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: **Illustrations of the curated MM-WLAuslan dataset.** 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.

Abstract

Isolated Sign Language Recognition (ISLR) focuses on identifying individual sign language signs. Considering the diversity of sign languages across geographical regions, developing region-specific ISLR datasets is crucial for supporting communication and research. Auslan, as a sign language specific to Australia, still lacks a dedicated large-scale word-level dataset for the ISLR task. To fill this gap, we curate **the first** large-scale Multi-view Multi-modal Word-Level Australian Sign Language recognition dataset, dubbed MM-WLAuslan. Compared to other publicly available datasets, MM-WLAuslan exhibits three significant advantages: (1) **the largest amount** of data, (2) **the most extensive** vocabulary, and (3) **the most diverse** of multi-modal camera views. Specifically, we record **282K+** sign videos covering **3,215** commonly used Auslan glosses presented by **73** signers in a studio environment. Moreover, our filming system includes two different types of cameras, *i.e.*, three Kinect-V2 cameras and a RealSense camera. We position cameras hemispherically around the front half of the model and simultaneously record videos using all four cameras. Furthermore, we benchmark results with state-of-the-art methods for various multi-modal ISLR settings on MM-WLAuslan, including multi-view, cross-camera, and cross-view. Experiment results indicate that MM-WLAuslan is a challenging ISLR dataset, and we hope this dataset will contribute to the development of Auslan and the advancement of sign languages worldwide. All datasets and benchmarks are available at MM-WLAuslan.

*Work done while visiting the University of Queensland.
†Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/079017-2227

1 Introduction

Sign language (SL) is the primary mode of communication for many deaf or hard-of-hearing individuals. Each sign language possesses its own vocabulary and grammatical rules, akin to spoken languages [1, 2, 3]. Notably, even within regions that share a commonly spoken language, such as the United States, Australia, and the United Kingdom, distinct native sign languages are prevalent. To facilitate communication between the deaf and hearing communities, Isolated Sign Language Recognition (ISLR) is highlighted as a fundamental sign language understanding task. ISLR aims to recognize an individual sign gloss, which is a written representation of signs using words from a spoken language, into a corresponding word or phrase in spoken languages [4, 5].

With emerging deep learning techniques [6, 7, 8, 9] and large-scale sign language datasets [4, 10, 11, 12, 13], ISLR achieves promising progress recently [14, 4, 15]. As shown in Table 1, researchers from various countries construct word-level sign language datasets and thus promote the development of ISLR in the respective sign languages, such as American Sign Language (ASL) [4, 16, 17, 18, 19, 10], British Sign Language (BSL) [20, 21], Chinese Sign Language (CSL) [22, 23] and German Sign Language (DGS) [12, 24]. Meanwhile, according to the Australian Federal Department of Health and Aged Care (DHAC)[3], as of 14 May 2024, one in six Australians, over 3.6 million people, had hearing loss affecting them, and the number is expected to reach 7.8 million people by 2060. However, to the best of our knowledge, there is no publicly available large-scale Auslan dataset for ISLR. Due to the regional nature of sign languages and the societal commitment to supporting individuals with hearing impairments, word-level Australian Sign Language (Auslan) datasets are inevitably and urgently needed in order to investigate automatic recognition.

Moreover, the volume of data, the breadth of data categories, and the diversity of data modalities are three critical points that influence the fundamental quality of an ISLR dataset. The larger the volume, the wider the range of categories, and the richer the modalities of data mean the higher the value of the dataset for scientific research and practical applications, such as sign language education [25] and dictionary [26]. Specifically, a large volume of data and an extensive gloss dictionary within the dataset enhance the robustness and capability of the sign recognition system. Additionally, the captured multi-view sign data and depth information improve the accuracy of recognizing complex hand movements and reduce the issues caused by occlusion and single-view ambiguities. However, most existing publicly available ISLR corpora either contain the limited gloss videos and vocabulary [5, 16, 27, 12, 13, 24, 28] or are only captured in a single viewpoint without depth information [4, 10, 19, 18, 20].

In this work, we record the first word-level Auslan recognition dataset, named MM-WLAuslan, that contains the largest number of data samples, the most extensive vocabulary, and the most diverse multi-modal camera views compared to other publicly available datasets, as shown in Table 1. Specifically, we select 3,215 commonly used glosses that contain a sufficient variety of classes and training instances for a practical word-level Auslan recognition model. We collect the glosses from *Auslan SignBank*[4] [29], the most authoritative Auslan dictionary in Australia. We ask Auslan experts to help select glosses that are widely used throughout Australia, including fingerspelling glosses[5], such as "*TV*" and "*NEWS*". The collected glosses correspond to over 7,900 English words or phrases, covering most of the vocabulary commonly used in daily life. We invite sign language experts, deaf individuals, and volunteers to participate in the recording process. After 2,500+ hours of preparation and recording, we capture over 282K+ high-quality isolated Auslan gloss videos with the assistance of 73 signers. Each video recording is supervised by at least one Auslan expert to ensure the precision of the sign language expression.

To record multi-view, multi-modal, and high-quality isolated Auslan gloss videos, we set up a recording studio equipped with a green screen backdrop. We position two different types of RGB-D cameras, *i.e.*, three Kinect-V2 cameras and a RealSense camera, hemispherically around the front half of the model. As shown in Figure 1, we place the cameras to the left-front, front, and right-front of the subject and simultaneously record videos. Unlike the previous dataset [24] that only provides depth video from the front view, we record both RGB-D videos from every camera.

[3] `https://www.health.gov.au/topics/ear-health/about`

[4] Auslan SignBank: `https://auslan.org.au/dictionary/`

[5] English words are signed letter by letter.

