SELF-SUFFICIENT FRAMEWORK FOR CONTINUOUS SIGN LANGUAGE RECOGNITION

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

[1]Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
[2]Hanyang University, Seoul, Republic of Korea

Project page with demo: https://mm.kaist.ac.kr/projects/ssslr

ABSTRACT

The goal of this work is to develop *self-sufficient* framework for Continuous Sign Language Recognition (CSLR) that addresses key issues of sign language recognition. These include the need for complex multi-scale features such as hands, face, and mouth for understanding, and absence of frame-level annotations. To this end, we propose (1) Divide and Focus Convolution (DFConv) which extracts both manual and non-manual features without the need for additional networks or annotations, and (2) Dense Pseudo-Label Refinement (DPLR) which propagates non-spiky frame-level pseudo-labels by combining the ground truth gloss sequence labels with the predicted sequence. We demonstrate that our model achieves state-of-the-art performance among RGB-based methods on large-scale CSLR benchmarks, PHOENIX-2014 and PHOENIX-2014-T, while showing comparable results with better efficiency when compared to other approaches that use multi-modality or extra annotations.

1. INTRODUCTION

The Continuous Sign Language Recognition (CSLR) task aims to recognise a gloss[1] sequence in a sign language video [1, 2, 3]. To capture the meaning of the sign expressions from a signer, recent works obtain manual and non-manual expressions by fusing RGB with other modalities such as depth [4], infrared maps [5] and optical flow [6], or by explicitly extracting multi-cue features [2, 7, 8, 9] or human keypoints [10] using off-the-shelf detectors. However, using such extra components introduce bottlenecks in both training and inference processes. In addition, most CSLR datasets only have sentence-level gloss labels without frame- or gloss- level labels [2, 11, 12]. To overcome insufficient annotations, the Connectionist Temporal Classification (CTC) [13] loss has been traditionally opted to consider all possible underlying alignments between the input and target sequence. However, using the CTC loss without true frame-level supervision produces temporally spiky attention which can make the model fail to localise important temporal segments [14].

Accordingly, we develop *self-sufficient* framework for CSLR, which provides meaningful gloss supervision while capturing helpful multi-cue information *without additional modalities* or *annotations*. To this end, we propose two novel methods: Divide and Focus Convolution (DFConv) and Dense Pseudo-Label Refinement (DPLR). DFConv is a task-aware convolutional layer which extracts visual multi-cue features by dividing spatial regions to focus on partially specialised features. Note that DFConv is designed to leverage prior knowledge about the structure of human bodies without any additional networks or modalities. In addition, DPLR elaborately refines an initially predicted gloss sequence from the model by referring a ground-truth gloss, and propagates frame-level gloss supervision without additional networks, unlike [6, 15]. We emphasise that DPLR is generally applicable to other CSLR architectures or frameworks [1, 14] to bring performance gain by reducing missing glosses in predictions.

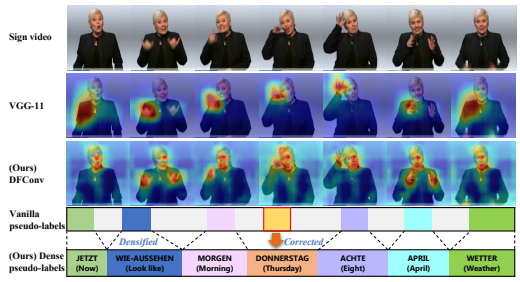

Fig. 1. Comparison of GradCAM between VGG-11 and DFConv, and an example of the generated pseudo-labels before and after DPLR. DFConv better highlights multiple individual elements (hands, faces) across the entire scene whereas VGG-11 simply highlights a small region (*i.e.*, the right hand). DPLR corrects the mispredicted gloss with the ground truth gloss (*e.g.*, red box in Vanilla pseudo-labels) and densifies the pseudo-labels with the nearest glosses, which results in a more informative supervision without external knowledge.

We extensively validate the effectiveness of DFConv and DPLR. We also show that the whole *self-sufficient* counterpart achieves state-of-the-art results among RGB-based methods and is comparable to other methods that use extra knowledge with better efficiency on two publicly available CSLR benchmarks [11, 12]. To summarise, our main contributions are as follows:

(1) We design a task-specific convolutional layer, named DFConv, that efficiently extracts non-manual and manual features without additional networks or annotations. (2) We also introduce DPLR, a novel pseudo-label generation method, to propagate frame-level supervision by using the combination of the ground truth gloss sequence and the predicted temporal segmentation information. (3) We conduct extensive experiments on two publicly available CSLR benchmarks, showing state-of-the-art performance compared to other RGB-based methods, and competitive results compared to other approaches that use multi-modality or additional knowledge with better efficiency.

