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

Youngjoon Jang[1]
wgs01088@kaist.ac.kr
Youngtaek Oh[1]
youngtaek.oh@kaist.ac.kr
Jae Won Cho[1]
chojw@kaist.ac.kr
Dong-Jin Kim[2]
djdkim@hanyang.ac.kr
Joon Son Chung[1]
joonson@kaist.ac.kr
In So Kweon[1]
iskweon77@kaist.ac.kr

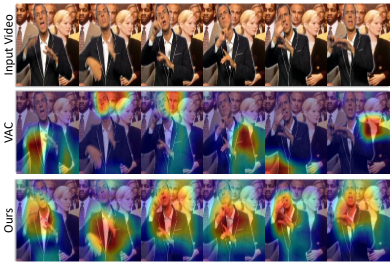
[1] Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea
[2] Hanyang University
Seoul, South Korea

Abstract

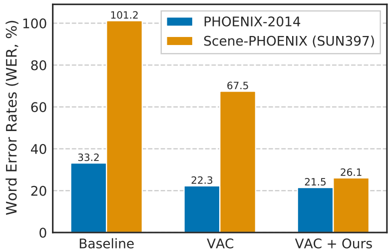
The goal of this work is background-robust continuous sign language recognition. Most existing Continuous Sign Language Recognition (CSLR) benchmarks have fixed backgrounds and are filmed in studios with a static monochromatic background. However, signing is not limited only to studios in the real world. In order to analyze the robustness of CSLR models under background shifts, we first evaluate existing state-of-the-art CSLR models on diverse backgrounds. To synthesize the sign videos with a variety of backgrounds, we propose a pipeline to automatically generate a benchmark dataset utilizing existing CSLR benchmarks. Our newly constructed benchmark dataset consists of diverse scenes to simulate a real-world environment. We observe even the most recent CSLR method cannot recognize glosses well on our new dataset with changed backgrounds. In this regard, we also propose a simple yet effective training scheme including (1) background randomization and (2) feature disentanglement for CSLR models. The experimental results on our dataset demonstrate that our method generalizes well to other unseen background data with minimal additional training images. Our dataset is available here.

1 Introduction

Most publicly available CSLR benchmarks are curated from either studio or TV broadcasts, where background images are fixed and monochromatic [19, 28, 35]. In a deployment scenario, these backgrounds are dissimilar to situations where real world communications occur,



(a) Grad-CAM [49] activation maps.



(b) WER scores from test benchmarks.

Figure 1: We evaluate CSLR models trained on PHOENIX-2014 [35] under our newly proposed Scene-PHOENIX benchmark. **(a)** VAC trained on monochromatic background sign videos fails to attent to the signer in the video and attends to other people in the background. **(b)** Both Baseline (ResNet18 [21] + LSTM) and VAC [41] severely degrade when tested on Scene-PHOENIX. In contrast, our framework can still capture signer's dominant expressions and favorably close the gap between test splits of the original PHOENIX-2014 and Scene-PHOENIX. We further detail experimental setup in Sec. 5.

potentially limiting the practicality of CSLR models. A naïve solution to this would be constructing a new dataset outside the studio, but the cost of extensive gloss annotations as well as collecting sign videos with skilled signers present significant challenges.

To tackle this issue, we first propose an automatic pipeline for CSLR benchmark dataset generation that re-uses existing CSLR datasets to synthesize a new dataset with various backgrounds for evaluating the robustness of background changes with minimal human intervention. During this process, we select natural scene images from scene datasets [56, 58] and carefully incorporate them on the test set of existing benchmark with a person mask. We make variants of development and test splits of PHOENIX-2014 [35] with our automated pipeline and name our benchmark dataset with diverse backgrounds *Scene-PHOENIX*.

