華中科技大學 HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY · 南昌大学 NANCHANG UNIVERSITY · 浙江大学 ZHEJIANG UNIVERSITY · 香港中文大學（深圳） The Chinese University of Hong Kong, Shenzhen

Explore More Guidance: A Task-aware Instruction Network for Sign Language Translation Enhanced with Data Augmentation

Yong Cao[1]*, Wei Li[2]*, Xianzhi Li[1]*, Min Chen[1#], Guangyong Chen[3], Long Hu[1], Zhengdao Li[4] and Hwang Kai[4]

[1]Huazhong University of Science and Technology, [2]Nanchang University, [3]Zhejiang University, [4]The Chinese University of Hong Kong, Shenzhen

Introduction

Task definition:

- Sign language recognition and translation first uses a recognition module to generate glosses from sign language videos and then employs a translation module to translate glosses into spoken sentences.
- In our work, we focus on sign language translation (SLT).

Contribution:

- We present a novel TIN-SLT network.
- A learning-based feature fusion strategy is proposed.
- A multi-level augmentation scheme is designed.
- SOTA Results are obtained.

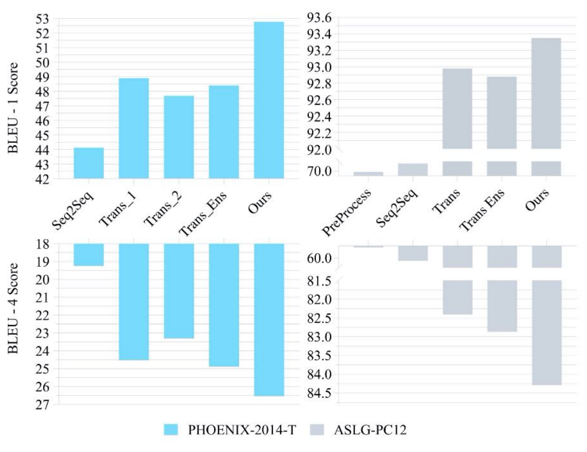

Figure 1: Comparing the sign language translation performance on two challenging datasets, i.e., PHOENIX-2014-T (blue) and ASLG-PC12 (gray), in terms of BLEU-1 and BLEU-4 metrics. Clearly, our approach achieves the highest scores on both datasets compared with others. The experiments section contains more results and analysis.

Challenges

Challenge 1: Limited annotated corpus

- The data resources of sign languages are scarce, thus the SLT models often suffer from overfitting with poor generalization.

Challenge 2: Discrepancy between glosses and texts

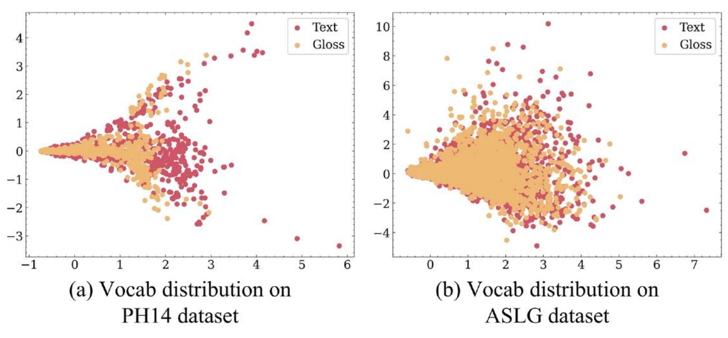

Figure 2: Comparing the sample distribution between the input sign glosses (yellow dots) and the output translated texts (red dots) on two datasets.

- The representation space of sign glosses is clearly smaller than that of the target spoken language, thus increasing the difficulty of network learning.