Table 1: **Comparison between MM-WLAuslan and existing ISLR datasets.**

Dataset	Country	Signs	Signers	Videos	Ave.Videos/Sign	Cross-Cam	Depth	Source
Purdue RVL-SLLL [16]	USA	39	14	0.5K	14	✗	✓	Studio
RWTH-BOSTON 50 [27]	USA	50	3	0.5K	9.66	✓	✗	Studio
ASLLVD [17]	USA	3,000	6	9.8K	3.27	✓	✗	Studio
WLASL [4]	USA	2,000	119	21.1K	10.54	✗	✗	Web
MS-ASL [10]	USA	1,000	222	25.5K	25.51	✗	✗	Web
ASL Citizen [19]	USA	2,731	52	83.9K	30.73	✗	✗	Webcam
PopSign ASL v1.0 [18]	USA	250	47	214.3K	857.30	✗	✗	Smartphone
BSL-1K [20]	GBR	1,064	40	273.0K	257	✗	✗	Web
DEVISIGN-L [23]	CHN	2,000	8	24.0K	12.00	✗	✓	Studio
CSL 500 [11]	CHN	500	50	125.0K	250.00	✗	✓	Studio
DGS Kinect 40 [12]	DEU	40	14	2.8K	70.00	✗	✓	Studio
SMILE [24]	DEU/CHE	100	30	-	-	✓	✓	Studio
GSL 982 [30]	GRC	982	1	4.9K	5.00	✗	✗	Studio
INCLUDE [31]	ISR	263	7	4.3K	16.30	✗	✗	Studio
KL-MV2DSL [28]	ISR	200	-	5.0K	25	✓	✗	Studio
LSA64 [13]	ARG	64	10	3.2K	50.00	✗	✗	Studio
LSE-Sign [32]	ESP	2,400	2	2.4K	1.00	✓	✗	Studio
LSFB-ISOL [33]	FRA/BEL	395	100	47.6K	120.38	✗	✗	Studio
BosphorusSign22K [34]	TUR	744	6	22.5K	30.30	✗	✓	Studio
AUTSL [35]	TUR	226	43	38.3K	169.63	✗	✓	Studio
Auslan-Daily [5]	AUS	600	21	3.0K	5.00	✗	✗	Web
MM-WLAuslan	**AUS**	**3,215**	**73**	**282.9K**	**88.00**	✓	✓	**Studio**

Furthermore, for an unbiased performance evaluation of ISLR systems, we involve nearly 20 signers in the test set who are not exposed to the training and validation sets. Concurrently, we split the test set into four distinct subsets to mimic the various scenarios in the real world. Videos in three subsets are designed to incorporate diverse backgrounds or potential temporal disturbances. After obtaining the realistic test sets, we utilize the collected multi-modal, multi-view, and multi-camera videos to benchmark various multi-modal ISLR settings. Extensive experiments demonstrate the limitations of current state-of-the-art (SOTA) methods when these methods are applied across various cameras and views. This manifests the potential of MM-WLAuslan to advance the future research and development of ISLR systems. Overall, our major contributions are summarized as follows:

- 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.

2 Related Work

2.1 Isolated Sign Language Recognition Datasets

As shown in Table 1, several datasets are developed to facilitate research and application development of ISLR. However, most datasets have limitations in gloss dictionary size, depth information, and recording perspectives. For example, Purdue RVL-SLLL dataset [16] exhibits methodological rigor in a laboratory setting, but its applicability for sign language recognition is limited because it only covers 39 signs. Furthermore, despite ASLLVD dataset [17] including a large lexicon of 3,000 glosses, it is limited by its single perspective and lack of depth information, crucial for capturing the three-dimensional motion of sign language. WLASL [4] and MS-ASL [10] datasets expand on the number of signs and signers but still restrict their recordings to single-view videos without depth, missing critical spatial dynamics essential for accurate sign interpretation. In contrast, datasets like CSL 500 [11] and DGS Kinect 40 [12] include depth information but cover only a small number of glosses, limiting their usefulness for extensive sign language applications.

Different from all of the above datasets, the proposed MM-WLAuslan dataset is a comprehensive ISLR dataset. It encompasses 3,215 signs from 73 signers, with each sign captured from four distinct viewpoints along with depth information, significantly enhancing the diversity and utility of the dataset. Moreover, MM-WLAuslan is currently the largest sign language recognition dataset in Australia, with extensive lexicon and high-quality data.

https://doi.org/10.52202/079017-2227

2.2 Isolated Sign Language Recognition Methods

ISLR aims to identify the gloss labels of short-term videos.Previous research can be categorized into three types based on the input modality: pixel-based, pose-based and multi-modal-based approaches.
Pixel-based ISLR: Significant advances in CNN-based action recognition inspire the development of pixel-based ISLR models. Early efforts [36, 37, 38] utilize convolutional neural networks (CNN) to extract frame-wise features, which are then temporally encoded using recurrent neural networks to capture time-series information. Meanwhile, 3D CNNs, such as C3D model [39, 40, 41] and I3D model [6], are commonly used in ISLR [10, 4, 42, 19, 43, 44].
Pose-based ISLR: Unlike RGB pixel-based methods, pose-based ISLR models are robust against background clutters, lighting conditions, and occlusions, while explicitly depicting human hand and limb movements [45, 46, 47, 48]. ST-GCN [47], the first to apply a spatio-temporal graph convolutional network for action recognition, encodes motions across the human kinetic chain. Subsequent studies utilize this spatio-temporal architecture, employing both graph convolutional networks [4, 49, 50, 51] and Transformers [52, 53, 54, 55] to embedding and analyze sign pose data.
Multi-modal-based ISLR: Recent studies show that combining pose, depth, and RGB modalities significantly improves ISLR. Zuo et al. [14] use the S3D model to extract RGB and pose heatmap features, enhancing recognition on the WLASL [4] dataset. Moreover, Jiang et al. [15] integrate depth information into the model, enabling recognition results to exceed 99% on the AUTSL dataset [35].

