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

Youngjoon Jang
wgs01088@kaist.ac.kr
Youngtaek Oh
youngtaek.oh@kaist.ac.kr
Jae Won Cho
chojw@kaist.ac.kr
Dong-Jin Kim
djdkim@hanyang.ac.kr
Joon Son Chung
joonson@kaist.ac.kr
In So Kweon
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,
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, 28, 54] 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 [13, 28, 54], 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% → 101.2%). Even

\(^{1}\text{WER} = \frac{\text{#substitutions} + \text{#deletions} + \text{#insertions}}{\text{#words in reference}}\)
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 network2. 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

2https://github.com/thuyngch/Human-Segmentation-PyTorch
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., $h^q_s$ and $h^k_s$) should pull each other and the background features (e.g., $h^q_b$ and $h^k_b$) push against each other. We employ the cosine similarity losses $L_{\text{cos}}^{\text{pos}}$.

**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.
Figure 4: The overall architecture of the proposed model. The original video passes through Teacher Network, and the background-randomized video passes through Student Network. In the latent space, the signer features \((h^s_k, h^s_b)\) 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 \(L_{\text{sim}}^{\text{neg}}\) for pulling and pushing respectively. The similarity loss is formulated as follows:

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

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

\[
L_{\text{sim}} = L_{\text{sim}}^{\text{pos}}(h^q_s, h^s_b) + L_{\text{sim}}^{\text{neg}}(h^q_b, h^s_k).
\]

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 \(h^q_s\) and \(h^k_s\). 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, \(L_{\text{rec}}\) is measured as L1 distances of respective reconstructed features:

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

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

Finally, only \(h^q_s\) 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 [28] network is as follows:

\[
L_{\text{total}} = \underbrace{L_{\text{CTC}} + L_{\text{VE}} + \alpha L_{\text{VA}}}_{L_{\text{VAC}}} + \underbrace{L_{\text{sim}} + L_{\text{rec}}}_{L_{\text{DAE}}},
\]
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 $\alpha$ accelerates the convergence of the model as a whole and set $\alpha$ 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. WER$_{\text{LSUN}}$ and WER$_{\text{SUN}}$ 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 $256 \times 256$, followed by random crop to $224 \times 224$ 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 ($\text{WER}_{\text{Test}}^{\text{SUN}}: 101.2\% \rightarrow 72.7\%$), 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 ($\text{WER}_{\text{Test}}^{\text{LSUN}}: 76.6\% \rightarrow 29.9\%, \text{WER}_{\text{Test}}^{\text{SUN}}: 72.7\% \rightarrow 28.6\%$) 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-PHOENIX, but also achieves better performances in WER of Dev and Test splits in
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 Our 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 WER\(^{\text{SUN}}\) 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 WER\(^{\text{SUN}}\): 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 WER\(^{\text{SUN}}\): 29.4%), the model trained with BR + DAE without pose supervision...
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.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>WER</th>
<th>WER^SUN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>pretrain</td>
<td>Pose</td>
<td>BR</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>31.2</td>
<td>33.2</td>
</tr>
<tr>
<td></td>
<td>28.4</td>
<td>27.9</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>23.6</td>
<td>24.0</td>
</tr>
<tr>
<td>ResNet-18 w/ BR + DAE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23.1</td>
<td>23.2</td>
</tr>
</tbody>
</table>

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

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 $h^q_s$ and $h^q_b$ in Fig. 5. We see that the model is able to focus on the regions that make up the signer’s features $h^q_s$ (e.g., hands and face) while the background feature $h^q_b$ 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 $h^s_b$ and background feature $h^b_b$ in our framework. By virtue of our Disentangling Auto-Encoder, $h^s_b$ and $h^b_b$ 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