Based on our Scene-PHOENIX dataset, we find that current CSLR approaches are not robust to background shifts. For example, we observe that VAC [41], a state-of-the-art CSLR method, severely degrades on our new evaluation criteria. Hence, we find the need for addressing background shift issue to improve practicality. To this end, we further propose a simple yet effective training scheme. First, we employ background randomization, where a sign video for training is combined with a scene image via mixup [60] to simulate background shifts. Then, we design a Disentangling Auto-Encoder (DAE), which aims to disentangle the signer feature and the background feature in the latent space. We emphasize that we use only a few background images during training, and the DAE can disentangle the signers from the background without additional annotations (*e.g.*, body keypoint, person mask) as shown in Fig. 1(a). From the experimental results based on Scene-PHOENIX in Fig. 1(b), we show that our method greatly reduces the gap between performance on test set with diverse backgrounds and with monochromatic backgrounds while boosting the performance on the original test set as well.

Our contributions are as follows: (1) To the best of our knowledge, we are the first to study the background shift issue in CSLR. We propose an automatic benchmark dataset generation pipeline that can be applied to any CSLR dataset and generate our dataset, Scene-PHOENIX. (2) We propose a new training scheme for CSLR, including background randomization and Disentangling Auto-Encoder (DAE) to improve robustness to background shifts.

Our approach can be readily integrated to any CSLR models using only a small number of extra images, without any annotations. (3) We experimentally show even the recent state-of-the-art CSLR models suffer significant performance degradation on Scene-PHOENIX. We also show that our approach effectively improves the robustness to background shifts while maintaining the performance from the original test data.

2 Related Work

Background Bias in Sign Videos. The Continuous Sign Language Recognition (CSLR) task [5, 28, 36, 47] aims to predict a sequence of glosses from a sign video. To efficiently construct benchmarks for the CSLR task on a large-scale, sign videos are often crawled from TV programs [1, 2, 35]. Sign videos are also captured in lab environments, to obtain multi-view videos [13, 34] or body pose and depth [13, 28, 54]. Based on such benchmarks, numerous CSLR studies have been proposed [5, 7, 20, 28, 37, 41, 44, 57, 61]. In isolated SLR tasks, Carneiro et al. [6] leverage background replacement for data augmentation at train-time. However, robustness to *background shift* has not been explored in the CSLR field [11, 28, 35]. As most CSLR benchmarks have monochromatic backgrounds [11, 28, 35], we observe that models are biased toward the background of the training data and cannot be generalized in videos with diverse background distributions (*e.g.*, indoor or outdoor scenes common in daily life). In this regard, we release a new benchmark dataset, named Scene-PHOENIX, to measure the robustness of CSLR models to background shifts without the need for collecting the new sign videos.

Robustness to Distribution Shift. Addressing the distribution shift is a crucial research problem since deep learning models are fragile to testing distribution different from the training [51]. In this aspect, various benchmarks have been proposed to measure the robustness under distribution shifts [9, 14, 23, 25, 26, 29, 30, 45, 48, 50], and this problem has been extensively studied in broad research fields [3, 4, 10, 15, 16, 24, 38, 39, 40, 43, 52, 55, 62]. Among them, benchmarking robustness [23] and resolving scene bias [10, 42] or distribution shift [43, 59] are the most related to our problem setup. Different from the aforementioned works, we first explore the background shift issue in the CSLR task with a newly synthesized benchmark. We then propose a simple and versatile training scheme to train a background-agnostic CSLR model while preserving performance in the original test data without background change.

3 Benchmarking Background Robustness

Background Bias. Continuous Sign Language Recognition (CSLR) videos are collected from weather forecasts [35] or studio recordings [13, 28], causing backgrounds to be fixed and monochromatic as in Fig. 2. As a result, CSLR models are trained and evaluated on a dataset with the same background distribution. However, the model is biased toward the background of the training data and fails to generalize in sign videos with *unseen* backgrounds. To support the hypothesis, we train CSLR models on PHOENIX-2014 [35], and compare the Word Error Rate (WER)[1] scores [35] between the original test split and our new test split, *Scene-PHOENIX*. As shown in Fig. 1(b), the performance of the baseline CSLR model *completely* degrades on Scene-PHOENIX (WER: 33.2% $\rightarrow$ 101.2%). Even

[1] WER = (#substitutions + #deletions + #insertions) / (#words in reference)

Figure 2: Comparison of videos in Continuous Sign Language Recognition (CSLR) dataset [11, 13, 35]. All videos display monochromatic backgrounds.