Method

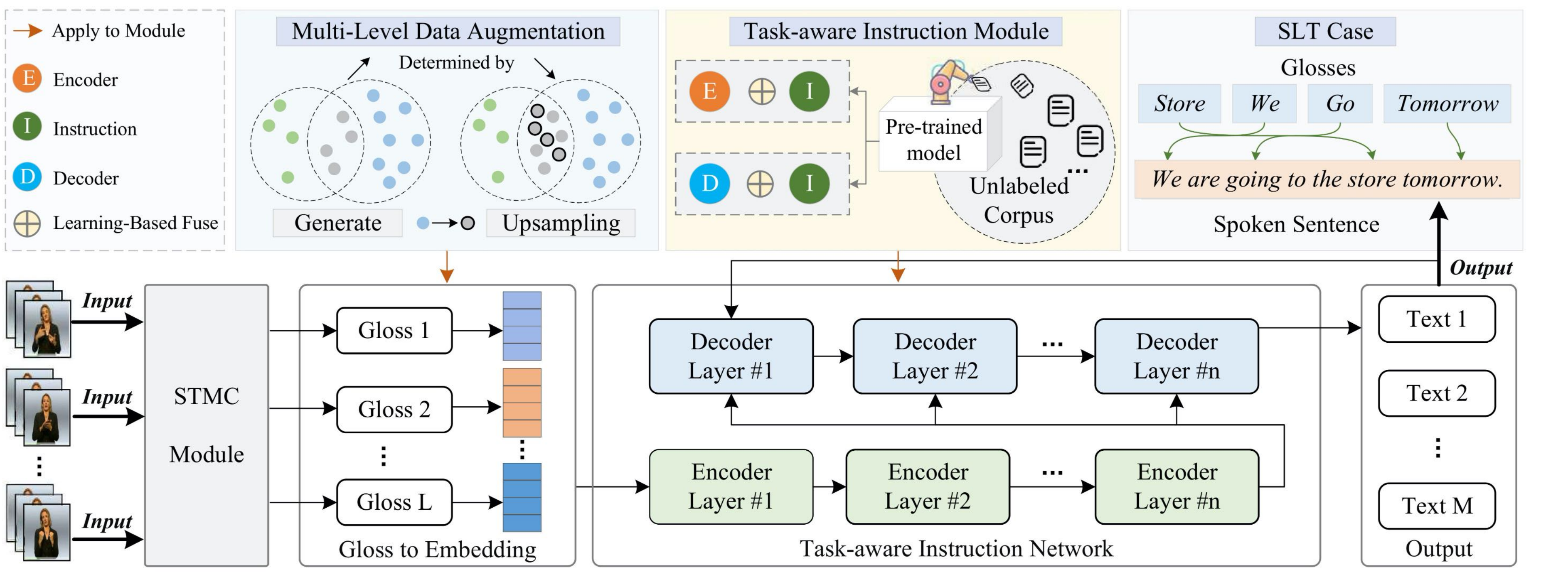

Figure 3: Network architecture of TIN-SLT. As shown in the bottom row, we first employ STMC model (Zhou et al., 2020) to recognize sign language videos to independent glosses. Next, we design a multi-level data augmentation scheme to enrich existing data pool for better feature embedding from glosses. Then, we design a task-aware instruction network with a novel instruction module to translate glosses into a complete spoken sentence.

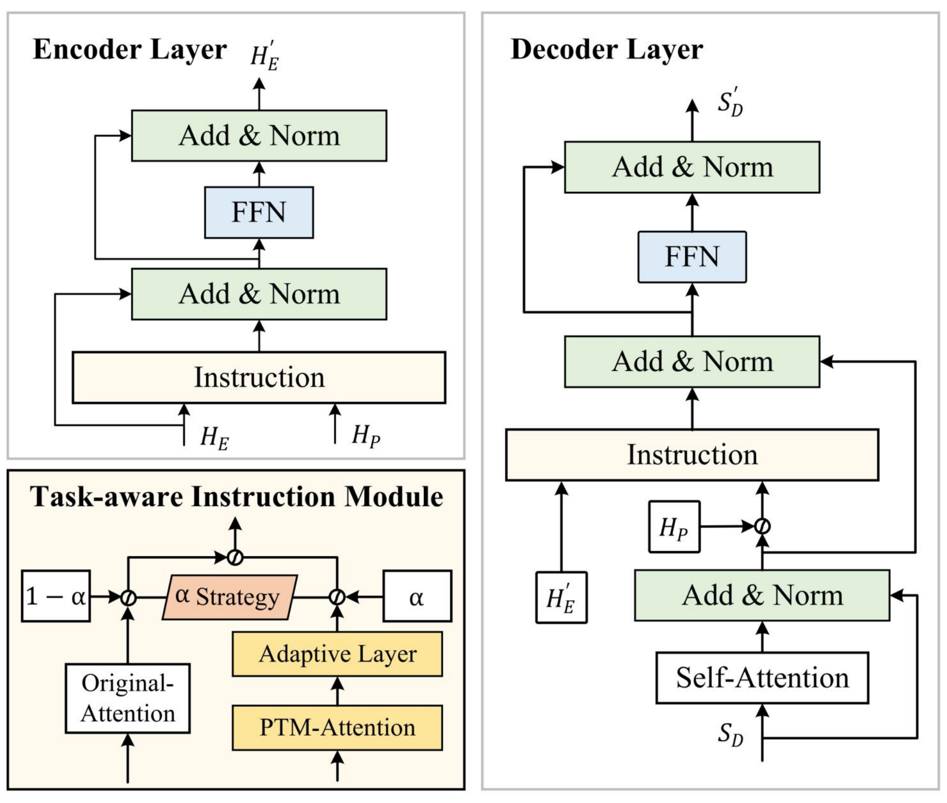

Figure 4: Details of Encoder layer, Decoder layer, and Task-aware Instruction Module.