2.3 Multi-view and Multi-modal Action Recognition

Previous research [56] argues that Action Recognition (AR) methods can be applied on sign language recognition. To build an effective and robust real-world ISLR and AR system, initiating multi-view and multi-modal learning is essential [28]. Recent advancements in AR introduce various approaches for multi-view learning [57], including dictionary learning [58], neural networks with adjustable views [59], convolutional neural networks [60], and attention mechanisms [61]. Additionally, Zhu *et al.* [62] adopt vision transformer models as robust solutions for multi-view learning. Recent approaches [63, 64] develop robust view-invariant representations for downstream tasks, while DA-Net [65] merges view-specific and independent modules for effective prediction. A feature factorization approach in [66] and a cascaded residual autoencoder in [67] address challenges in RGB-D action recognition and incomplete view classification, respectively.

3 Proposed MM-WLAuslan Dataset

In this section, we describe our recording setup and workflow, detail the data processing and augmentation, and provide statistics for the MM-WLAuslan[6] dataset.

3.1 Recording Setup and Workflow

Our recording setup is located in a studio environment surrounded by a green screen. In the studio, we position Kinect-V2 cameras at the left-front, front, and right-front views, along with a centrly placed RealSense camera. Both Kinect-V2 and RealSense are capable of recording high-quality videos with depth information. In the Appendix, we compare the different parameters of these two types of cameras. Most importantly, the imaging principles of Kinect-V2 and RealSense cameras are different. The former employs time-of-flight technology to measure depth, while the latter utilizes stereo vision to capture depth information. Moreover, Kinect-V2 offers high resolution and excellent depth sensing, while RealSense provides a higher frame rate and portability. We record data using these two types of RGB-D cameras to investigate the cross-camera robustness of methods.

We recruit signers with diverse experience in Auslan, including Auslan experts, deaf individuals who use Auslan, and volunteers interested in sign language, to sign glosses[7]. The involvement of Auslan experts and deaf individuals ensures the precision of a subset of the data, which is crucial for precise research and applications of sign language. The extensive participation of volunteers enhances the

[6]Our dataset follows the copyright **Creative Commons BY-NC-SA 4.0** license ©. Please note that we obtain the consent of the signers before recording them.

[7]***Auslan experts*** refers to non-deaf individuals who are proficient in Australian Sign Language, while ***non-expert deaf signers*** only refers to deaf individuals who use Auslan.

Figure 2: **Demonstrations of test subsets.** "STU", "ITW", "SYN", and "TED" represent the studio set, in-the-wild set, synthetic background set and temporal disturbance set, respectively.

diversity of the signers, reflecting the natural variability in the deaf community. Moreover, we design an interactive interface for dataset recording and present the interface in the Appendix. We record videos of sign language imitated by volunteers. Each sign is supervised and checked by at least one expert to ensure the precision of the sign language expression.

3.2 Data Processing and Augmentation

After recording all the sign language videos, we notice that a significant portion of the footage consists of a green screen background. Therefore, we utilize the keypoints estimated by AlphaPose [68, 69, 70] to remove the background that is irrelevant to the sign language. We crop videos based on a fixed-size box that can cover every signer and align their eyes on the same horizontal level.

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, and temporal disturbance (TED) set. Each subset encompasses videos for all gloss vocabulary. The **STU set** includes consistent scene settings with the training set. In the **ITW set**, green screens are removed and replaced with dynamic or static backgrounds to simulate videos recorded in diverse environments, as shown in Figure 2. We utilize the Background Remover[8] to extract signers from videos and synthesize indoor and outdoor backgrounds in the **SYN set**. The **TED set** simulates potential recording time discrepancies in real-world scenarios by randomly adjusting video segments through removal or altering playback speed.

Overall, each data sample in our dataset includes: (1) RGB-D videos captured by a Kinect-V2 camera or a RealSense camera; (2) intrinsic and extrinsic parameters for the captured camera; (3) pose sequences corresponding to the RGB video; (4) gloss identities; (5) spoken English words or phrases corresponding to the gloss and (6) signer identities. These various views and modalities of sign language video samples can be further investigated for different word-level Auslan ISLR settings.

3.3 Data Statistics

We select 3,215 commonly used Auslan glosses, corresponding to over 7,900 English words or phrases. As illustrated in Figure 3(b), there are more than 2,000 glosses with multiple meanings, highlighting the contextual variability of sign language similar to natural languages. Additionally, these terms are finely categorized into 49 groups, including health, education, and others, as shown in Figure 3(d). The extensive vocabulary and semantic richness of MM-WLAuslan demonstrate its potential to advance sign language research and applications.

Table 2: **Key statistics of MM-WLAuslan dataset splits.** "BG" and "TP" represent background and temporal, respectively. "OOS" indicates the signers only occur in the test set.

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	✗	✗	✗	✗	✗	✓

After over 2,500 hours of recording, we capture 282,900 videos by 73 signers. Specifically, for 3,215 commonly used word-level Auslan glosses, we record every gloss 22 times utilizing 4 different cameras (3215×20×4). Unlike other datasets [4, 10], our dataset maintains a consistent number of videos per Auslan gloss, thereby establishing a uniform ISLR dataset. We split the samples of a gloss

