Supplementary Material

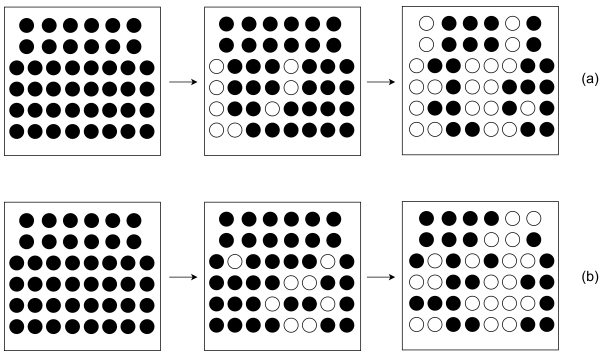

Figure 1. The variation of learnable parameters in a three-layer GCN within SPD. (a) represents the early stage of training, while (b) shows the model after convergence. Hollow points indicate the learned weak joints. Each row in the matrix corresponds to a partitioned VS, from top to bottom representing left arm, right arm, left fingers-1, left fingers-2, right fingers-1, and right fingers-2.

1. Additional Experimental Details

In VSformer, the GCN utilizes a shared adjacency matrix. Specifically, four learnable matrices are set for each video. After being input to the GCN, the channels are divided into four parts, which are individually convolved and then concatenated. Before calculating attention in TWA-l, we add relative positional biases to the features of each group. The down-sample layer employs a 2D convolution with a kernel size of (7,1) and a stride of 2, reducing the temporal dimension by half with each application.

2. More Experiments

2.1. Fine-tuning

Skeleton	MSASL			WLASL		
	100	200	1000	100	300	2000
type1	86.26	84.62	70.71	84.50	78.29	53.54
type1(fine-tune)	89.56	86.46	74.90	87.21	82.63	55.25

Table 1. Evaluating the Performance of our type-1 grouping (with and without fine-tuning) on the test set.

In the final stage, following [1], we fine-tune the model on the validation set as a reference to evaluate our model. Training is stopped when the training loss in the fine-tuning experiment decreases to the same level as the best model from the training phase. Partial results are shown in Tab. 1.

mouth	arm	hand	WLASL300	
			Top-1	Top-5
		✓	78.44	93.86
	✓	✓	79.04	94.01
✓	✓	✓	79.94	93.41

Table 2. Selection of key points for the face and arms.

2.2. Keypoints Selection

In previous experiments, we selected 44 upper body keypoints, including 40 hand points and 4 arm points. In this section, we discuss the scenario where only hand points are used on the WLASL300 dataset. Meanwhile, although we believe that facial expressions of the signer carry individual habits, and that using facial keypoints in insufficiently diverse datasets can reduce the model's generalization ability, many previous works have demonstrated significant improvements in sign language recognition by incorporating facial keypoints. Therefore, we added 12 facial points and present the final results in the table. As shown, the inclusion of facial keypoints led to a notable improvement, which can be attributed to the rich facial expressions in the WLASL dataset, where some signers articulate the gloss pronunciation while performing the sign language gestures.

2.3. More ablation study.

Methods	Ours	WLASL300		WLASL2000	
		Top-1	Top-5	Top-1	Top-5
GAN_1	SPD	74.25(↓4.8)	91.47	46.46(↓7.1)	77.24
GAN_2	SPA	76.20(↓2.8)	92.96	51.01(↓2.5)	81.90
$AGCN$	SPA	78.14(↓0.9)	93.41	51.63(↓1.9)	81.58
TC	TWA	77.25(↓1.8)	93.11	51.70(↓1.8)	82.56

Table 3. Comparing our models with existing models.

As shown in Tab. 3, GAN_1 and GAN_2 respectively refers to the use of graph attention networks as a replacement for SPD (dropping block) and SPA. Similarly, AGCN (adaptive GCN) replaces SPA, and TC (temporal convolution layer with a kernel size of 5) replaces TWA.

3. Visualization

As shown in the Fig. 1, we present the learning process of weak joints by the GCN in SPD. It is evident that there is a significant difference between the early and late stages of training. In (b), the learned weak joints exhibit better symmetry between the left and right hands, demonstrating that the model has effectively captured the correlations between joints.

References

[1] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *CVPR*, pages 3413–3423, 2021. 1