WED-PM-242

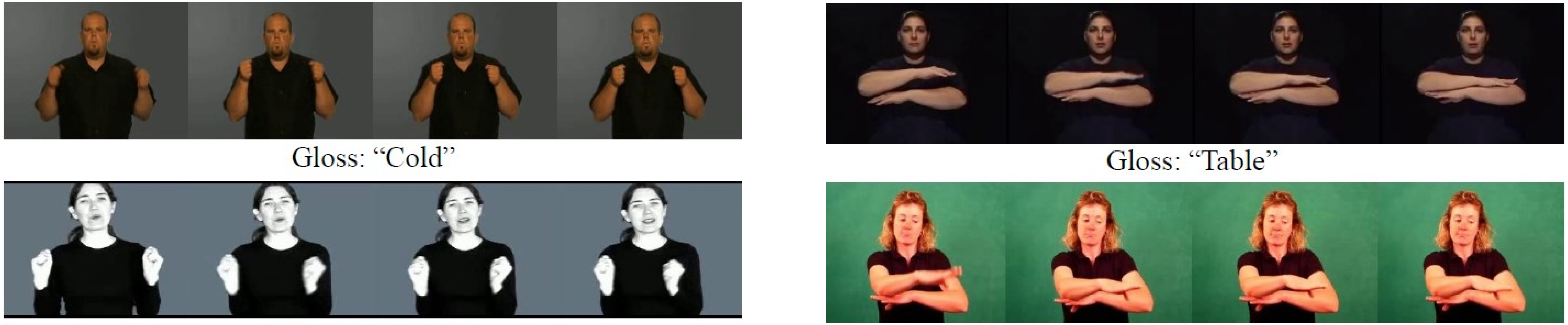
Natural Language-Assisted Sign Language Recognition

Ronglai Zuo[1] Fangyun Wei[2†] Brian Mak[1]

[1]The Hong Kong University of Science and Technology [2]Microsoft Research Asia

Introduction

- Primary communication method among deaf communities.
- Sign language recognition (SLR) is the most basic task.
- Visually indistinguishable signs (VISigns)



Gloss: “Cold”

Gloss: “Winter”

(a) VISigns may have *similar* semantic meanings.

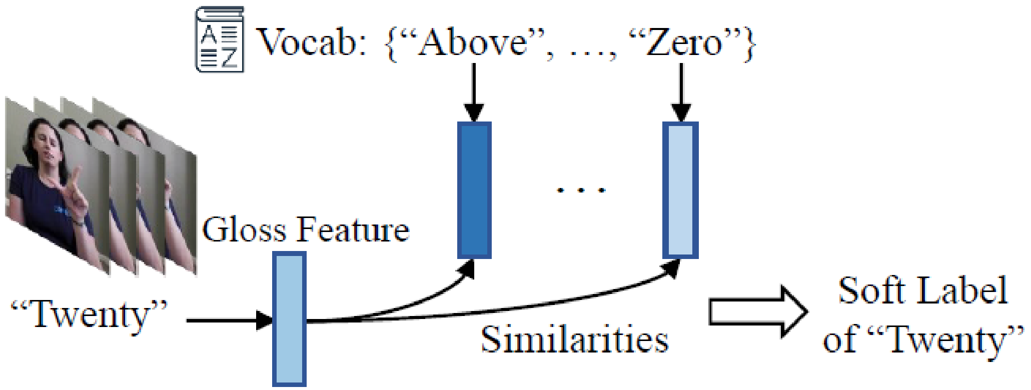
Gloss: “Table”

Gloss: “Afternoon”