[8] https://github.com/nadermx/backgroundremover

https://doi.org/10.52202/079017-2227

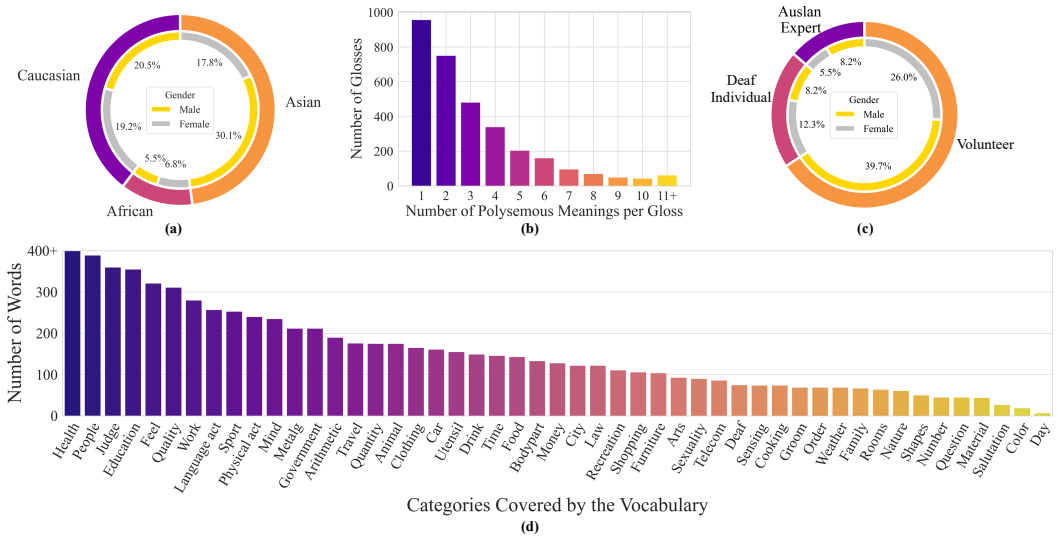

Figure 3: **Statistics of signers and glosses.** (a) Ethnicity and gender distribution. (b) Frequency of polysemous glosses. (c) Distribution of Auslan proficiency. (d) Categories of words.

into training, validation, and testing sets following a ratio of 6:1:4. Note that the test set contains 18 signers who do not appear in either the training or validation sets. Additionally, we further divide the testing set into the STU set, the ITW set, the SYN set, and the TED set in a 1:1:1:1 ratio. The detailed split statistics are demonstrated in Table 2.

Moreover, as illustrated in Figure 3(a), we provide the ethnic and gender distribution of signers in MM-WLAuslan. The signers are categorized into three primary ethnic groups: Caucasian, African, and Asian. The male-to-female ratios are relatively balanced across the different ethnic groups. The near-equitable gender balance within each ethnic group not only enhances the representativeness of the dataset but also underscores its gender fairness. Meanwhile, we include a broader range of ethnicities to enhance the inclusivity and representativeness of the dataset further. Thus, this composition ensures that the ISLR models developed from this dataset mitigate biases and offer equitable performance across the diverse Australian population. Furthermore, as shown in Figure 3(c), we demonstrate the distribution of participants involved in recording, segmented by their proficiency in Auslan. We make concerted efforts to include as many Auslan experts and deaf individuals as possible for the quality of the recordings. Additionally, we recruit many volunteers to further increase the diversity of the signers, and thus, enrich the representativeness of the dataset.

4 MM-WLAuslan Benchmark

In this section, we present and analyze benchmark results of various multi-modal ISLR settings on MM-WLAuslan. More experiments and details are included in the Appendix.

4.1 Isolated Sign Language Recognition Settings

Single-view RGB-based ISLR involves recognizing isolated sign language from video sequences captured from a single fixed camera. The input consists of RGB frames, denoted as $\{F_1, F_2, \ldots, F_n\}$, where n represents the total number of frames in a video sequence. The single-view RGB setting utilizes spatial and temporal information from a singular perspective.

Single-view RGB-D-based ISLR aims to enhance the recognition of isolated signs by incorporating depth information along with RGB data. The input is represented as $\{(F_1, D_1), (F_2, D_2), \ldots, (F_n, D_n)\}$, where D_i indicates the depth information corresponding to the i-th frame. This approach facilitates a richer interpretation of spatial dynamics.

Multi-view RGB-based ISLR employs multiple cameras to capture the sign language videos. The input from each camera k is represented as a sequence of RGB frames $\{F_1^k, F_2^k, \ldots, F_n^k\}$. Multi-view RGB data helps in mitigating issues related to occlusions and varied angles.

Table 3: **The baseline of Single-view RGB-based ISLR on MM-WLAuslan.** "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
ResNet2+1D [7]	Pixel	58.71	77.03	13.83	18.37	26.14	39.58	51.14	69.97	37.45	51.24
TSN [71]	Pixel	51.17	68.60	11.06	23.75	31.01	45.89	40.40	69.10	33.41	51.84
I3D [6]	Pixel	63.97	84.93	14.18	26.52	36.17	57.22	60.96	80.63	43.82	62.33
S3D [8]	Pixel	75.55	94.11	29.41	55.11	44.60	71.34	62.21	85.26	52.94	76.46
SlowFast [9]	Pixel	80.68	96.08	32.22	64.81	53.17	78.30	66.21	82.18	58.07	80.34
Timesformer [72]	Pixel	73.20	81.40	21.14	56.44	41.88	65.83	68.40	79.67	51.15	70.84
UMDR [73]	Pixel	80.86	95.88	13.57	28.66	13.99	31.01	**82.69**	**95.67**	47.78	62.81
KVNet-V [14]	Pixel	**84.51**	**97.57**	**39.88**	**68.00**	**56.56**	**82.18**	70.31	90.86	**62.82**	**84.65**
TGCN [4]	2D pose	68.62	86.30	58.01	74.74	63.50	81.38	47.68	68.82	62.11	77.81
SL-GCN [15]	2D pose	71.07	91.21	66.59	89.5	63.20	86.94	**69.98**	88.99	67.71	89.16
SPOTER [74]	2D pose	72.81	92.69	64.12	86.36	66.81	88.11	69.42	**90.94**	68.29	89.53
DSTA-SLR [50]	2D pose	82.33	96.31	74.96	93.98	78.10	93.78	66.84	88.99	75.55	93.26
STC-SLR [51]	2D pose	79.92	95.88	74.35	93.92	76.02	93.50	63.11	87.33	73.35	92.65
KVNet-K [14]	2D pose	**82.88**	**96.70**	**76.29**	**94.56**	**79.07**	**94.07**	69.05	89.80	**76.82**	**93.78**
SAM-SLR [15]	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 [14]	2D pose + Pixel	**86.32**	**97.79**	**79.05**	**94.91**	**84.26**	**96.16**	**77.98**	**91.76**	**81.90**	**95.16**

