Signing Outside the Studio: Benchmarking Background Robustness for Continuous Sign Language Recognition

Poster Number : 322

BMVC 2022

Youngjoon Jang[1] Youngtaek Oh[1] Jae Won Cho[1] Dong–Jin Kim[2] Joon Son Chung[1] In So Kweon[1]

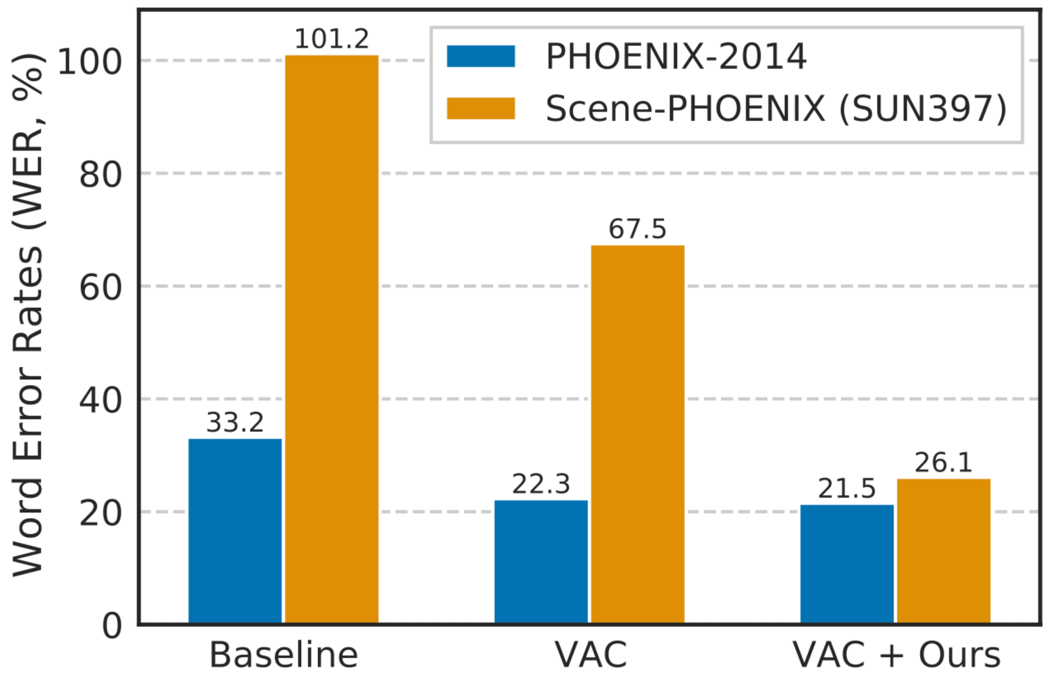
[1]KAIST, Daejeon, Republic of Korea [2]Hanyang Uni., Seoul, Republic of Korea

Sign Language Recognition

Motivation:

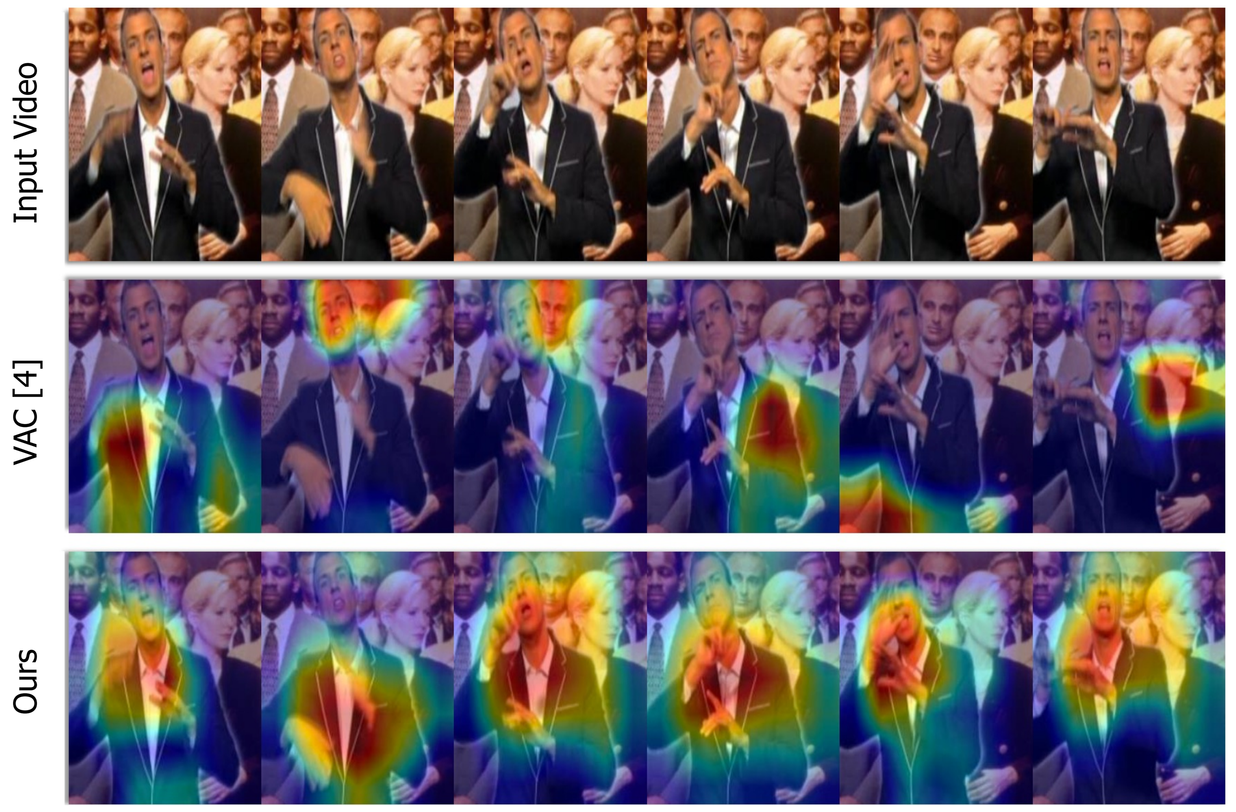
- Most existing Continuous Sign Language Recognition (CSLR) benchmarks are filmed in studios with a monochromatic background.

- **Observation:** Even the recent state-of-the art models suffer significant performance degradation on random background videos



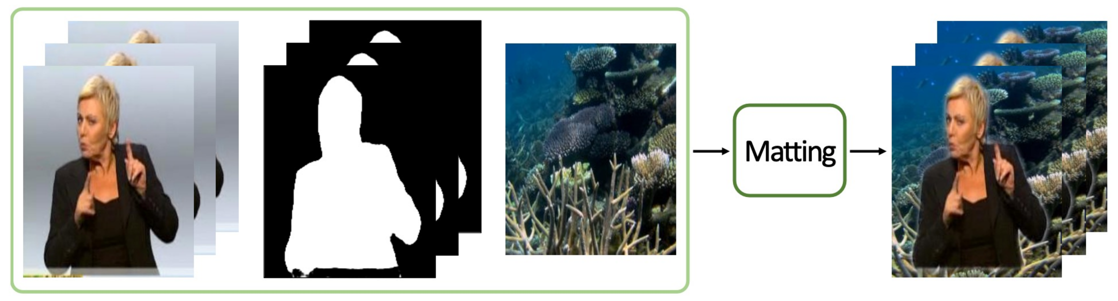
Our Contributions

- We propose an automatic benchmark dataset generation pipeline that can be applied to any CSLR dataset (Scene-PHOENIX).
- We propose a new training scheme for CSLR, including Background Randomization (BR) and Disentangling Auto-Encoder (DAE).
- We experimentally show that our approach effectively improves the robustness to background shifts while maintaining the performance.