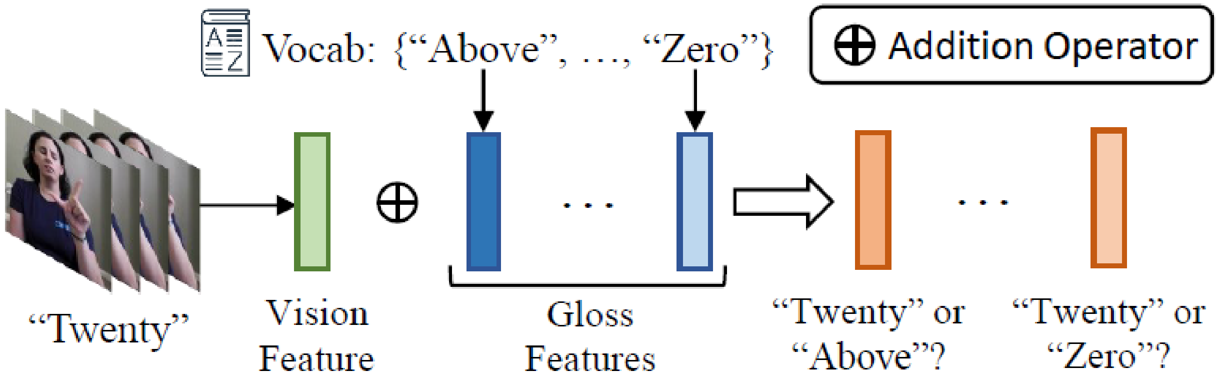
(b) VISigns may have *distinct* semantic meanings.

Introduction

- Two techniques to address the two types of VISigns

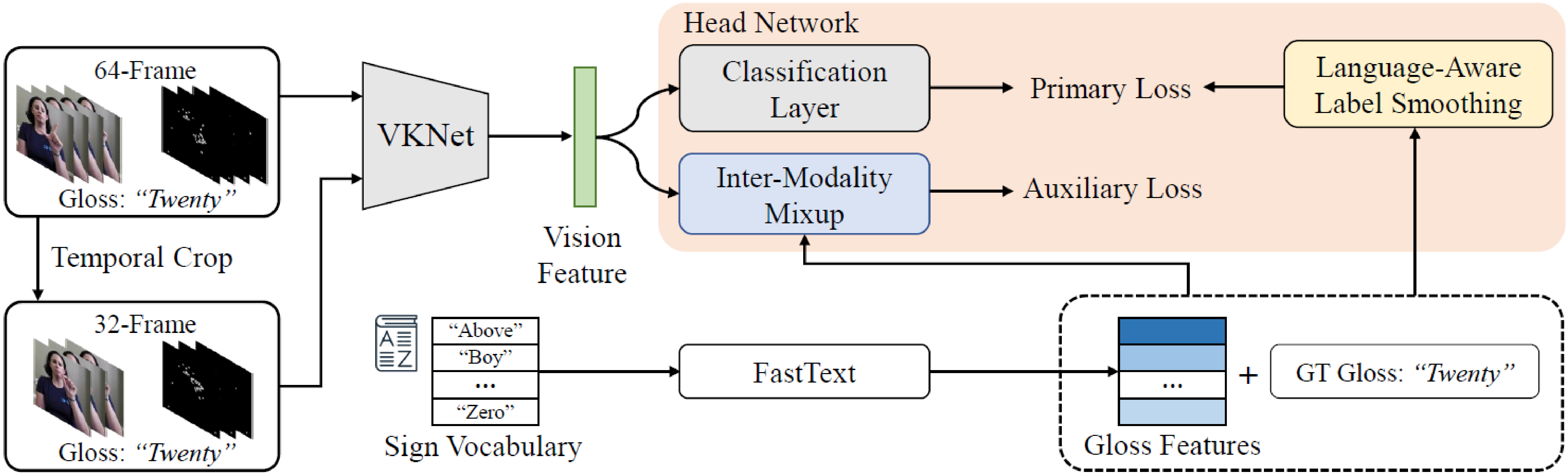


(a) Language-aware label smoothing.