VAC [41] degrades in performance (WER: 22.3% → 67.5%). Furthermore, feature activations in Fig. 1(a) show that VAC is unable to consistently attend to the signer. On the other hand, our method significantly improves WER on background shifts, further closing the gap from the original test split.

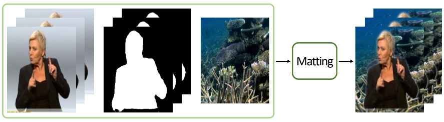
Robust Dataset Construction. In order to tackle the severe performance degradation of CSLR models during background shifts, we propose a new CSLR benchmark dataset for background shift evaluation. As collecting and annotating new sign language videos are expensive and time-consuming, we propose an effective alternative, an automatic CSLR benchmark dataset generation algorithm that utilizes existing datasets as shown in Fig. 3(a).

For a given sign video in PHOENIX-2014, a set of person mask is obtained for every frame by employing an open-source pretrained semantic segmentation network[2]. Then, we apply background matting with the masks and replace the background with external natural scene images across all frames. We utilize LSUN [58] and SUN397 [56], which contain a wide range of indoor and outdoor scenes of 10 and 397 classes respectively, as our background images. Note that we uniformly distribute each scene class across the sign samples – for the test split of PHOENIX-2014 with 629 samples, 62-63 videos are assigned for each class of LSUN, and 1-2 videos are assigned for each class of SUN397. Each synthesized set is called a *Split* and we generate three *Splits* for each dataset (*i.e.*, LSUN and SUN397) for reliable evaluation. Note that this method can be applied to other CSLR datasets to create additional benchmarks.

4 Background Agnostic Framework

We design a new framework that enables models to learn CSLR in a background agnostic manner as shown in Fig. 4. Our framework comprises of (1) *Background Randomization* (BR), which simply generates a sign video with new background via mixup [60] to simulate background shift, and (2) *Disentangling Auto-Encoder* (DAE) that aims to disentangle the signer from videos with background in latent space obtained by the query encoder e^q and the key encoder e^k. The Teacher Network from Fig. 4 serves as the target network, taking the original video and then provides positive and negative pairs from spatially encoded features, inspired by [18, 22]. Finally, the sequence model is provided with the disentangled signer

[2] https://github.com/thuyngch/Human-Segmentation-PyTorch



(a) Generation of Scene-PHOENIX.

(b) Background randomization *for training*.

Figure 3: Data generation. **(a)** For Scene-PHOENIX, background matting is performed with scene images using person masks. **(b)** For training set, we apply mixup [60] between a sign video and a scene image *without person masks* to reduce additional labeling cost in training.

component of latent feature h_s^q. We adapt the meta-architecture (CNN+LSTM) used in [7, 20, 41, 61] as our *Baseline* and train this model via CTC loss [17].

4.1 Background Randomization

We propose to leverage additional natural scene images to create videos with randomized backgrounds to bridge the gap from real world shifts. However, adding scenes used at test time during training can be seen as a *trivial* way to enhance robustness. Thus, to better test robustness and reduce potential costs, we limit the number of scene images available during training. In detail, we denote a variable K, where K is the number of images that we sample per class within the LSUN dataset. For example, if $K = 1$, we select 1 image per class, and LSUN has 10 classes, resulting in a total of 10 images for BR. Then, as shown in Fig. 3(b), we obtain a convex sum [60] of the target video with random background images. While Carneiro et al. [6] use person masks for changing backgrounds in training ISLR models, we emphasize BR is done without masks to reduce additional labor costs.

4.2 Disentangling Auto-Encoder

While Background Randomization (BR) improves a CSLR model's background shift robustness, we additionally propose a light-weight Disentangling Auto-Encoder (DAE) that further improves this robustness. In designing DAE, we hypothesize that the input video can be separated into a signer feature and a background feature in the embedding space.