Table 4: **The baseline of Single-view RGB-D-based ISLR on MM-WLAuslan.**

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 [6]	Pixel + Depth	65.74	88.57	21.71	41.32	61.06	85.41	47.25	65.71	48.94	70.25
S3D [8]	Pixel + Depth	79.70	95.93	64.97	89.16	76.38	92.67	66.11	88.62	71.79	91.60
KVNet-V [14]	Pixel + Depth	82.22	96.75	38.79	66.11	57.88	82.92	66.94	88.58	61.46	83.59
UMDR [73]	Pixel + Depth	**91.65**	**98.81**	**72.52**	**90.46**	**83.77**	**95.18**	**88.35**	**98.07**	**84.07**	**95.63**
TGCN [4]	3D pose	70.19	89.78	59.52	76.59	66.35	84.06	51.48	71.17	61.88	80.40
SPOTER [74]	3D pose	74.95	95.88	66.75	89.41	70.22	91.23	71.65	92.36	70.89	92.22
SL-GCN [15]	3D pose	**77.76**	**96.98**	**72.26**	**91.49**	**74.91**	**92.57**	**72.27**	**94.88**	**74.30**	**93.98**
NLA-SLR [14]	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 [15]	3D pose + Pixel + Depth	**87.05**	**98.93**	**81.29**	**96.92**	83.03	**95.86**	**85.07**	93.53	**84.11**	**96.31**

Multi-view RGB-D-based ISLR incorporates depth data in a multi-view setup, the input for each camera k is represented as $\{(F_1^k, D_1^k), (F_2^k, D_2^k), \ldots, (F_n^k, D_n^k)\}$. This method enhances the model's capability to interpret complex gestures from multiple perspectives.

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.

Cross-View ISLR requires the model to recognize signs from views not seen during training. With training views denoted as V_{train} and testing views as V_{test}, the model must handle the appearance changes due to different viewing angles, thus testing its view-invariance capabilities.

4.2 Evaluation Metric

Top-*k* Accuracy is quantitatively defined as the proportion of test instances for which the true label is among the top *k* predictions made by the model. It is particularly suitable for ISLR [4, 10, 5] task with a large set of possible outcomes. The expression is shown by the following equation:

$$\text{Top-}k \text{ Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(y_i \in \hat{Y}_i^k), \tag{1}$$

where N is the total number of instances in the test set, $\mathbf{1}$ is a binary indicator that returns 1 if the true label of the i-th instance y_i is within the set of the top-k predicted labels $\hat{Y}_i^k$ for that instance.

4.3 Benchmark Results

All single-view experiments in this section are conducted on the data captured by front Kinect-V2.

https://doi.org/10.52202/079017-2227

Table 5: **The baseline of Multi-view RGB-based ISLR on MM-WLAuslan.** "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 [73]	Pixel	**92.56**	**99.09**	23.78	44.22	22.12	42.61	**90.13**	**98.23**	57.15	71.04
KVNet-V [14]	Pixel	91.57	99.00	**62.25**	**86.19**	**70.90**	**90.97**	79.78	94.68	**76.13**	**92.71**
SPOTER [74]	2D pose	76.92	95.55	67.79	89.98	69.21	92.16	74.34	**94.14**	72.06	92.96
DSTA-SLR [50]	2D pose	**91.68**	97.22	**87.06**	95.86	85.67	**96.34**	**79.15**	92.14	**85.89**	95.39
STC-SLR [15]	2D pose	90.11	96.28	86.82	94.91	**86.09**	96.29	75.13	90.76	84.53	94.56
KVNet-K [14]	2D pose	90.45	**98.56**	86.23	**97.77**	85.73	95.47	77.26	93.93	84.92	**96.43**
SAM-SLR [15]	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 [14]	2D pose + Pixel	**94.62**	**99.31**	**89.75**	**98.60**	**88.94**	**96.98**	**85.19**	**96.69**	**89.63**	**97.90**

Table 6: **The baseline of Multi-view RGB-D-based ISLR on MM-WLAuslan.**

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 [73]	Pixel + Depth	**93.25**	**99.11**	**74.98**	**92.19**	**86.14**	**96.24**	**90.42**	**97.39**	**86.20**	**96.36**
KVNet-V [14]	Pixel + Depth	87.67	98.22	66.01	88.80	83.06	95.27	74.23	92.28	77.74	93.64
SPOTER [74]	3D pose	79.91	**96.91**	73.44	91.29	**76.41**	**93.58**	76.87	94.45	76.66	94.06
ST-GCN [47]	3D pose	**81.77**	95.07	**77.34**	**93.13**	76.38	92.83	**79.36**	**96.73**	**78.71**	**94.44**
SAM-SLR [15]	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 [14]	2D pose + Pixel + Depth	**94.43**	**99.37**	**88.95**	**98.49**	**89.52**	**97.14**	**85.13**	**96.46**	**89.51**	**97.87**

Single-view RGB-based ISLR: Following previous works [4, 10, 18], we adopt this setting as a central focus of ISLR research. We utilize publicly available ISLR models, such as KVNet [14], SPOTER [74], DSTA-SLR [50], STC-SLR [51], SAM-SLR [15] and NLA-SLR [14]. Meanwhile, we incorporate models that have demonstrated strong performance in action recognition, including I3D [6], SlowFast [9] and Timesformer [72]. As indicated by Table 3, pixel-based models perform well in controlled STU. This suggests that pixel models are effective in settings with minimal noise and well-defined conditions. Conversely, pose-based models are robust in challenging environments, like ITW and SYN, because they focus on structural rather than textural information. Furthermore, NLA-SLR [14] is the SOTA model for ISLR. It ensembles the high-performance KVNet-V and KVNet-K models for pixel and pose data, respectively. The model demonstrates high accuracy across all test subsets consistently.