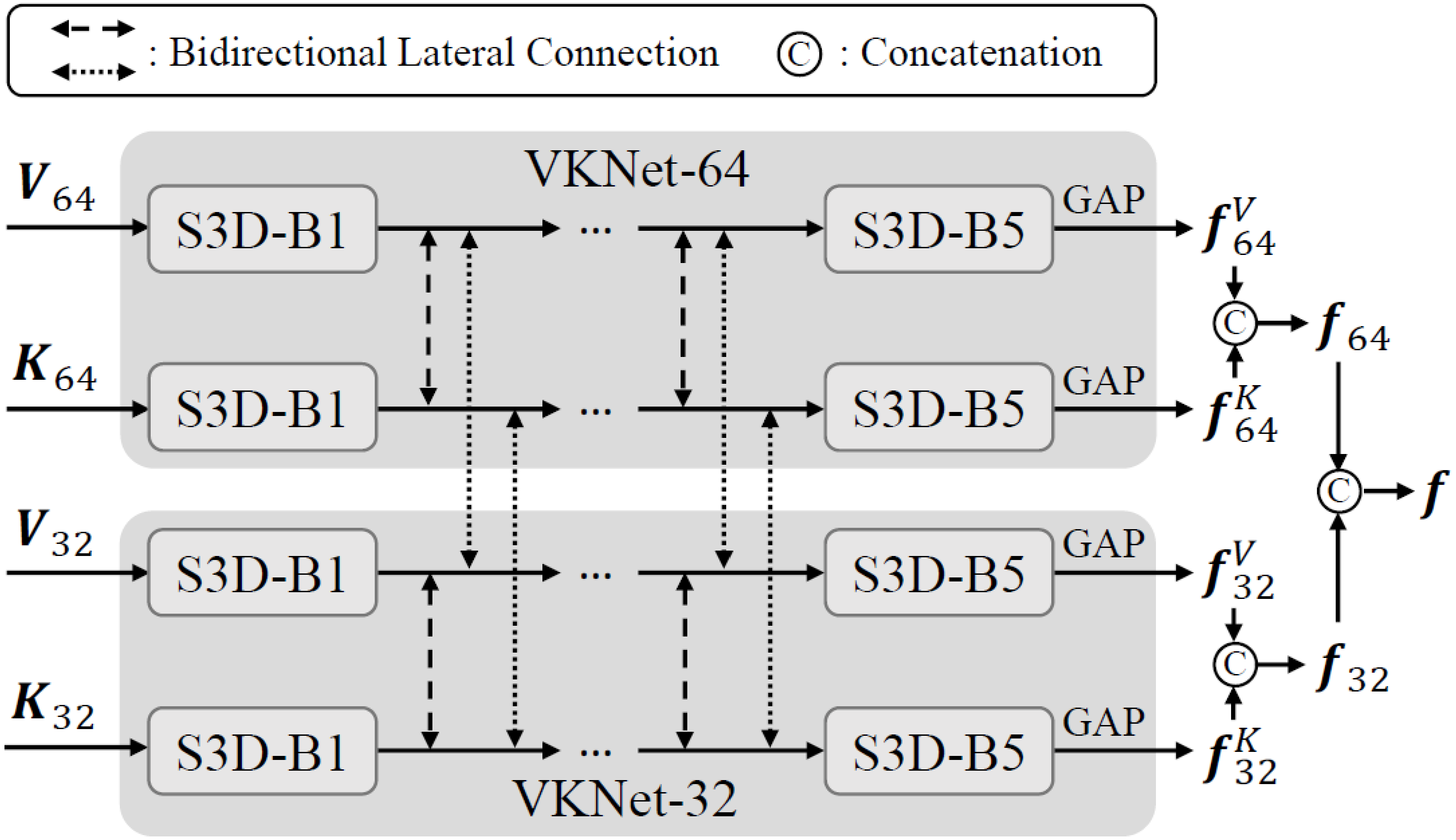


(b) Inter-modality mixup.