As shown in Fig. 4, our framework consists of a *Teacher Network* and a *Student Network* [8, 22, 27, 31, 32] that both have the same network architecture. The teacher network takes the original sign videos, and the student network takes the background-randomized sign videos as their inputs. Each input videos pass through a spatial feature extractor (2D CNN) and then are flattened without average pooling to obtain D dimensional vectors f^k and f^q. Sequentially, key and query encoders embed f^k and f^q into h^k and h^q respectively. Here, we physically divide each latent feature (h^k, h^q) into two parts and we assume that the divided latent features consist of signer feature h_s and background feature h_b.

In order to embed more discriminative latent features, we give an additional training objective so that the signer features (*e.g.*, h_s^q and h_s^k) should pull each other and the background features (*e.g.*, h_b^q and h_b^k) push against each other. We employ the cosine similarity losses L_{sim}^{pos}

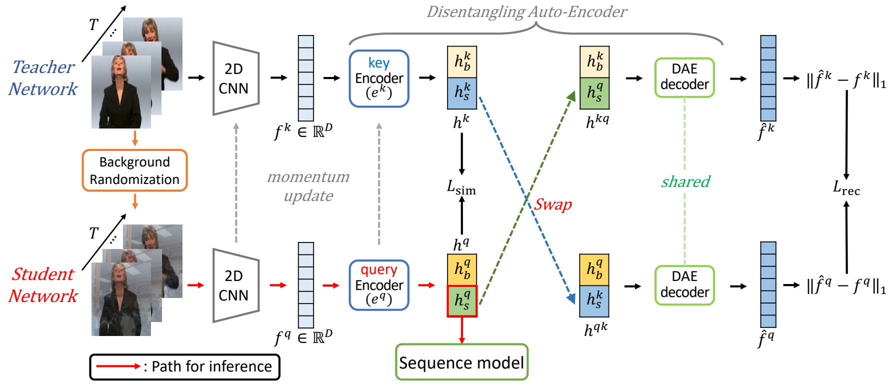


Figure 4: The overall architecture of the proposed model. The original video passes through Teacher Network, and the background-randomized video passes through Student Notwork. In the latent space, the signer features (h_s^k, h_s^q) are swapped with each other. Then, the swapped features (h^{kq}, h^{qk}) are input to the shared DAE decoder for reconstructing the original features f^k and f^q. Note the Red arrows show the path during inference.

and L_{sim}^{neg} for pulling and pushing respectively. The similarity loss is formulated as follows:

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

where Δ is a hyperparameter for the margin, which penalizes based on its value. The final similarity loss is given by:

$$L_{\text{sim}} = L_{\text{sim}}^{pos}(h_s^q, h_s^k) + L_{\text{sim}}^{neg}(h_b^q, h_b^k). \tag{2}$$

In the case that the latent features (h^q, h^k) are perfectly disentangled into signer feature (h_s) and background feature (h_b), there should be no difference between the signer features h_s^q and h_s^k. To try to enforce this, we first *swap* the signer features of the teacher network and the student network. h^{kq} and h^{qk} denote the swapped features. Here, we train our DAE decoder that re-projects the latent features (h^q, h^k) back into the original feature space after the 2D CNN by reconstructing f^q and f^k from the swapped features h^{kq} and h^{qk}. To enforce such reconstruction, L_{rec} is measured as L1 distances of respective reconstructed features:

$$L_{\text{rec}} = \|\hat{f}^q - f^q\|_1 + \|\hat{f}^k - f^k\|_1, \tag{3}$$

where $\hat{f}^q$ and $\hat{f}^k$ are re-projected features from the DAE decoder.

Finally, only h_s^q is passed through the sequence model to predict gloss sequences by propagating CTC loss [17], so the network can focus more on the signer in a background agnostic manner. During inference, the teacher network and DAE decoder are discarded, causing the inference overhead to be negligible. Note that the teacher network is updated by momentum update [22].