Grad-CAM [5] activation maps

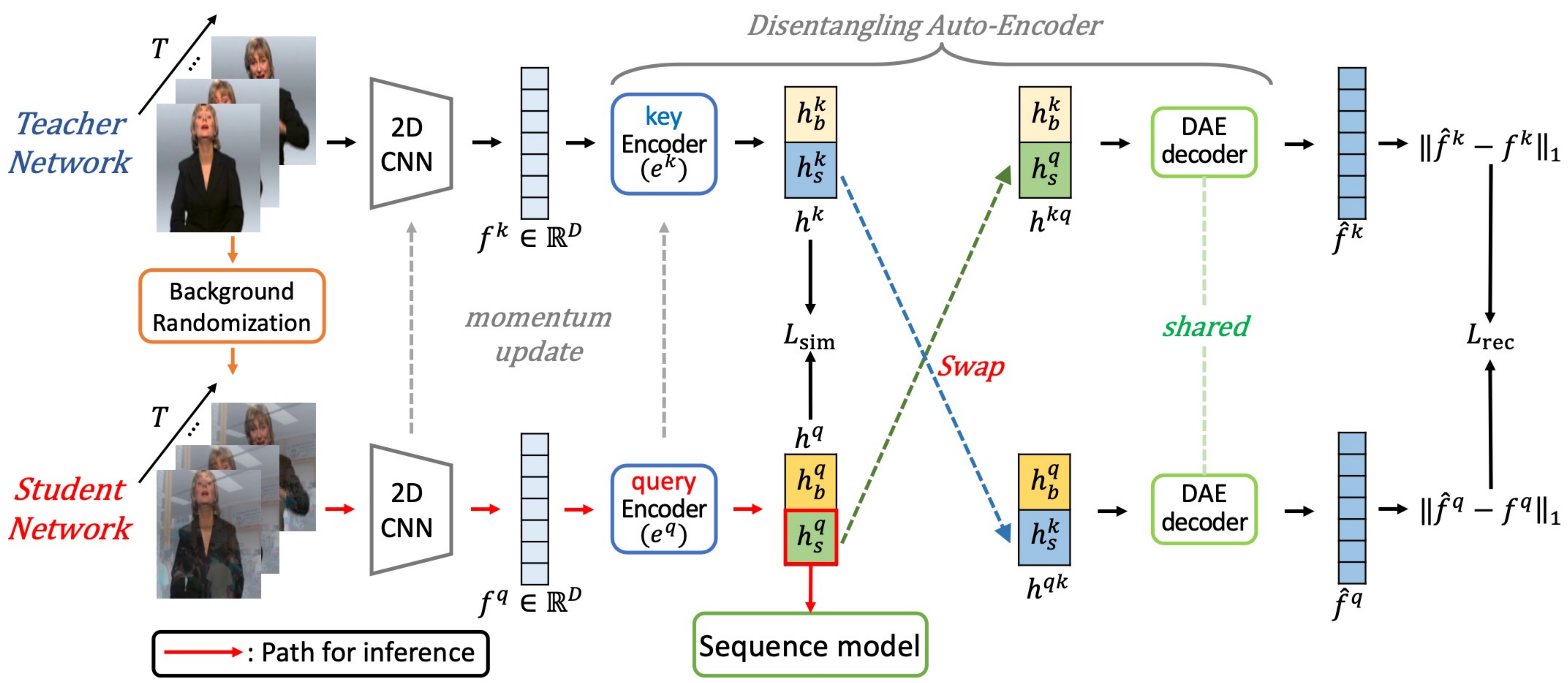
Background Agnostic Framework



Scene-PHOENIX dataset generation

Background randomization for training

Disentangling Auto-Encoder



$$L^{pos}_{\text{sim}}(x_1,x_2) = 1 - \cos(x_1,x_2), \quad L^{neg}_{\text{sim}}(x_1,x_2) = \max(0, \cos(x_1,x_2) - \Delta)$$

$$L_{\text{sim}} = L^{pos}_{\text{sim}}(h^q_s, h^k_s) + L^{neg}_{\text{sim}}(h^q_b, h^k_b)$$

$$L_{\text{rec}} = \|\hat{f}^q - f^q\|_1 + \|\hat{f}^k - f^k\|_1$$

$$L_{\text{total}} = \underbrace{L_{\text{CTC}} + L_{\text{VE}} + \alpha L_{\text{VA}}}_{L_{\text{VAC}}} + \underbrace{L_{\text{sim}} + L_{\text{rec}}}_{L_{\text{DAE}}}$$