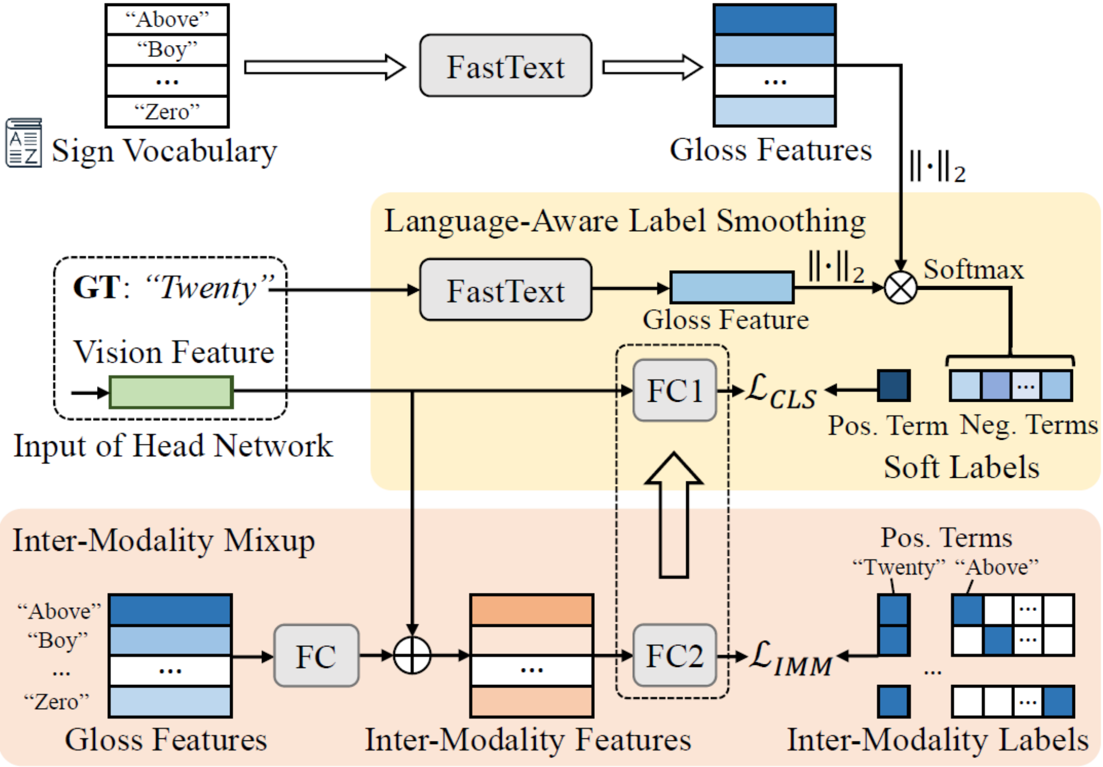
Framework Overview



Video-Keypoint Network (VKNet)



Head Network



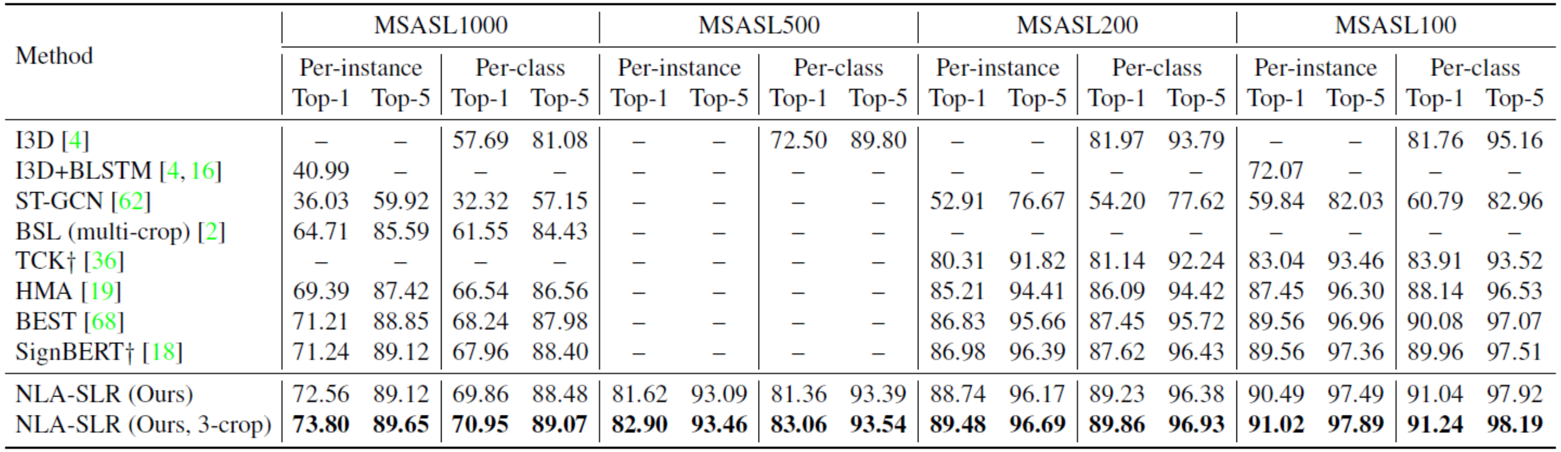
$$\boldsymbol{y}[i] = \begin{cases} 1-\epsilon & \text{if } i = b, \\ \epsilon \cdot \frac{\exp\left(\boldsymbol{s}[i]/\tau\right)}{\sum_{i=1, i\neq b}^{N} \exp\left(\boldsymbol{s}[i]/\tau\right)} & \text{otherwise} \end{cases}$$