4.3 Objective Function.

The final objective of Ours when integrated with VAC [41] network is as follows:

$$L_{\text{total}} = \underbrace{L_{\text{CTC}} + L_{\text{VE}} + \alpha L_{\text{VA}}}_{L_{\text{VAC}}} + \underbrace{L_{\text{sim}} + L_{\text{rec}}}_{L_{\text{DAE}}}, \tag{4}$$

where the first three terms correspond to VAC, and the final two terms are used for train the DAE. In our framework, we empirically find that lower value of α accelerates the convergence of the model as a whole and set α to 3.

5 Experiments

Dataset. We utilize the publicly available PHOENIX-2014 [35] dataset to validate our framework. In order to evaluate the robustness of models to background changes, we construct Scene-PHOENIX by synthesizing the background and perform various experiments on both PHOENIX-2014 and Scene-PHOENIX. To generate Scene-PHOENIX, we use two large-scale scene datasets, LSUN [58] and SUN397 [56], which consist of about 1M images with 10 classes and about 0.1M images with 397 classes, respectively.
Evaluation Protocol. We measure the performance of CSLR models by WER score. Scene-PHOENIX consist of one Dev *Split*, and three Test *Splits* respectively on both LSUN and SUN397. We report the average of WERs from all Test *Splits*. $\mathrm{WER}^{\mathrm{LSUN}}$ and $\mathrm{WER}^{\mathrm{SUN}}$ denote WER in the test set in which Scene-PHOENIX video background is randomized with LSUN and SUN397 respectively.
Implementation details. The proposed Disentangling Auto-Encoder (DAE) has an encoder-decoder architecture, and both the encoder and the decoder consists of two fully-connected layers. The input video frames are first resized to 256×256, followed by random crop to 224×224 at the same location of all the frames and random horizontal flip with 50% probability. We then randomly insert duplicated frames up to 20% in total length, followed by random deletion of frames up to 20% of the whole length. We train the network using Adam optimizer [33] with batch size 2 and weight decay 10^{-4} for 100 epochs. The initial learning rate is 10^{-4} and the cosine scheduler is adopted to decay the rate. Our framework is implemented using PyTorch [46].

5.1 Main Results

VAC-Oracle. We first introduce a VAC [41] based Oracle as a frame of reference. The Oracle set uses all of the images within the LSUN dataset (over $1M$ images), and instead of using our simple BR method, we generate all images in the Oracle set using background matting as illustrated in Figure 3(a), identical to the Scene-PHOENIX set. Then, we report the VAC model trained on this Oracle set as VAC-Oracle in Table 1.
Baseline. Table 1 shows the comparison of our framework including BR and DAE with other methods. The Baseline model gives the WER scores larger than 100% in all splits of Scene-PHOENIX. While using pretrained feature extractor on ImageNet [12] (w/ pretrain) can be beneficial for the background shifts (WER^{SUN}_{Test}: 101.2% → 72.7%), there still exists large performance gap between PHOENIX-2014 and Scene-PHOENIX. In contrast, applying our BR and DAE to the pretrained Baseline dramatically improves the performance on Scene-PHOENIX (WER^{LSUN}_{Test}: 76.6% → 29.9%, WER^{SUN}_{Test}: 72.7% → 28.6%) and closes the gap in performance between PHOENIX-2014 and Scene-PHOENIX significantly.
Evaluation on VAC. We evaluate the VAC model under same condition as the Baseline experiments. When both BR and DAE are applied to VAC, Ours significantly improves the performance on Scene-PHOENIX, even with only 10 scene images in total (*e.g.*, $K = 1$).

Moreover, we find that the DAE not only improves the performance of our model on Scene-PHEONIX, but also achieves better performances in WER of Dev and Test splits in

		PHOENIX-2014 WER		Scene-PHOENIX $\mathrm{WER}^{\mathrm{LSUN}}$		$\mathrm{WER}^{\mathrm{SUN}}$	
Method	K	Dev	Test	Dev	Test	Dev	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**

Table 1: Experimental results on PHOENIX-2014 and Scene-PHOENIX. VAC-Oracle is a VAC model that is trained on all LSUN background matted images. While the performance of the baselines severely degrades under Scene-PHOENIX, the proposed Background Randomization (BR) shows significant performance improvements. Our final model (BR + DAE) shows the best performance among the baseline models. Note that our final model with $K = 1$ outperforms all VAC w/ BR models. Moreover, Our with $K = 1000$ surpasses the VAC-Oracle and VAC in both dataset without any off-the-shelf human segmentation masks.