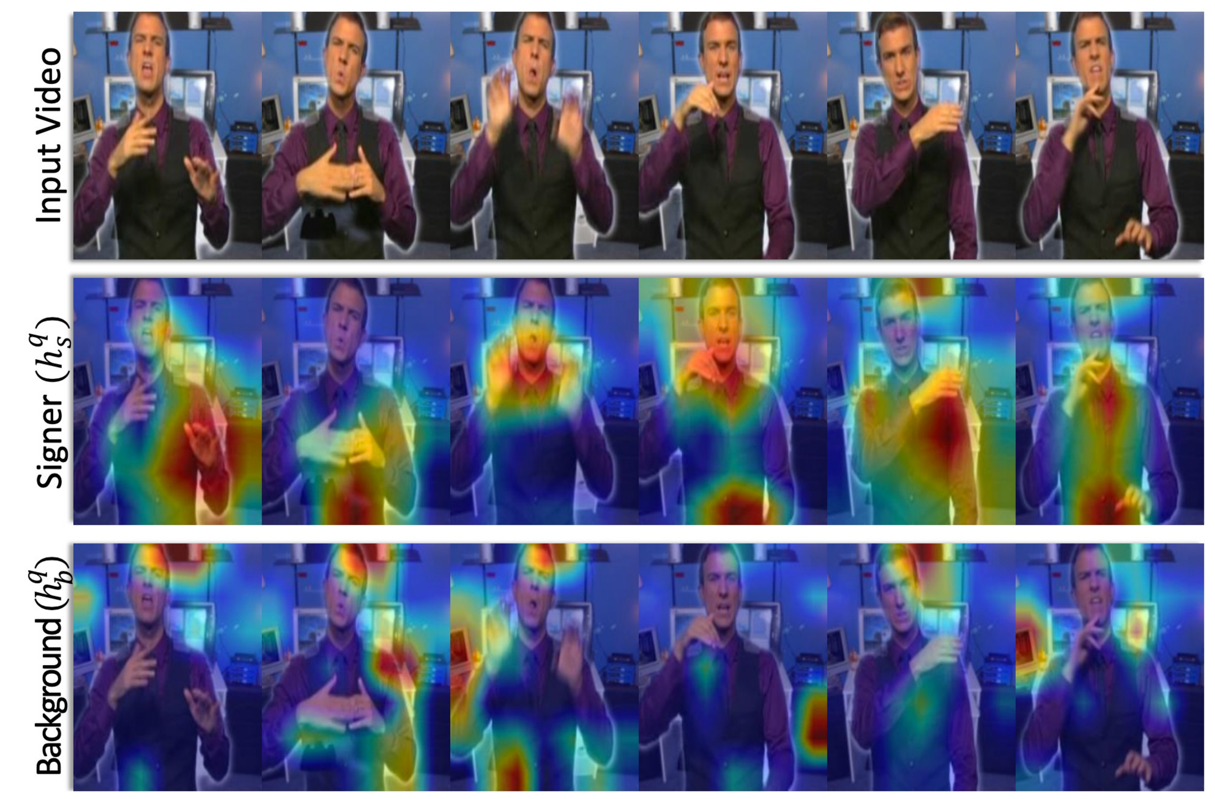
Quantitative Results

Method	K	PHOENIX-2014 WER Dev	PHOENIX-2014 WER Test	Scene-PHOENIX WER^{LSUN} Dev	Scene-PHOENIX WER^{LSUN} Test	Scene-PHOENIX WER^{SUN} Dev	Scene-PHOENIX WER^{SUN} Test
VAC-Oracle [41]	0.1M+	21.5	22.0	24.3	24.2	23.8	24.1
Baseline	-	31.2	33.2	101.1	101.0	100.9	101.2
w/ pretrain	-	25.4	26.1	71.0	76.6	69.9	72.7
w/ BR + DAE (Ours)	10	23.1	23.2	30.0	29.9	27.8	28.6
VAC	-	21.2	22.3	65.0	68.8	66.7	67.5
w/ BR	1	21.9	22.9	30.0	32.4	30.5	30.5
w/ BR	10	21.2	22.4	30.1	32.0	29.5	30.4
w/ BR	100	21.5	21.8	30.0	31.9	31.7	30.7
w/ BR	1000	22.4	22.9	27.7	29.2	28.5	28.6
w/ BR + DAE (Ours)	1	**20.6**	**21.5**	26.4	27.7	26.3	26.1
w/ BR + DAE (Ours)	10	20.9	21.5	26.7	27.4	26.4	26.1
w/ BR + DAE (Ours)	100	21.5	21.9	23.7	24.0	23.3	23.6
w/ BR + DAE (Ours)	1000	20.8	21.7	**22.9**	**23.4**	**22.5**	**23.1**

