Supplementary Materials of "Pose Guided Fine-Grained Sign Language Video Generation"

1. More Experiment results

1.1. High-resolution results

The reason why we used 128*128 resolution is to reduce training costs. To verify the effectiveness of our method at higher resolutions, we conducted experiments on Phoenix 2014T at 256*256 (with a reduced training dataset, using only one pair of source and driving images per video, which is 1/5 of the original amount). As shown in Fig.1, and Tab.1, our method has great potential at high resolution, with significant improvements in clear hand region details compared to the baseline.

1.2. Results on large scale dataset

We tested our method on the large-scale sign language dataset How2Sign (using the same setting as 1.1, with one-fifth of the original training size), and the results are shown in Fig.2 and Tab.1. We achieved satisfactory generation results. We believe that for SLVG tasks, a larger vocabulary (representing action diversity) may bring less generalizability improvement compared to a larger number of signers (representing signer diversity).

1.3. Back-translation result

We did not use back-translation because the checkpoint of the SLT method used in the original paper is not open-sourced, making it difficult for future work to compare fairly. Instead, we requested the checkpoint of the CorrNet+ method [2], which the author promised would be released soon, to test our results at 256x256, as shown in Tab.1. For intelligibility, we will collaborate with sign language experts to conduct intelligibility assessments in our future research.

1.4. Cross-dataset results

In fact, our method has good generalizability, as shown in Fig.2, where we used a model trained on WLASL2000 and tested it on How2Sign in a zero-shot setting. WLASL2000 consists of sign language data from over 100 signers, and we believe that increasing the number of signers in the training data will lead to better generalizability according to our experiments.

2. More discussion

2.1. Heatmap selection

We were inspired by the Two-stream SLR method [1] in using heatmaps for pose guidance, which has the advantage of having the same dimension as image features, making it easy to interact with image features. All the experimental results show that using dense heatmaps provides more spatial information than using OpenPose's sparse graph or keypoint coordinates, which helps guide the generation of detailed parts.

2.2. Cross-person optical flow quality

We employ TPSMM for Motion Generation. Specifically, we estimate 100 keypoints using ResNet-18 (corresponding to 20 sets of TPS transformations, with each transformation having 5 keypoints). These keypoints are then used to generate 20 TPS transformations, thereby generating optical flow information.

During cross-person generation, the semantic information corresponding to keypoints from different persons remains consistent (as illustrated in Fig.1). Consequently, the quality of the generated optical flow is sufficiently

[1] Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. In: Advances in Neural Information Processing Systems. vol. 35, pp. 17043–17056 (2022) 1

[2] Hu, L., Feng, W., Gao, L., Liu, Z., Wan, L.: Cornet+: Sign language recognition and translation via spatial-temporal correlation (2024) 1

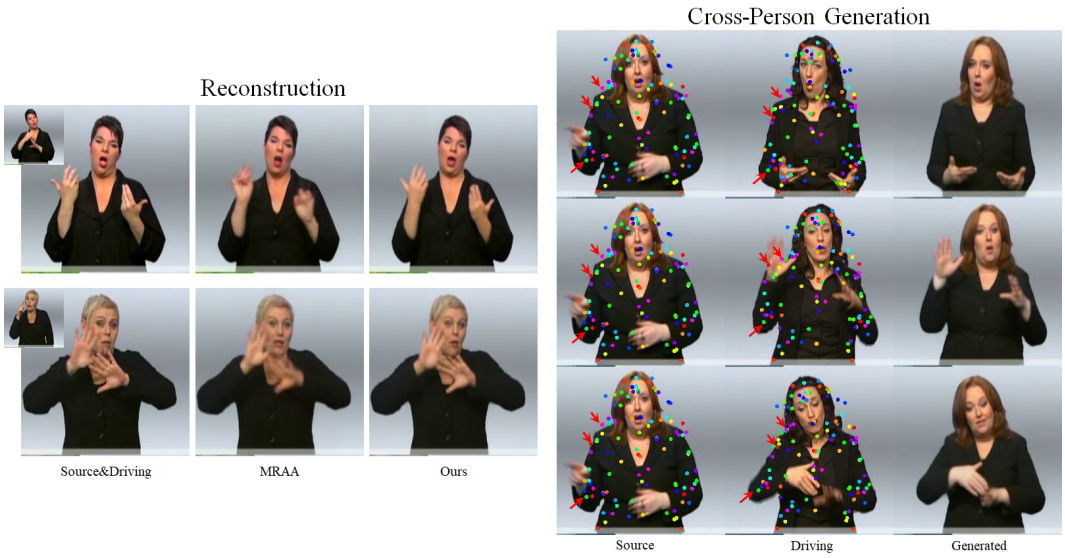

Figure 1. Qualitative results on Phoenix 2014T at 256*256. You can enlarge the image to obtain clearer details.