PHOENIX-2014. This indicates that the disentanglement procedure in DAE with BR not only improves the model's robustness to background shifts but also improves the performance of sign videos in monochromatic backgrounds.

Variation on scene images K for training. By gradually increasing the value of K, we also observe that the WER reductions in Scene-PHOENIX are consistent, and DAE indirectly helps BR. We also highlight that Ours has superior performance to the oracle when $K = 1000$ in all metrics. While the oracle has access to more than $1M$ images during training, our BR + DAE approach is $1000\times$ more efficient in terms of the number of scene images required, even without using off-the-shelf human segmentation masks.

5.2 Additional Analysis

Ablation on additional training data. We systematically study the WER scores by ablating additional costs such as pretraining, using pose data, and BR for CSLR model training in Table 2. Note that we fix the feature extractor as ResNet-18, and $K = 10$ for BR. First, pose supervision with extra annotations [61] contributes to both WER and $\mathrm{WER}^{\mathrm{SUN}}$ significantly. Moreover, pretraining feature extractor on ImageNet (denoted as pretrain), while using pose supervision can further reduce the WER scores in Test split: (WER: 24.0% and $\mathrm{WER}^{\mathrm{SUN}}$: 46.6%), which shows the best performance throughout the experiments without using extra scene images. A similar observation can be made when we additionally apply BR. While the model trained with BR, pretrain, and pose can greatly improve the performance: (WER: 23.9% and $\mathrm{WER}^{\mathrm{SUN}}$: 29.4%), the model trained with BR + DAE without pose supervision

pretrain	Pose	BR	WER Dev	WER Test	WER^{SUN} Dev	WER^{SUN} Test
			31.2	33.2	100.9	101.2
	✓		28.4	27.9	73.4	73.5
✓	✓		23.6	24.0	51.4	46.6
		✓	29.9	31.0	48.0	48.3
	✓	✓	29.0	29.1	35.1	34.8
✓	✓	✓	23.8	23.9	29.5	29.4
ResNet-18 w/ BR + DAE			23.1	23.2	27.8	28.6

Table 2: Ablation on additional training data. Using DAE is more efficient in annotation cost compared to using pose, which requires extra annotation. We emphasize using additional 100 scene images for BR is much cheaper than annotating pose for training.

Backbone	WER Dev	WER Test	WER^{SUN} Dev	WER^{SUN} Test
GoogLeNet [53]	26.2	27.9	76.0	80.0
w/ BR + DAE	24.4	26.9	27.8	27.9
ResNet18 [21]	25.4	26.1	69.9	72.7
w/ BR + DAE	23.1	23.2	27.8	28.6

Table 3: Comparison of performances with different feature extractors: GoogLeNet [53] and ResNet18 [21]. Our framework consistently works well with different feature extractors.

outperforms it with 23.2% and 28.6% WER and WER^{SUN} in Test split, respectively. We conclude that DAE along with BR outperforms in all metrics with annotation-efficiently, compared to training CSLR models with pose supervision.

Different backbones. To show that our proposed framework of BR and DAE are generalizable, we show the results obtained by using different feature extractors in Table 3. While GoogLeNet [53] shows slightly worse performance in WER compared to ResNet-18, it shows greater improvements in WER^{SUN} when our BR + DAE is applied.

5.3 Qualitative Results

Latent Feature Disentanglement. To validate the DAE's ability to distinguish between the signer and background, we visualize the Grad-CAM [49] of h_s^q and h_b^q in Fig. 5. We see that the model is able to focus on the regions that make up the signer's features h_s^q (*e.g.*, hands and face) while the background feature h_b^q focuses on the background region (*i.e.*, the region outside the signer).