Language-aware Label Smoothing

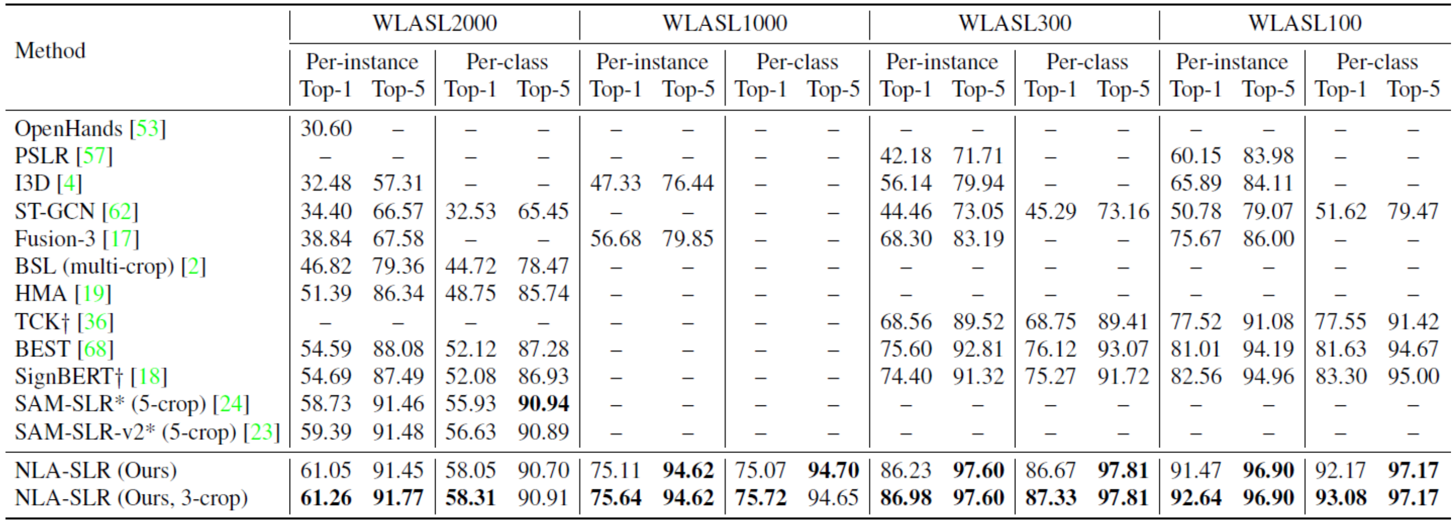
$$\theta_1, \theta_2 \leftarrow \text{optimizer}(\theta_1, \theta_2, \nabla_{\theta_1}\mathcal{L}, \nabla_{\theta_2}\mathcal{L}, \eta)$$
$$\theta_1 \leftarrow \mu\theta_1 + (1-\mu)\theta_2,$$