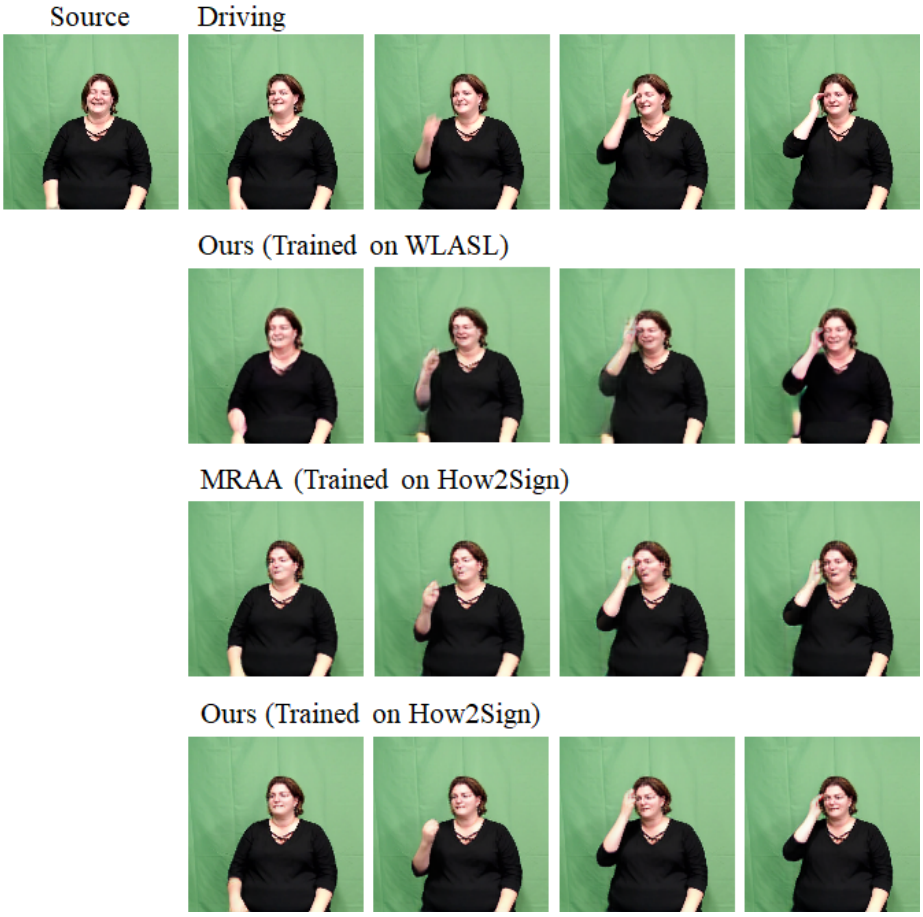

Figure 2. Reconstruction results on How2Sign. You can enlarge the image to obtain clearer details.

Table 1. Quantitative results.

Dataset	Model	L1	SSIM	LPIPS	FVD	TCD	B1	B2	B3	B4	ROUGE
Phoenix14T test(256)	SIGNGAN	-	0.759	-	-	-	-	-	-	18.78	40.60
	SignDiff	-	0.849	-	-	-	-	-	-	22.15	**46.82**
	MRAA	0.02116	0.8868	0.05243	277.499	0.109	31.64	21.89	16.10	12.58	30.57
	Ours	**0.01898**	**0.8944**	**0.04562**	**161.925**	**0.107**	**45.38**	**34.57**	**27.46**	**22.64**	44.18
How2Sign test (128)	MRAA	0.02446	0.8584	0.04485	294.057	0.155	-	-	-	-	-
	Ours	**0.01954**	**0.8944**	**0.03230**	**205.296**	**0.140**	-	-	-	-	-