Gloss Prediction. We qualitatively show our method's robustness to background shifts in Fig. 6. we visualize the predicted glosses from three sign language videos, which have the ground truth with different backgrounds. Without background shift, the original VAC [41] correctly predicts gloss sequence while it fails when the background shifts. In contrast, Ours correctly predicts gloss sequences in all test videos regardless of background.

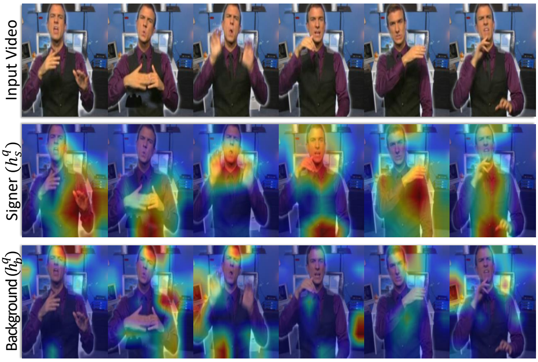


Figure 5: Grad-CAM [49] comparison of the signer feature h_s^q and background feature h_b^q in our framework. By virtue of our Disentangling Auto-Encoder, h_s^q and h_b^q in latent space consistently focus on the signer and background area respectively. Verifying that the two components are disentangled in latent space from the background randomized video.

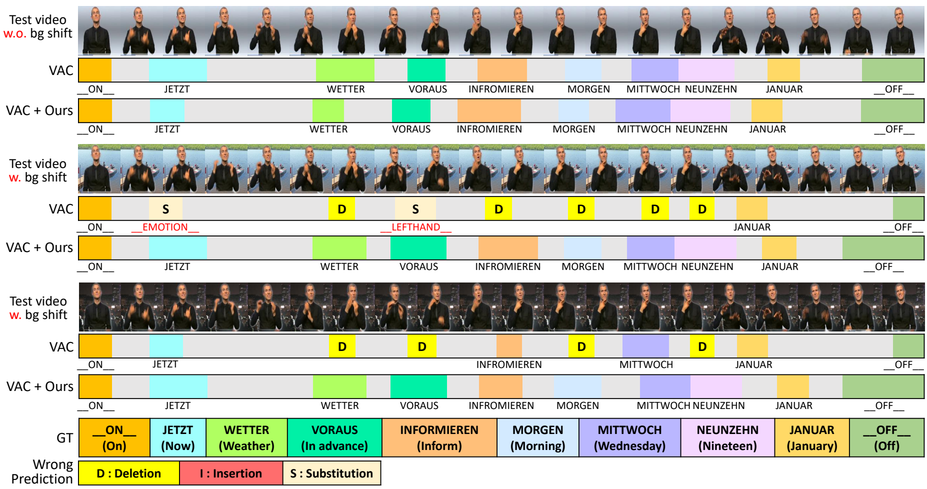


Figure 6: Comprehensive comparison of gloss predictions between VAC [41] and Ours. We visualize the frame-level gloss predictions from the models and show the difference when the background shifted. We observe that VAC fails to predict correct glosses with different backgrounds, while our method consistently recognizes glosses regardless of backgrounds.

6 Conclusion

In this paper, we introduce a new Scene-PHOENIX benchmark, composed of background-synthesized sign videos, and a pipeline for automatically generating CSLR benchmark with backgrounds to expand the possibilities for real world applications for CSLR. We also propose a simple but effective Background Randomization (BR) that shows dramatic performance gains without human segmentation masks and Disentangling Auto-Encoder (DAE) that disentangles the signer features from the background in latent space. Experiments on two datasets (PHOENIX-2014, Scene-PHOENIX) demonstrate strong robustness to background shifts while maintaining the existing model performance.

7 Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT, No. 2022-0-00989, Development of Artificial Intelligence Technology for Multi-speaker Dialog Modeling).

References

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*, pages 35–53. Springer, 2020.

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Hannah Bull, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. BOBSL: BBC-Oxford British Sign Language Dataset". *arXiv preprint arXiv:2111.03635*, 2021.

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[4] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3722–3731, 2017.

[5] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3075–3084, 2017.