Classifier Integration

SOTA Performance

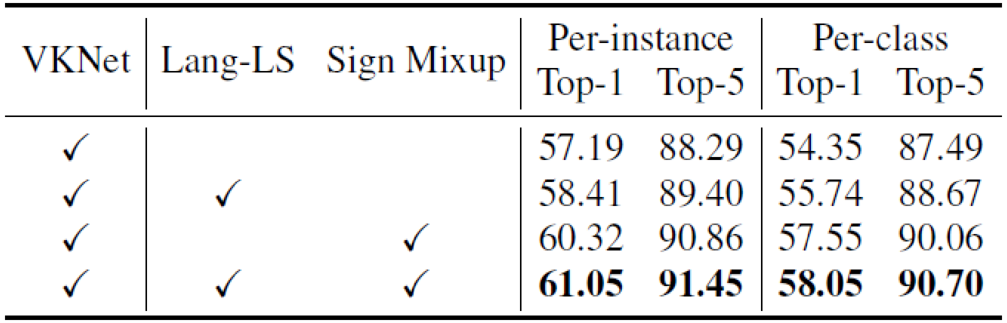


Method	MSASL1000				MSASL500				MSASL200				MSASL100			
	Per-instance		Per-class		Per-instance		Per-class		Per-instance		Per-class		Per-instance		Per-class	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
I3D [4]	–	–	57.69	81.08	–	–	72.50	89.80	–	–	81.97	93.79	–	–	81.76	95.16
I3D+BLSTM [4, 16]	40.99	–	–	–	–	–	–	–	–	–	–	–	72.07	–	–	–
ST-GCN [62]	36.03	59.92	32.32	57.15	–	–	–	–	52.91	76.67	54.20	77.62	59.84	82.03	60.79	82.96
BSL (multi-crop) [2]	64.71	85.59	61.55	84.43	–	–	–	–	–	–	–	–	–	–	–	–
TCK† [36]	–	–	–	–	–	–	–	–	80.31	91.82	81.14	92.24	83.04	93.46	83.91	93.52
HMA [19]	69.39	87.42	66.54	86.56	–	–	–	–	85.21	94.41	86.09	94.42	87.45	96.30	88.14	96.53
BEST [68]	71.21	88.85	68.24	87.98	–	–	–	–	86.83	95.66	87.45	95.72	89.56	96.96	90.08	97.07
SignBERT† [18]	71.24	89.12	67.96	88.40	–	–	–	–	86.98	96.39	87.62	96.43	89.56	97.36	89.96	97.51
NLA-SLR (Ours)	72.56	89.12	69.86	88.48	81.62	93.09	81.36	93.39	88.74	96.17	89.23	96.38	90.49	97.49	91.04	97.92
NLA-SLR (Ours, 3-crop)	**73.80**	**89.65**	**70.95**	**89.07**	**82.90**	**93.46**	**83.06**	**93.54**	**89.48**	**96.69**	**89.86**	**96.93**	**91.02**	**97.89**	**91.24**	**98.19**



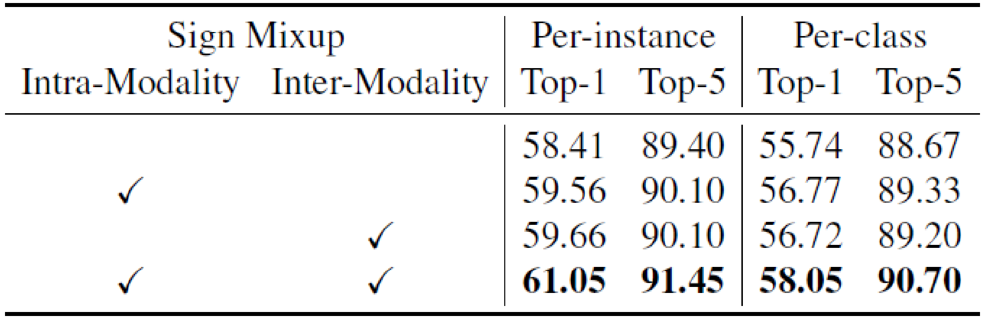
Method	WLASL2000				WLASL1000				WLASL300				WLASL100			
	Per-instance		Per-class		Per-instance		Per-class		Per-instance		Per-class		Per-instance		Per-class	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
OpenHands [53]	30.60	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
PSLR [57]	–	–	–	–	–	–	–	–	42.18	71.71	–	–	60.15	83.98	–	–
I3D [4]	32.48	57.31	–	–	47.33	76.44	–	–	56.14	79.94	–	–	65.89	84.11	–	–
ST-GCN [62]	34.40	66.57	32.53	65.45	–	–	–	–	44.46	73.05	45.29	73.16	50.78	79.07	51.62	79.47
Fusion-3 [17]	38.84	67.58	–	–	56.68	79.85	–	–	68.30	83.19	–	–	75.67	86.00	–	–
BSL (multi-crop) [2]	46.82	79.36	44.72	78.47	–	–	–	–	–	–	–	–	–	–	–	–
HMA [19]	51.39	86.34	48.75	85.74	–	–	–	–	–	–	–	–	–	–	–	–
TCK† [36]	–	–	–	–	–	–	–	–	68.56	89.52	68.75	89.41	77.52	91.08	77.55	91.42
BEST [68]	54.59	88.08	52.12	87.28	–	–	–	–	75.60	92.81	76.12	93.07	81.01	94.19	81.63	94.67
SignBERT† [18]	54.69	87.49	52.08	86.93	–	–	–	–	74.40	91.32	75.27	91.72	82.56	94.96	83.30	95.00
SAM-SLR* (5-crop) [24]	58.73	91.46	55.93	**90.94**	–	–	–	–	–	–	–	–	–	–	–	–
SAM-SLR-v2* (5-crop) [23]	59.39	91.48	56.63	90.89	–	–	–	–	–	–	–	–	–	–	–	–
NLA-SLR (Ours)	61.05	91.45	58.05	90.70	75.11	**94.62**	75.07	**94.70**	86.23	**97.60**	86.67	**97.81**	91.47	**96.90**	92.17	**97.17**
NLA-SLR (Ours, 3-crop)	**61.26**	**91.77**	**58.31**	90.91	**75.64**	**94.62**	**75.72**	94.65	**86.98**	**97.60**	**87.33**	**97.81**	**92.64**	**96.90**	**93.08**	**97.17**