➢ Sign language recognition

- We empirically adopt the spatial-temporal multi-cue (STMC) network to recognize sign language videos to independent glosses.

➢ Multi-level data augmentation scheme

- Token Level $\phi_v = 1 - \frac{|W_\mathcal{G}|}{|W_\mathcal{G} \cup W_\mathcal{S}|}$ $\phi_r = 1 - \frac{\sum_{\mathcal{G}\in W_\mathcal{G}} \#(Counter(\mathcal{G}) < \tau_r)}{|W_\mathcal{G} \cup W_\mathcal{S}|}$
- Sentence Level $r_i = \frac{|\mathcal{G}_i \cap \mathcal{S}_i|}{|\mathcal{S}_i|}, \quad \phi_s = 1 - \frac{1}{N}\sum_{i, r_i > \tau_c} r_i$
- Dataset Level $\phi_d = 1 - \frac{\sum_i |\mathcal{G}_i|}{\sum_i |\mathcal{S}_i|}$

➢ Task-aware instruction network

- **Encoder.** Given the recognized glosses, we fuse instruction features encoded by the the pre-trained model (PTM).

$$\hat{h}_t = (1-\alpha)Attn_E(h_t, H_E, H_E) + \alpha h_i$$

- **Decoder.** The hidden states are passed to a masked self-attention and then generate gloss into spoken sentence.

$$\tilde{s}_t = Attn_D(s_t, s_{1:t}, s_{1:t})$$

- **Learning-based feature fusion.**

$$\alpha_{t+1} = \Gamma(\alpha_t, g_t)$$

Experimental Results

➢ TIN-SLT achieves the highest scores on most evaluation metrics with a significant margin.

Model	Dev Set BLEU-1	Dev Set BLEU-2	Dev Set BLEU-3	Dev Set BLEU-4	Dev Set ROUGH-L	Dev Set METEOR	Test Set BLEU-1	Test Set BLEU-2	Test Set BLEU-3	Test Set BLEU-4	Test Set ROUGH-L	Test Set METEOR
PHOENIX-2014-T Dataset Evaluation												
Raw Data (Yin and Read 2020)	13.01	6.23	3.03	1.71	24.23	13.69	11.88	5.05	2.41	1.36	22.81	12.12
Seq2seq (Camgoz et al. 2018)	44.40	31.93	24.61	20.16	46.02	-	44.13	31.47	23.89	19.26	45.45	-
Transformer (Camgoz et al. 2020)	50.69	38.16	30.53	25.35	-	-	48.90	36.88	29.45	24.54	-	-
Transformer (Yin and Read 2020)	49.05	36.20	28.53	23.52	47.36	46.09	47.69	35.52	28.17	23.32	46.58	44.85
Transformer Ens. (Yin and Read 2020)	48.85	36.62	29.23	24.38	49.01	46.96	48.40	36.90	29.70	24.90	48.51	46.24
DataAug (Moryossef et al. 2021b)	-	-	-	-	-	-	-	-	-	23.35	-	-
TIN-SLT(Ours)	**52.35**	**39.03**	**30.83**	**25.38**	48.82	**48.40**	**52.77**	**40.08**	**32.09**	**26.55**	**49.43**	**49.36**
ASLG-PC12 Dataset Evaluation												
Raw data (Yin and Read 2020)	54.60	39.67	28.92	21.16	76.11	61.25	54.19	39.26	28.44	20.63	75.59	61.65
Preprocessed data (Yin and Read 2020)	69.25	56.83	46.94	38.74	83.80	78.75	68.82	56.36	46.53	38.37	83.28	79.06
Seq2seq (Arvanitis et al. 2019)	-	-	-	-	-	-	86.70	79.50	73.20	65.90	-	-
Transformer (Yin and Read 2020)	**92.98**	89.09	83.55	**85.63**	82.41	95.93	92.98	89.09	85.63	82.41	95.87	96.46
Transformer Ens.(Yin and Read 2020)	92.67	88.72	85.22	81.93	**96.18**	**95.95**	92.88	89.22	85.95	82.87	**96.22**	**96.60**
TIN-SLT (Ours)	92.75	**88.91**	**85.51**	82.33	95.17	95.21	**93.35**	**90.03**	**87.07**	**84.29**	95.39	95.92

Table 1: Comparing the translation performance of TIN-SLT against state-of-the-art techniques on PHOENIX-2014-T and ASLG-PC12 datasets. Clearly, our TIN-SLT achieves the best performance on most metrics.