Single-view RGB-D-based ISLR: As shown in Table 4, the combination of pixel and depth data generally improves recognition accuracy on most methods, highlighting the benefits brought by depth data. However, the performance of the KVNet-V [14] model declines, indicating its insufficient processing of depth information alongside pixel data. In contrast, the UMDR [73] model, a SOTA model for RGB-D action recognition, leads to significant performance improvements across various test subsets. Additionally, pose-based models with 3D pose data as the input also show improved performance, further supporting the benefits of integrating depth information into pose-based models.

Multi-view RGB-based & RGB-D-based ISLR: In Table 5 and Table 6, we show performances of several RGB-based and RGB-D-based models on multi-view ISLR. The results highlight that using multiple views and additional modalities generally improves model performance. Models like UMDR and SAM-SLR, incorporating depth or 3D pose data, consistently achieve better results. This suggests these models effectively capture more comprehensive gesture information. However, these benefits come at the cost of increased model complexity. The introduction of multi-view RGBD data inevitably raises the training costs of the model. Additionally, information redundancy in the data can potentially interfere with the model's learning process. For instance, the recognition accuracy of the NLA-SLR model, when trained on multi-view RGBD data, is lower compared to when it is trained solely on RGB data. For future research, we focus on developing more efficient methods to optimize performance without increasing complexity for multi-view and multi-modal data.

Cross-camera ISLR: As illustrated in Table 7, there is a challenge in cross-camera ISLR on the MM-WL Auslan dataset. The results show a significant decline in accuracy when models trained on one type of camera are tested on the other one. Although two models, *i.e.*, KVNet-V [14] and

Table 7: **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-V [14]	*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 [73]	*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

Table 8: **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-V [14]	*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 [73]	*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

UMDR [73], perform well with data from the same camera, their performance drops across the cameras. This highlights the substantial differences between the two cameras, emphasizing the complexity of achieving robust ISLR across varied hardware. The challenge of cross-camera ISLR underscores the need for developing models that can better generalize on data from various cameras.

Cross-view ISLR: We report the performance of two models, *i.e.*, KVNet-V [14] and UMDR [73], training and evaluating on the data from different Kinect-v2 views, as shown in Table 8. UMDR, incorporating depth information alongside pixel data, generally exhibits greater resilience and performance compared to KVNet-V. Both models exhibit high accuracy under the single-view setting of our dataset, yet experience a significant drop in accuracy in the cross-view context. This indicates that models capable of adapting to diverse visual inputs are necessary to address the challenges posed by cross-view.

5 Limitation and Future Work

Limited Diversity in Data: As shown in Figure 3(a) of the main paper, we analyze the distribution of Caucasian, African, and Asian signers within the MM-WLAuslan dataset. We observe that the proportion of African signers is significantly lower than that of Caucasian and Asian signers. Consequently, the signers in our recordings do not fully represent the demographic diversity of the Auslan community. Australia, being a multi-cultural nation, encompasses a wide range of ethnicities, and the representation of these ethnicities in our dataset is crucial. Therefore, we will continue recording to achieve a more balanced representation.

Lack of Real-world Scenarios: Although we attempt to simulate real-life environments by altering backgrounds and capturing some data "in the wild", these settings still fall short of fully representing the complexities of real-world scenarios, such as multi-person interactions and intricate backgrounds. Moving forward, we intend to capture real-world Auslan glosses for a more authentic dataset. This initiative aims to more accurately reflect the dynamic and diverse contexts in which Auslan is naturally used, thereby improving the relevance and applicability of the dataset.

https://doi.org/10.52202/079017-2227

Existing Model Limitations: In this work, we utilize publicly available deep learning models, some of which are not specifically designed for sign language. Consequently, developing more effective multi-modal fusion and multi-view techniques tailored to the unique characteristics of our dataset is essential. This approach will enhance the accuracy and applicability of the models, ensuring they are better suited to address the specific challenges and nuances of isolated sign language recognition.

Investigating Isolated Sign Language Production Task: Sign Language Production is currently a popular task, involving not only the generation of isolated glosses [75] but also continuous sign language [76, 77]. Unlike previous datasets, ours incorporates multi-view and multi-modal capabilities, enabling the creation of more accurate 2D or 3D sign language representations. We plan to further explore this task using our dataset in future research. This will enhance the precision and effectiveness of sign language modelling, providing more robust tools for communication within the deaf and hard-of-hearing community.

6 Conclusion

In this work, we introduce the first large-scale, multi-view, multi-modal word-level dataset for Australian Sign Language (Auslan), named MM-WLAuslan. The dataset includes 282K+ videos encompassing 3,215 distinct Auslan glosses performed by 73 signers. To the best of our knowledge, MM-WLAuslan has the largest amount of data, the most extensive vocabulary, and the most diverse set of multi-modal camera views. We position four RGB-D cameras, *i.e.*, three Kinect-V2 cameras and a RealSense camera, hemispherically around the signers. Extensive experiments demonstrate the validity and challenges of MM-WLAuslan. Thanks to the cross-camera, multi-view, and multi-modal gloss videos, our dataset can be used for practical applications related with Auslan. Furthermore, the presented benchmark results can act as strong baselines for future research.

Acknowledgement

This research is funded in part by ARC-Discovery grant (DP220100800 to XY), ARC-DECRA grant (DE230100477 to XY) and Google Research Scholar Program. We express our gratitude to Professor Trevor Cohn for his valuable feedback on this work. We also gratefully thank all the anonymous reviewers and ACs for their constructive comments.