2. RELATED WORKS

Multi-cue fusion methods for CSLR task can be categorised into *multi-semantic* and *multi-modal* methods. Multi-semantic works [3, 12, 8, 9] utilised hand-crafted or weak-labeled features such as detected hands, trajectories of hands, and body parts, then integrate these features into frames to predict the gloss sequences. On the other hand, multi-modal works [5, 4] use color, depth, and optical flow to extract orthogonal features. [6] proposed a multi-modality integration framework of appearance and motion cues by using both RGB frames and optical flow. Most recently, [10] fused human body keypoints extracted by an off-the-shelf network [16].

†Corresponding authors

[1]Glosses are the smallest units having independent meaning in sign language.

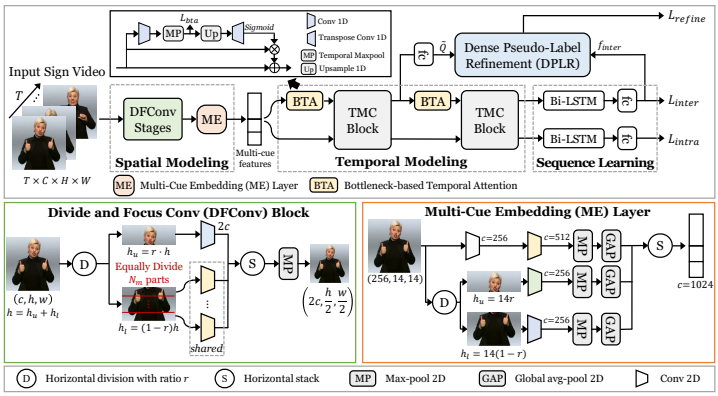

Fig. 2. **Overall architecture**. In Spatial Modeling, non-manual and manual features are extracted through DFConv followed by a Multi-Cue embedding layer. In Temporal Modeling, temporal features are extracted for gloss sequence prediction by integrating each element. Finally, the gloss probability vectors are obtained through sequence learning.

Unlike the methods listed above, we design DFConv that captures both manual and non-manual expressions from RGB video, without relying on *any additional hand-crafted features* or *multi-modal data*.

In addition, the CSLR task [2, 7, 15, 17, 18] naturally corresponds to weakly-supervised learning problem due to the lack of frame-level gloss annotations. The challenge lies in the ambiguous semantic boundary of the adjacent glosses from sign videos [2, 12, 19]. To address this issue, some works in CSLR field generate frame-level pseudo-labels from sparse gloss annotations [3, 20], which can be inherently noisy and reliant on the model's performance. Most recently, the CTC loss [13] is employed to facilitate end-to-end training of a deep learning model [21, 22], and consider all the feasible underlying alignments between the predictions and labels. However, as observed in [14, 23], directly optimising CTC can cause spiky attention in predictions, favoring more blank glosses. Recent works tackle this issue by balancing the blank output and meaningful glosses [1], and by directly supervising the visual features via visual alignment constraint [14] and mutual knowledge transfer [24]. In contrast, we propose Dense Pseudo-Label Refinement (DPLR) that provides dense and reliable supervision signals obtained by gloss predictions of the model to visual features.

3. METHOD

CSLR task aims to map a given input video to its corresponding gloss sequence $g = \{g_n\}_{n=1}^{N}$ with N glosses. As shown in Fig. 2, a sign video is fed into the *spatial modeling* module consisting of several Divide and Focus Convolution (DFConv) layers, and a multi-cue embedding layer to extract manual and non-manual features. The multi-cue features of all the frames are passed through the *temporal modeling* module, that is comprised of Bottleneck-based Temporal Attention (BTA), which captures more important information among adjacent frames, and Temporal Multi-Cue (TMC) blocks of [10]. Then, the output of last TMC block is passed through the *sequence learning* stage, which is composed of a Bi-LSTM layer [25] and FC layer to predict the gloss sequence from the final model output. Finally, the Dense Pseudo-Label Refinement (DPLR) module is introduced to effectively train the latent representations by generating corrected and densified frame-level pseudo-labels.