➢ Performance comparison of different feature fusion strategies, and other hyper-parameters.

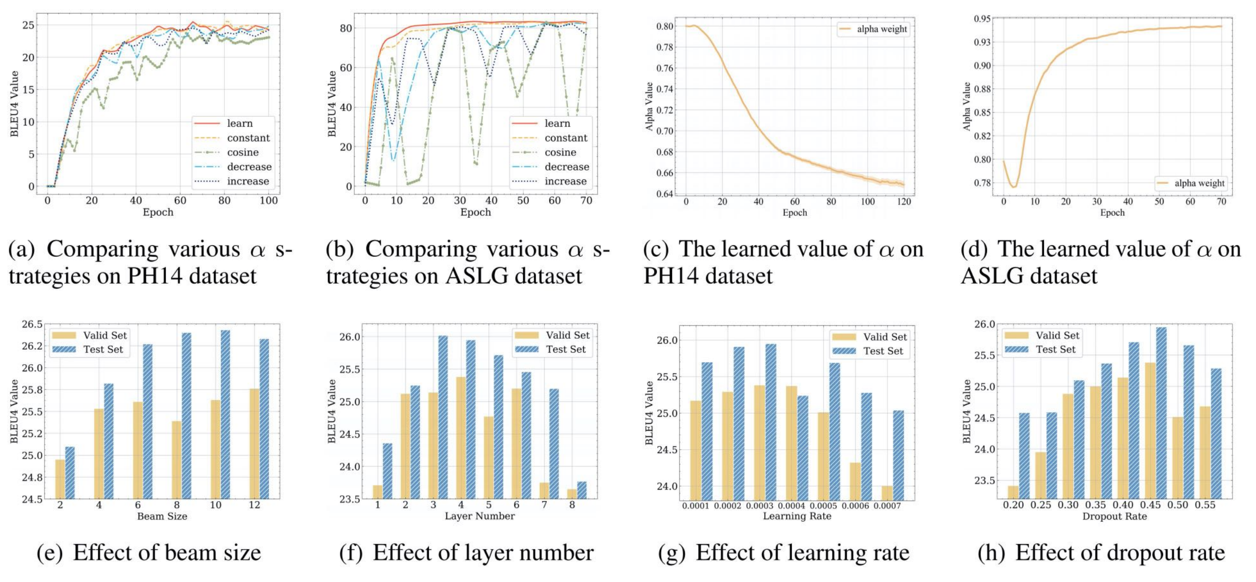

Figure 5: Various analysis results. (a) & (b) present the results by using different feature fusion strategies on two datasets, respectively. (c) & (d) show our learned value of α during the training process on the two datasets, respectively. (e)-(h) explore how beam size, layer number, learning rate, and dropout rate affect the model performance.

➢ Analysis on major network components.

Model	Test Set BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGH-L	METEOR
PHOENIX-2014-T Dataset Evaluation						
Baseline	47.69	35.52	28.17	23.32	46.58	44.85
w/ DataAug	50.77	37.85	29.88	24.57	47.39	46.95
w/ Encoder	51.05	37.94	29.91	24.63	47.59	47.13
w/ Decoder	50.99	38.47	30.48	25.08	48.78	48.20
Full pipeline	**52.77**	**40.08**	**32.09**	**26.55**	**49.43**	**49.36**
ASLG-PC12 Dataset Evaluation						
Baseline	92.98	89.09	85.63	82.41	**95.87**	96.46
w/ DataAug	92.60	89.15	85.80	83.05	95.08	95.33
w/ Encoder	92.77	89.22	86.23	83.40	95.22	**96.87**
w/ Decoder	93.15	89.80	86.49	83.89	95.34	95.67
Full pipeline	**93.35**	**90.03**	**87.07**	**84.29**	95.39	95.92

Table 3: Ablation analysis of our major network components on the G2T task.

- Table 3 proves that data augmentation can improve performance and our full model achieves the best performance.

➢ Case Study

Type	Content	BLEU-4
GT Gloss	X-I WANT IRELAND TO REMAIN AT HEART DECISION MAKE IN EUROPE .	57.58
GT Text	i want ireland to remain at the heart of decision making in europe .	
Pred Text	i want ireland to remain at the heart of the decision made in europe .	

Table 5: Qualitative evaluation of translation performance in different BLEU-4 scores on ASLG dataset.

- Table 5 presents an intuitive case on ASLG. The translation quality is good, even the translated texts with low BLEU-4 still convey valid information.