S^2Net: Skeleton-aware SlowFast Network for Efficient Sign Language Recognition Supplementary Material

Yifan Yang[1], Yuecong Min[2,3], and Xilin Chen[2,3]

[1] Huazhong University of Science and Technology, Wuhan, China
[2] Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China
[3] University of Chinese Academy of Sciences, Beijing, China
yyf_355@hust.edu.cn, {minyuecong, xlchen}@ict.ac.cn

We first deliver a more detailed description and pseudo code of graph-structured projectors in Sec A, and then we introduce the concrete realization, analysis, and discussion of S^2Net in Sec B. Afterward, we conduct additional ablation studies for S^2Net in Sec C. Finally, we perform more qualitative results and analysis in Sec D.

A More Details for Graph-structured projectors

We further elaborate on the detailed process of extracting graph-structured visual representations, as shown in Fig A. For the index-based projector, we first standardize the estimated 2D coordinates. Specifically, 'Resize' adjusts the spatial dimensions of the coordinates to 224×224 to match the resolution of the input image, while 'Clamp' is responsible for adjusting any invalid coordinates to the nearest valid boundary value. Afterward, we index the values from the pre-extracted feature maps at the corresponding locations based on the standardized coordinates and finally map these values to the graph-structured space through a linear projection. For the query-based projector, we first expand the learnable embeddings along the temporal dimension to T. Then, the cross-attention mechanism is utilized to aggregate elements related to specific graph-structured representation. In this process, the query vector is generated from the feature maps $\boldsymbol{V'}$, while the Key and Value vectors are generated from the learnable embeddings $\boldsymbol{E}$. Ultimately, these elements are also projected into the graph-structured space.

B Detailed Implementations

Implementation details of the S^2Net. For the slow pathway, all input image frames maintain a spatial resolution of 256×256, with video frames selected at a temporal stride of α=2. As for data augmentation, we apply random cropping (224×224), horizontal flipping(50%), and random temporal rescaling(20%). Only center cropping (224×224) is adopted during inference. Graph-structured

```
# X(V)[t, 3, h, w]          - video data after data augmentation
# X(J)[t, k, 2]             - estimated skeleton data after data augmentation
# E[t, k, c]                - learnable embedding
# Wq[c, cin]                - learnable query matrix to obtain query vector
# Wk[c, cin]                - learnable key matrix to obtain key vector
# Wv[c, cin]                - learnable value matrix to obtain value vector
# Visual_Module             - shallow layers of pre-trained ResNet18
# V(0)[t, c, k]             - graph-structrued representation

# extract mid-level visual feature maps
V' = Visual_Module(X(V))  # [t, c, h'*w']  h' = h*η  w' = w*η
If Index-based:
    # standardize 2D coordinates
    Jcoor = Clamp(Resize(X(J)))
    # convert 2D coordinates into 1D Indices[t, c, k]
    Indices = Index(Jcoor)
    # aggregate graph-structured representation
    XIndex = torch.gather(V', Indices, dim=2)
    # project representation into graph-structured space
    V(0) = Linear_Projection(XIndex)
If Query-based:
    # cross attention mechanism
    Q, K, V = V' · Wq, E · Wk, E · Wv      # Q[t, h'*w', cin]  K[t, k, cin]  V[t, k, cin]
    Attn = Softmax(Q · KT/(cin)1/2)
    XQuery = Feed_Forward(Attn · V) + Attn · V
    # project representation into graph-structured space
    V(0) = Linear_Projection(XQuery)
```

Fig. A: Torch-like pseudocode for the core of an implementation of graph-structured projectors.

representation is derived from mid-level feature maps extracted after the first two layers of a ResNet18 model pre-trained on the relevant sign language dataset based on SMKD [1]. For the fast pathway, we use MMPose to obtain the whole body skeleton data as previous work does [2]. 2D joint coordinates with confidence scores are concatenated as inputs, and only random temporal rescaling (20%) is applied for augmentation. The entire model is trained for 40 epochs with a minimum batch size of 4 and optimized by the AdamW optimizer with a weight decay of 1×10^{-4}. The initial learning rate starts as 4×10^{-4} and is reduced by a factor of 10 at 20 and 35 epochs, respectively. The loss weights λ_{CR} and λ_{KL} are set to 2 and 10, respectively. The pathway achieving the best performance on the dev set is selected as the final prediction.

Implementation details of the baseline and graph-structured projectors. We first employ the complete Resnet18 model as a frame-level feature extractor for the slow pathway to evaluate the impact of adopting pre-trained weights and freezing the visual extractor on performance. We then ablate the index layer of ResNet18 when generating the index-based graph-structured vi-

Table A: Performance (WER %) of each pathway on ablation results of index layer on Phoenix14. '-' denotes using the entire ResNet18 as the visual extractor without indexing. The index layer number indicates the number of residual layers from ResNet18 utilized.