[6] AL Cavalcante Carneiro, L Brito Silva, and DH Pinheiro Salvadeo. Efficient sign language recognition system and dataset creation method based on deep learning and image processing. In *Thirteenth International Conference on Digital Image Processing (ICDIP 2021)*, volume 11878, pages 11–19. SPIE, 2021.

[7] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *European Conference on Computer Vision (ECCV)*, pages 697–714, 2020.

[8] Jae Won Cho, Dong-Jin Kim, Jinsoo Choi, Yunjae Jung, and In So Kweon. Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[9] Jae Won Cho, Dong-jin Kim, Hyeonggon Ryu, and In So Kweon. Generative bias for visual question answering. *arXiv preprint arXiv:2208.00690*, 2022.

[10] Jinwoo Choi, Chen Gao, Joseph C. E. Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, 2019.

[11] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793, 2018.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009.

[13] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2735–2744, 2021.

[14] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning (ICML)*, pages 1802–1811. PMLR, 2019.

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[16] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

[17] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pages 369–376, 2006.

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.

[19] Soomin Ham, Kibaek Park, YeongJun Jang, Youngtaek Oh, Seokmin Yun, Sukwon Yoon, Chang Jo Kim, Han-Mu Park, and In So Kweon. Ksl-guide: A large-scale korean sign language dataset including interrogative sentences for guiding the deaf and hard-of-hearing. In *International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2021.

[20] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 11303–11312, October 2021.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[24] Dan Hendrycks*, Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple method to improve robustness and uncertainty under data shift. In *International Conference on Learning Representations (ICLR)*, 2020.

[25] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021.

[26] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, June 2021.

[27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[28] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018.

[29] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, Youngjin Yoon, and In So Kweon. Disjoint multi-task learning between heterogeneous human-centric tasks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.

[30] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[31] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision (ECCV)*, 2020.

[32] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Acp++: Action co-occurrence priors for human-object interaction detection. *IEEE Transactions on Image Processing (TIP)*, 30:9150–9163, 2021.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[34] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683, 2019.

[35] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.

[36] Oscar Koller, Hermann Ney, and Richard Bowden. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3793–3802, 2016.

[37] Oscar Koller, Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[38] Boyi Li, Felix Wu, Ser-Nam Lim, Serge Belongie, and Kilian Q. Weinberger. On feature normalization and data augmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12383–12392, June 2021.

[39] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5542–5550, 2017.

[40] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[41] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 11542–11551, October 2021.

[42] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *arXiv preprint arXiv:2108.00049*, 2021.

[43] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8690–8699, 2021.

[44] Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *European Conference on Computer Vision (ECCV)*, pages 172–186, 2020.

[45] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9786–9796, 2022.

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, page 8026–8037, 2019.

[47] Junfu Pu, Wengang Zhou, and Houqiang Li. Dilated convolutional network with iterative optimization for continuous sign language recognition. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 3, page 7, 2018.

[48] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, pages 5389–5400. PMLR, 2019.

[49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[50] Arda Senocak, Junsik Kim, Tae-Hyun Oh, Hyeonggon Ryu, Dingzeyu Li, and In So Kweon. Audio-visual fusion layers for event type aware video recognition. *arXiv preprint arXiv:2202.05961*, 2022.

[51] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[52] Inkyu Shin, Dong-Jin Kim, Jae Won Cho, Sanghyun Woo, Kwanyong Park, and In So Kweon. Labor: Labeling only if required for domain adaptive semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8588–8598, 2021.

[53] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[54] Ulrich Von Agris and Karl-Friedrich Kraiss. Towards a video corpus for signer-independent continuous sign language recognition. *Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, May*, 11:2, 2007.

[55] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

[56] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119(1):3–22, 2016.

[57] Zhaoyang Yang, Zhenmei Shi, Xiaoyong Shen, and Yu-Wing Tai. Sf-net: Structured feature network for continuous sign language recognition. *arXiv preprint arXiv:1908.01341*, 2019.

[58] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[59] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2100–2110, 2019.

[60] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[61] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 13009–13016, 2020.

[62] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations (ICLR)*, 2021.