Broader Impact

The development of the word-level Australian Sign Language (Auslan) dataset has several impacts on technology, education, and society. Our proposed MM-WLAuslan, recorded using multi-view RGB-D cameras and focused on isolated Auslan glosses, brings about a wide range of positive effects:

- **Improveing Accuracy and Efficiency in ISLR**: The high-quality data provided by multi-view RGB-D cameras enhance the detailed capture of sign language gestures, which is crucial for developing efficient and accurate ISLR systems.
- **Facilitating Social Integration for the Deaf:** Improved by our MM-WLAuslan dataset, the ISLR technology can provide the deaf and hard-of-hearing community with more efficient communication capabilities.
- **Expanding Educational Resources:** Our dataset can support Auslan education [25, 78]. By providing multi-view demonstrations, the dataset allows Auslan learners to observe signs from different views, enhancing their understanding and accuracy in sign language.
- **Driving Technological Innovation:** Our dataset offers valuable resources for research in computer vision and machine learning, promoting technological development and innovation in these fields [79, 80, 81, 82].
- **Preserving and Promoting Culture:** By recording and utilizing the MM-WLAuslan dataset, we preserve and disseminate the unique cultural heritage of Australian Sign Language, enhancing public awareness of its cultural value [83].

These societal impacts demonstrate that the development and application of the Auslan dataset are not only technically significant but also have profound positive values on social and cultural levels.

References

[1] Natasha Abner, Carlo Geraci, Shi Yu, Jessica Lettieri, Justine Mertz, and Anah Salgat. Getting the upper hand on sign language families: Historical analysis and annotation methods. *FEAST. Formal and Experimental Advances in Sign language Theory*, 3:17–29, 2020.

[2] William C Stokoe Jr. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37, 2005.

[3] William C Stokoe. Sign language structure. *Annual review of anthropology*, 9(1):365–390, 1980.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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*, volume 11219 of *Lecture Notes in Computer Science*, pages 318–335. Springer, 2018.

[9] 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.

[10] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A large-scale data set and benchmark for understanding american sign language. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 100. BMVA Press, 2019.

[11] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Attention-based 3d-cnns for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2822–2832, 2018.

[12] Kinect gesture dataset. `https://www.microsoft.com/en-us/download/details.aspx?id=52283`, 2019. Accessed: 2019-07-16.

[13] Franco Ronchetti, Facundo Manuel Quiroga, César Estrebou, Laura Lanzarini, and Alejandro Rosete. Lsa64: an argentinian sign language dataset. *arXiv preprint arXiv:2310.17429*, 2023.

[14] 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.

[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] Aleix M. Martínez, Ronnie B. Wilbur, Robin Shay, and Avinash C. Kak. Purdue RVL-SLLL ASL database for automatic recognition of american sign language. In *4th IEEE International Conference on Multimodal Interfaces (ICMI 2002), 14-16 October 2002, Pittsburgh, PA, USA*, pages 167–172. IEEE Computer Society, 2002.

 https://doi.org/10.52202/079017-2227

[17] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan P. Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. The american sign language lexicon video dataset. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2008, Anchorage, AK, USA, 23-28 June, 2008*, pages 1–8. IEEE Computer Society, 2008.

[18] Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, and et al. Popsign ASL v1.0: An isolated american sign language dataset collected via smartphones. 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.

[19] Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard E. Ladner, Hal Daumé III, Alex X. Lu, Naomi Caselli, and Danielle Bragg. ASL citizen: A community-sourced dataset for advancing isolated sign language recognition. 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.

[20] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: scaling up co-articulated sign language recognition using mouthing cues. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 35–53. Springer, 2020.

[21] Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Watch, read and lookup: learning to spot signs from multiple supervisors. In *ACCV*, 2020.

[22] 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.

[23] Xiujuan Chai, Guang Li, Yushun Lin, Zhihao Xu, Yili Tang, Xilin Chen, and Ming Zhou. Sign language recognition and translation with kinect. In *IEEE conf. on AFGR*, volume 655, page 4, 2013.

[24] Sarah Ebling, Necati Cihan Camgöz, Penny Boyes Braem, Katja Tissi, Sandra Sidler-Miserez, and et al. SMILE swiss german sign language dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018.

[25] 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.

[26] Chenchen Xu, Dongxu Li, Hongdong Li, Hanna Suominen, and Ben Swift. Automatic gloss dictionary for sign language learners. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 83–92. Association for Computational Linguistics, 2022.

[27] Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In Walter G. Kropatsch, Robert Sablatnig, and Allan Hanbury, editors, *Pattern Recognition, 27th DAGM Symposium, Vienna, Austria, August 31 - September 2, 2005, Proceedings*, volume 3663 of *Lecture Notes in Computer Science*, pages 401–408. Springer, 2005.

[28] Suneetha Mopidevi, M. V. D. Prasad, and Polurie Venkata Vijay Kishore. Multiview meta-metric learning for sign language recognition using triplet loss embeddings. *Pattern Anal. Appl.*, 26(3):1125–1141, 2023.

[29] Steve Cassidy, Onno Crasborn, Henri Nieminen, Wessel Stoop, Micha Hulsbosch, Susan Even, Erwin Komen, and Trevor Johnson. Signbank: Software to support web based dictionaries of sign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018.

[30] Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and Richard Bowden. Sign language recognition using sequential pattern trees. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2200–2207. IEEE Computer Society, 2012.

[31] Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. Include: A large scale dataset for indian sign language recognition. MM '20. Association for Computing Machinery, 2020.

[32] Eva Gutierrez-Sigut, Brendan Costello, Cristina Baus, and Manuel Carreiras. Lse-sign: A lexical database for spanish sign language. *Behavior Research Methods*, 48:123–137, 2016.

[33] Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. LSFB-CONT and LSFB-ISOL: two new datasets for vision-based sign language recognition. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE, 2021.

[34] Ogulcan Özdemir, Ahmet Alp Kindiroglu, Necati Cihan Camgöz, and Lale Akarun. Bosphorussign22k sign language recognition dataset. *CoRR*, abs/2004.01283, 2020.

[35] Ozge Mercanoglu Sincan and Hacer Yalim Keles. AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020.