3.1. Divide and Focus Convolution

We observe from various CSLR datasets [11, 12, 26] that non-manual expressions occur frequently in the upper region of the image, while manual expressions occur mainly in the lower region. As shown in Fig. 1, despite the importance of both non-manual and manual expressions appearing in the entire image area, the conventional 2D convolution layer tends to capture the only one most dominant information (*i.e.*, right hand) over the whole image. To address this issue, we propose a novel Divide and Focus Convolution (DFConv) layer designed to independently capture non-manual features and manual features solely from RGB modality.

The structure of the DFConv is illustrated in Fig. 2. Inspired by the observation in [27], DFConv physically limits the receptive field that increases as the network deepens by subdividing an image [8] into upper (for non-manual expressions) and lower regions (for manual expressions) with the division ratio of r, where r is the ratio of spatial height of the upper region h_u to the original spatial height h given by $r = \frac{h_u}{h}$. To precisely capture the dynamic manual expressions, we further subdivide the lower region into N_m groups. For the upper region, this kind of subdivision is not required since non-manual expressions do not consist of the same amount of dynamics. We empirically observe that subdividing the upper region reduces the performance as well. Note that *different convolution weights are used for each upper and lower regions*, and *the weights are shared within the subdivided lower regions*. This helps the model to focus more on visually meaningful areas that represent complex sign expressions in the segmented image.

Unlike other methods that leverage external knowledge, we only introduce two hyper-parameters r and N_m, which make our method significantly more efficient. By virtue of simply splitting the frames horizontally, DFConv efficiently captures multi-cue features simultaneously without equipping costly human pose estimator like STMC [10] that increases model complexity and inference time (See project page). To further embed the outputs of stage 3 in Fig. 2 into the three individual multi-cue vectors (*i.e.*, full-frame, non-manual and manual), a simple and effective Multi-Cue Embedding (ME) layer is employed. The full-frame features containing global information are passed through two 2D convolution layers, and the remaining features (non-manual and manual) are passed through only single 2D convolution layer. Finally, all these features are vectorised by max pooling with a 2×2 kernel followed by an average pooling layer.

3.2. Dense Pseudo-Label Refinement

Most existing sign language datasets do not have temporally localised gloss labels [2, 11, 12, 28]. Due to the characteristics of the CTC loss used in training CSLR models without frame-level labels, the output sequence predictions of models are naturally induced to be sparse. As a result, it is difficult for CSLR models to receive direct and precise alignment supervision for each gloss token. In addition, without alignment supervision, CSLR models learn entire sequences as a whole instead of individual gloss words. This limits the robustness of models severely as

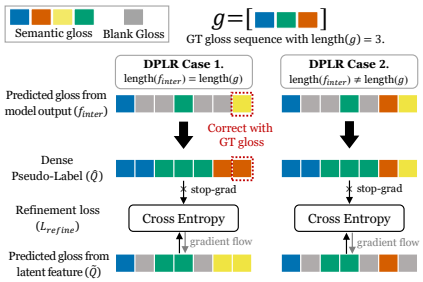

Fig. 3. Dense pseudo-label generation process in DPLR. **(Case 1)**: gloss sequence length predicted by the model and the ground truth are matched. **(Case 2)**: gloss sequence length predicted by the model and the ground truth is off by one word. DPLR provides latent features with *frame-level* supervision to compensate for the CTC loss while reducing the noise of pseudo-labels by the correction mechanism of Case 1.

they rely on entire sequences. In other words, models can easily confuse similar sequences with slightly different words. In order to mitigate these drawbacks, we introduce an additional training objective called Dense Pseudo-Label Refinement (DPLR) that uses the alignment information predicted by the model to generate Dense Pseudo-Labels (DPL). Then, the model is further refined with these generated pseudo-labels.

In DPLR process, we have two separate cases for generating DPL $\hat{Q}$ as illustrated in Fig. 3. We first compare the sequence length of non-blank predictions of the model with its corresponding ground truth gloss sequence. If the sequence length is matched, we go to **Case 1**, where we compare the predicted gloss sequence with the ground truth sequence. If a predicted gloss is wrong, we swap in correct gloss from the ground truth to increase the reliability of the pseudo-labels. As mentioned before, the predictions are sparse due to the nature of the CTC loss, and most of the predictions along the temporal axis are blanks. Here, we create DPL by filling each blank with the nearest predicted glosses. In the case where the predicted sequence length differs from the ground truth by one gloss length, we go to **Case 2**. Then, we simply densify the pseudo-label using the nearest gloss without swapping any glosses regardless of the correctness of the glosses. In the case that the sequence length differs by more than one gloss, we disregard that sequence as this might cause predictions of the model to degrade, so we do not propagate refinement loss L_{refine}.

