

MM-WLAuslan ISLR Benchmark

Cross-View ISLR requires the model to recognize signs from views not seen during training. The model must handle the appearance changes due to different viewing angles, thus testing its view-invariance capabilities.

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-V	*F*	*F*	Pixel	84.51	97.57	39.88	68.00	56.56	82.18	70.31	90.86	62.82	84.65
	L	*L*	Pixel	80.59	95.74	45.17	71.29	57.93	82.92	64.73	86.86	62.11	84.20
	R	*R*	Pixel	80.82	95.68	37.97	65.94	37.62	64.82	62.80	85.85	54.80	78.07
	F	*L+R*	Pixel	23.60	48.10	8.70	23.28	9.94	26.53	15.90	35.41	14.53	33.33
	L	*F+R*	Pixel	29.18	48.41	12.48	27.28	21.84	40.21	19.58	37.16	20.77	38.26
	R	*F+L*	Pixel	24.93	44.53	16.93	34.15	20.10	39.26	18.99	36.33	20.24	38.57
UMDR	*F*	*F*	Pixel + Depth	91.65	98.81	72.52	90.46	83.77	95.18	88.35	98.07	84.07	95.63
	L	*L*	Pixel + Depth	91.16	98.71	46.90	70.90	79.29	92.93	86.74	97.23	76.02	89.95
	R	*R*	Pixel + Depth	90.95	98.56	13.80	28.72	73.92	90.74	85.81	96.87	66.12	78.72
	F	*L+R*	Pixel + Depth	32.27	55.95	10.06	19.83	21.64	41.07	27.32	49.02	22.82	41.47
	L	*F+R*	Pixel + Depth	40.55	62.42	6.44	14.61	25.58	44.83	32.27	53.74	26.21	43.90
	R	*F+L*	Pixel + Depth	28.82	47.04	6.62	14.73	19.74	36.03	24.18	37.45	19.84	33.81

L, ***F*** and ***R*** represent left-front, front and right-front Kinect-v2, respectively.

Contents

- Background & Motivation
- MM-WLAuslan Dataset
- MM-WLAuslan Benchmark
- **Conclusion**

Summary of MM-WLAuslan

- We construct the first word-level Australian ISLR dataset, dubbed MM-WLAuslan. MM-WLAuslan consists of the largest number of gloss videos and the most extensive vocabulary.

- We provide the most diverse multi-modal camera views and enable the investigation of a variety of multi-modal ISLR settings, including multi-view, cross-camera and cross-view.

- We establish a leaderboard and an evaluation benchmark to promote future Australian ISLR research and development of applications.

Thanks for Watching!

For more details, please refer to our paper and appendix.

You are welcome to visit our project page at: **uq-cvlab.github.io/MM-WLAuslan-Dataset/**