[36] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1610–1618. IEEE Computer Society, 2017.

[37] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Trans. Multim.*, 21(7):1880–1891, 2019.

[38] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2306–2320, 2019.

[39] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In Kien A. Hua, Yong Rui, Ralf Steinmetz, Alan Hanjalic, Apostol Natsev, and Wenwu Zhu, editors, *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, pages 675–678. ACM, 2014.

[40] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1725–1732. IEEE Computer Society, 2014.

[41] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4489–4497. IEEE Computer Society, 2015.

[42] Al Amin Hosain, Panneer Selvam Santhalingam, Parth H. Pathak, Huzefa Rangwala, and Jana Kosecká. Hand pose guided 3d pooling for word-level sign language recognition. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 3428–3438. IEEE, 2021.

[43] Hongyu Fu, Chen Liu, Xingqun Qi, Beibei Lin, Lincheng Li, Li Zhang, and Xin Yu. Sign spotting via multi-modal fusion and testing time transferring. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13808 of *Lecture Notes in Computer Science*, pages 271–287. Springer, 2022.

[44] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6204–6213. Computer Vision Foundation / IEEE, 2020.

[45] Philippe Weinzaepfel, Zaïd Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3164–3172. IEEE Computer Society, 2015.

[46] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. 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 I*, volume 11205 of *Lecture Notes in Computer Science*, pages 106–121. Springer, 2018.

[47] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 7444–7452. AAAI Press, 2018.

 https://doi.org/10.52202/079017-2227

[48] Shannan Guan, Xin Yu, Wei Huang, Gengfa Fang, and Haiyan Lu. DMMG: dual min-max games for self-supervised skeleton-based action recognition. *IEEE Trans. Image Process.*, 33:395–407, 2024.

[49] Anirudh Tunga, Sai Vidyaranya Nuthalapati, and Juan P. Wachs. Pose-based sign language recognition using GCN and BERT. In *IEEE Winter Conference on Applications of Computer Vision Workshops, WACV Workshops 2021, Waikoloa, HI, USA, January 5-9, 2021*, pages 31–40. IEEE, 2021.

[50] 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.

[51] 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.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[53] Matyás Bohácek and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops, Waikoloa, HI, USA, January 4-8, 2022*, pages 182–191. IEEE, 2022.

[54] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11087–11096, 2021.

[55] Taeryung Lee, Yeonguk Oh, and Kyoung Mu Lee. Human part-wise 3d motion context learning for sign language recognition. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20683–20693. IEEE, 2023.

[56] Noha Sarhan and Simone Frintrop. Transfer learning for videos: from action recognition to sign language recognition. In *2020 IEEE International Conference on Image Processing*, pages 1811–1815. IEEE, 2020.

[57] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1077–1086, 2019.

[58] Zan Gao, Hai-Zhen Xuan, Hua Zhang, Shaohua Wan, and Kim-Kwang Raymond Choo. Adaptive fusion and category-level dictionary learning model for multiview human action recognition. *IEEE Internet of Things Journal*, 6(6):9280–9293, 2019.

[59] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1963–1978, 2019.

[60] Tong Hao, Dan Wu, Qian Wang, and Jin-Sheng Sun. Multi-view representation learning for multi-view action recognition. *Journal of Visual Communication and Image Representation*, 48:453–460, 2017.

[61] Yisheng Zhu and Guangcan Liu. Fine-grained action recognition using multi-view attentions. *Vis. Comput.*, 36(9):1771–1781, 2020.

[62] Kaijun Zhu, Ruxin Wang, Qingsong Zhao, Jun Cheng, and Dapeng Tao. A cuboid cnn model with an attention mechanism for skeleton-based action recognition. *IEEE Transactions on Multimedia*, 22(11):2977–2989, 2019.

[63] Jingjing Zheng, Zhuolin Jiang, and Rama Chellappa. Cross-view action recognition via transferable dictionary learning. *IEEE Transactions on Image Processing*, 25(6):2542–2556, 2016.

[64] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.*, 103(1):60–79, 2013.

[65] Dongang Wang, Wanli Ouyang, Wen Li, and Dong Xu. Dividing and aggregating network for multi-view action recognition. In *Proceedings of the European conference on computer vision*, pages 451–467, 2018.

[66] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep multimodal feature analysis for action recognition in rgb+ d videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1045–1058, 2017.

[67] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414, 2017.

[68] 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.

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

[70] 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.

[71] 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.

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

[73] 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.

[74] 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.

[75] Rotem Shalev-Arkushin, Amit Moryossef, and Ohad Fried. Ham2pose: Animating sign language notation into pose sequences. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 21046–21056. IEEE, 2023.

[76] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5131–5141. IEEE, 2022.

[77] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 687–705, 2020.

[78] Chenchen Xu, Dongxu Li, Hongdong Li, Hanna Suominen, and Ben Swift. Automatic gloss dictionary for sign language learners. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 83–92. Association for Computational Linguistics, 2022.

[79] 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.

[80] 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.

[81] Maria Zelenskaya, Scott Whittington, Julie Lyons, Adele Vogel, and Jessica Korte. Visual-gestural interface for auslan virtual assistant. In June Kim, Miu Ling Lam, and Kouta Minamizawa, editors, *SIGGRAPH Asia 2023 Emerging Technologies, Sydney, NSW, Australia, December 12-15, 2023*, pages 21:1–21:2, 2023.

[82] 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.

[83] River Tae Smith, Louisa Willoughby, and Trevor Johnston. Integrating Auslan resources into the language data commons of Australia. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 181–186, Marseille, France, June 2022. European Language Resources Association.

 https://doi.org/10.52202/079017-2227

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work does not pose any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [No] We do not have theoretical results.
   (b) Did you include complete proofs of all theoretical results? [No] We do not have theoretical results.
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All datasets and benchmarks are available at MM-WLAuslan.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 3.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Appendix Section C.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite the papers of the model.
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]