Pathway	Freeze Weight	Index Layer	ImageNet Pre-trained		Phoenix14 Pre-trained	
			Dev	Test	Dev	Test
Slow		-	**22.3**	22.6	21.8	22.0
	✓	-	22.9	23.0	21.6	22.0
	✓	1	23.8	23.6	23.2	23.3
	✓	2	27.4	26.7	**20.1**	20.3
	✓	3	38.5	37.7	34.9	34.0
	✓	4	43.7	47.9	40.5	43.7
Fast		-	22.1	22.7	22.3	22.7
	✓	-	22.8	22.8	22.3	22.7
	✓	1	**20.5**	20.6	21.0	20.9
	✓	2	21.2	21.2	**20.5**	20.9
	✓	3	21.4	21.2	20.9	21.0
	✓	4	21.0	21.2	21.3	21.6
Average		-	**20.7**	21.7	20.3	20.7
	✓	-	21.5	21.7	20.2	20.8
	✓	1	20.8	20.8	20.6	20.9
	✓	2	21.9	22.7	**19.3**	19.4
	✓	3	25.6	26.0	24.2	24.1
	✓	4	29.5	29.7	29.0	29.7

sual representation which corresponds to Table 3 in the main paper. For both ablations above, we only utilize a straightforward fusion strategy where the prediction results of both pathways are summed and averaged, and the detailed results of each pathway are presented in Table A. In addition, we adopt indexing the extracted features after the second residual block pre-trained on Phoenix14 as our competitive baseline, with the performance of each pathway on the three datasets shown in Table B.

Performance of each pathway on three datasets. For the final evaluation, we employ four different prediction pathways: Slow pathway($\boldsymbol{y}^{(V)}$), Fast pathway($\boldsymbol{y}^{(J)}$), Frame-wise Fusion(FF) pathway($\boldsymbol{y}^{(F)}$), and Average pathway($\boldsymbol{y}^{(A)}$), where the performance of the Average pathway is derived by averaging the prediction results of the other three pathways. This process can be formulated as:

$$\boldsymbol{y}^{(A)} = \frac{\boldsymbol{y}^{(V)} + \boldsymbol{y}^{(J)} + \boldsymbol{y}^{(F)}}{3} \tag{A}$$

where $\boldsymbol{y}$ denotes the prediction result and superscript indicates the type of pathway. We select the pathway with the best performance on the dev set as the final prediction result. The performance of each pathway is detailed in Table B which corresponds to Table 1 in the main paper.

Table B: Performance (WER %) of each pathway on Phoenix14/14-T and CSL-Daily for two types of projectors and baseline. FF denotes the Frame-wise Fusion pathway. The final results we select are highlighted.

Pathway	PHOENIX14		PHOENIX14-T		CSL-Daily	
	Dev	Test	Dev	Test	Dev	Test
Baseline						
Slow	20.1	20.3	18.8	19.8	27.7	27.2
Fast	20.5	20.9	20.0	20.3	28.9	28.5
Average	**19.3**	**19.4**	**18.6**	**19.5**	**27.7**	**27.2**
Index-based						
Slow	18.3	18.3	18.3	18.8	26.8	25.7
Fast	18.2	18.2	18.1	18.9	27.3	26.2
FF	18.2	18.1	18.0	18.9	**25.4**	**24.5**
Average	**17.6**	**17.5**	**17.7**	**18.4**	28.1	26.9
Query-based						
Slow	18.1	18.2	18.3	18.9	26.8	25.6
Fast	18.2	18.5	**17.2**	**18.9**	26.4	25.9
FF	18.5	18.3	18.1	19.0	**25.8**	**24.5**
Average	**17.4**	**17.5**	17.6	18.2	27.0	26.1

Specifically, on the Phoenix14 dataset, both projectors achieve optimal performance on the Average pathway, with results of 17.6%/17.5% and 17.4%/17.5%, respectively. However, on the CSL-Daily dataset, the Frame-wise Fusion(FF) pathway exhibits superior performance over the Average pathway(from 28.1% to 25.4%, from 27.0% to 25.8%). On the Phoenix14-T dataset, for the index-based projector, the Average pathway demonstrates the best performance(17.7%/18.4%). In contrast, the Fast pathway performs best(17.2%/18.9%) for the query-based projector.

Implementation details of the frame-wise fusion. First, we perform temporal interpolation on the frame-wise features for the slow pathway to match the frame rates of both pathways. Next, we concatenate and fuse the features using an extra linear layer.