Ablation Studies



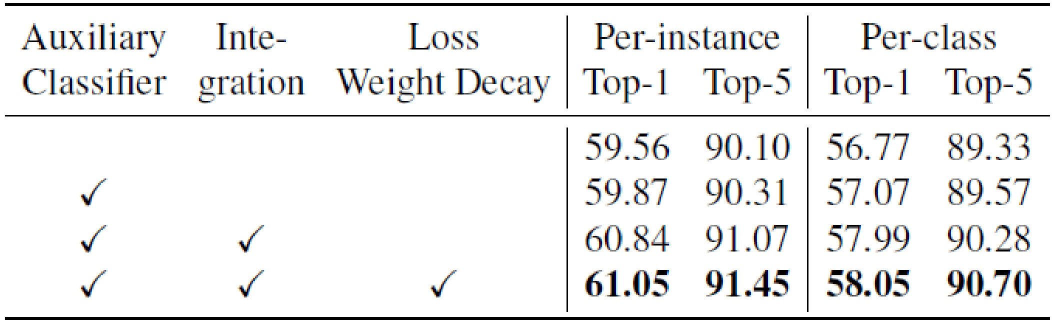
VKNet	Lang-LS	Sign Mixup	Per-instance Top-1	Per-instance Top-5	Per-class Top-1	Per-class Top-5
✓			57.19	88.29	54.35	87.49
✓	✓		58.41	89.40	55.74	88.67
✓		✓	60.32	90.86	57.55	90.06
✓	✓	✓	**61.05**	**91.45**	**58.05**	**90.70**

Major Components



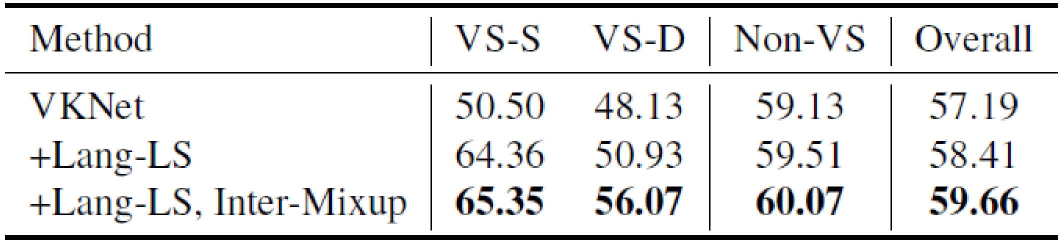
Sign Mixup Intra-Modality	Sign Mixup Inter-Modality	Per-instance Top-1	Per-instance Top-5	Per-class Top-1	Per-class Top-5
		58.41	89.40	55.74	88.67
✓		59.56	90.10	56.77	89.33
	✓	59.66	90.10	56.72	89.20
✓	✓	**61.05**	**91.45**	**58.05**	**90.70**

Sign mixup



Auxiliary Classifier	Integration	Loss Weight Decay	Per-instance Top-1	Per-instance Top-5	Per-class Top-1	Per-class Top-5
			59.56	90.10	56.77	89.33
✓			59.87	90.31	57.07	89.57
✓	✓		60.84	91.07	57.99	90.28
✓	✓	✓	**61.05**	**91.45**	**58.05**	**90.70**

Inter-modality mixup



Method	VS-S	VS-D	Non-VS	Overall
VKNet	50.50	48.13	59.13	57.19
+Lang-LS	64.36	50.93	59.51	58.41
+Lang-LS, Inter-Mixup	**65.35**	**56.07**	**60.07**	**59.66**

Quantitative results on VISigns

Thank you for your listening!