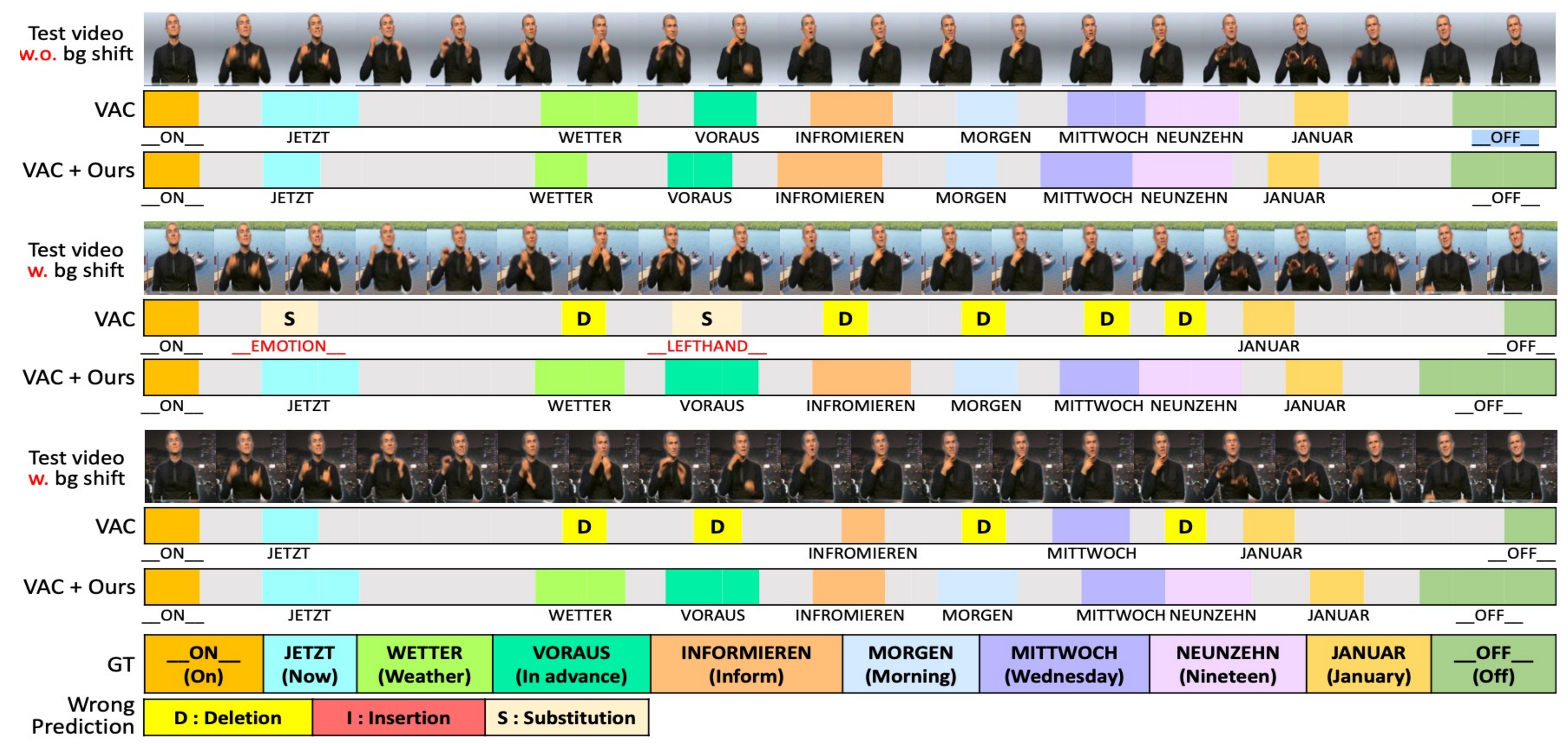
VAC-Oracle: VAC model trained on all LSUN [6] background matted videos.

DAE not only improves the performance on Scene-PHOENIX, but also achieves better performances on PHOENIX-2014

Qualitative Results



Grad-CAM comparison of the signer features and background features



Comprehensive comparison of gloss predictions between VAC and Ours

Reference

[1] Koller et al. "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers.", CVIU, 2015.
[2] Huang et al. "Video-based Sign Language Recognition without Temporal Segmentation", AAAI, 2018.
[3] Duarte et al. "How2sign: a large-scale multimodal dataset for continuous American sign language", CVPR, 2021.
[4] Min et al. "Visual alignment constraint for continuous sign language recognition", ICCV, 2021.
[5] Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization", ICCV, 2017.
[6] Yu et al. "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop" arXiv, 2015.