C Additional Ablation Studies

Ablation on pre-trained dataset. By utilizing a pre-trained visual module to construct the graph-structured representation, we hypothesize that the proposed method can also be used to assess the transferability across different datasets. We conduct an ablation on the pre-trained dataset and present results in Table C, on sign language datasets consistently outperforms pre-training on ImageNet by more than 0.4% WER on the dev set. We can also observe that pre-training the visual module on Phoenix14-T yields the best performance on the dev set at

Table C: Ablation results (WER %) of the pre-trained dataset on Phoenix14.

Pre-trained Dataset	ImageNet	Phoenix14	Phoenix14T	CSL-Daily
WER	18.7/18.7	17.7/17.7	**17.4**/17.9	18.2/18.0

Table D: Ablation results (WER %) of S^{2}Net on Phoenix14-T and CSL-Daily, GCA and FF denote Group-wise Cross-attention and Frame-wise Fusion, respectively.

Dataset	Graph-structured Projector		GCA	FF	Dev	Test
	Index-based	Query-based				
PHOENIX14-T	✓				19.5	19.7
	✓		✓		17.8	18.4
	✓			✓	18.5	18.5
	✓		✓	✓	17.7	18.4
		✓	✓	✓	**17.2**	18.9
CSL-Daily	✓				28.8	27.8
	✓		✓		26.2	25.4
	✓			✓	26.1	25.1
	✓		✓	✓	**25.4**	24.5
		✓	✓	✓	25.8	24.5

17.4%. This improvement may be attributed to the similarity between Phoenix14 and Phoenix14-T, both of which are collected from the same source.

More ablation on S^{2}Net designs. We further evaluated the effect of the S^{2}Net designs on the Phoenix14-T and CSL-Daily datasets, and the results are presented in Table D. Similarly, the adoption of both group-wise cross-attention and frame-wise fusion improves the recognition performance and the best results are achieved by adopting all of them. Moreover, adopting the query-based projector also achieves competitive performance on both datasets. The effectiveness of each component of S^{2}Net has been further demonstrated.

D More Qualitative Analysis

Visualizations of several learnable queries. The query-based projector aims to capture the visual feature for each joint without an explicit pose estimation stage. Although the input is a sequence of queries, we assume each learnable query has the ability to identify the corresponding region through the cross-attention between it and the extracted feature maps. We provide visualizations of the cross-attention maps for several queries, as shown in Fig B, which can relatively confirm our assumption. Subsequently, we organize the captured features into a graph sequence, feed it into the GCNs, and interact with the RGB path as the index-based projector does.

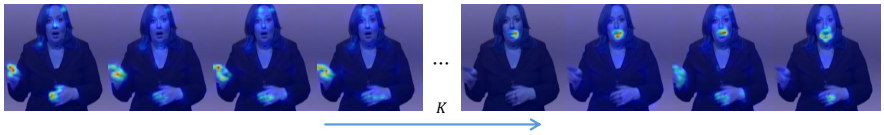

Fig. B: Visualizations of the cross-attention maps for several queries in the query-based projector, where K is the number of learnable queries.

Fig. C: Visualizations of the predictions from the different projectors. Wrong recognized glosses (except del) are marked in red.

Visualizations of the failure cases. We compare the predictions of the two types of projectors, as shown in Fig C. The query-based projector excels with words with complex and fast-moving movements (*e.g.*, regen) due to inaccuracies in pose estimation, but it struggles with capturing detailed static information (*e.g.*, nordwest). From a statistical perspective, the query-based projector reduces deletion errors (6.2% vs. 5.3%) while increasing substitution errors (8.8% vs. 9.5%) than the index-based projector. Besides, heatmaps provided in the main paper and Fig D show that the dispersion and incomplete attention regions of the query-based projector may also bring inevitable loss.

More visualizations on other datasets. To demonstrate the advantages of the query-based projector more clearly, we perform further visualization analyses on Phoenix14-T and CSL-Daily datasets, as shown in Fig D. In most cases, the query-based projector is more accurate in capturing the hand region and provides a wider field of view, especially on the CSL-Daily dataset. However, we also find that the query-based projector may have some shortcomings in capturing fine-grained details when focusing on global information. This lack of detail could be one of the reasons limiting its full potential.

References

1. Hao, A., Min, Y., Chen, X.: Self-mutual distillation learning for continuous sign language recognition. In: Int. Conf. Comput. Vis. pp. 11303–11312 (2021) 2
2. Jiao, P., Min, Y., Li, Y., Wang, X., Lei, L., Chen, X.: Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In: Int. Conf. Comput. Vis. pp. 20676–20686 (2023) 2

Fig. D: More visualizations of cross-attention maps and index maps generated from two types of projectors on Phoenix14-T and CSL-Daily.