Using pseudo-labels only from **Case 1** and **Case 2**, we refine the model with Cross Entropy (CE) loss on the latent features similar to [1] as follows:

$$L_{refine} = \text{CE}(\hat{Q}, \tilde{Q}), \tag{1}$$

where $\hat{Q}$ is dense pseudo-labels and, $\tilde{Q}$ is gloss probability acquired from latent features, which is the final output of the inter-cue path (See Fig. 2). Note that we demonstrate the efficacy of 'Densify' and 'Refine' processes in Table 4, and show that DPLR is generalisable to other models in our project page.

In addition, the quality of pseudo-labels generated from the model depend heavily on the model's performance. As the CSLR task aims to translate a sign language video into a gloss sequence by mapping several adjacent frames into one gloss, it is important to extract key frames in the video. Hence, we design the **Bottleneck-based Temporal Attention (BTA)** module to attend to the temporally salient frames among adjacent frames. BTA consists of a temporal-wise attention map using 1D convolution layers and a max pooling layer to capture the temporally salient frames. The CTC loss is then propagated to the bottleneck, after the max pooled features, hence the name is Bottleneck.

With our additional modules, our final loss function is as follows:

$$L_{total} = L_{inter} + \lambda_1 L_{intra} + \lambda_2 L_{refine} + \lambda_3 L_{bta}, \tag{2}$$

where, L_{inter}, L_{intra} and L_{bta} are all CTC losses. L_{bta} is the average of all the TMC block's CTC losses and λ_1, λ_2, and λ_3 are loss weights.

4. EXPERIMENTS

Dataset and Evaluation Metric. We conduct experiments on two publicly available CSLR benchmarks to validate our *self-sufficient* framework: PHOENIX-2014 [12] and PHOENIX-2014-T [11]. We adopt the Word Error Rate (WER)[2] [12] for evaluation. Furthermore, in our project page, we upload a demo video to visually demonstrate the effectiveness of DFConv.

Method	Extra Annotations	WER (%) Dev	WER (%) Test
DeepHand [17]	Hand	47.1	45.1
SubUNets [7]	Hand	40.8	40.7
Deep Sign [29]	Hand	33.3	38.8
Staged-Opt [21]	Hand	39.4	38.7
LS-HAN [2]	Hand	-	38.3
Align-iOpt [23]	-	37.1	36.7
SF-Net [30]	-	35.6	34.9
DPD+TEM [31]	-	35.6	34.5
cnn-lstm-hmm [3]	-	27.5	28.3
Re-sign [20]	-	27.1	26.8
DNF [6]	-	23.8	24.4
cnn-lstm-hmm [3]	Mouth	26.0	26.0
STMC [10]	-	25.0	-
SFL [22]	-	24.9	24.3
FCN [1]	-	23.7	23.9
DNF+SBD-RL [32]	-	23.4	23.5
DNF [6]	Flow	23.1	22.9
VAC [14]	-	21.2	22.3
CMA [33]	-	21.3	21.9
SMKD [24]	-	**20.8**	21.0
STMC [10]	Pose	21.1	**20.7**
Ours	-	**20.9**	**20.8**

Table 1. Comparison of performance in WER (%) on PHOENIX-2014 benchmark. Ours shows the comparable performances to the existing state-of-the-art methods using either pose [10] or algorithmic gloss segmentation [24] even without extra annotations.

Method	Extra Annotations	WER (%) Dev	WER (%) Test
cnn-lstm-hmm [3]	-	24.5	26.5
FCN [1]	-	23.3	25.1
SLRT [34]	-	24.9	24.6
SLRT [34]	Text	24.6	24.5
cnn-lstm-hmm [3]	Mouth+Hand	22.1	24.1
SMKD [24]	-	20.8	22.4
STMC [10]	Pose	**19.6**	**21.0**
Ours	-	**20.5**	**22.3**

Table 2. Comparison of performance in WER (%) on PHOENIX-2014-T benchmark. Our framework achieves the state-of-the-art performances among RGB-based approaches, while shows comparable performances with the pose-based multi-cue method [10].

4.1. Experimental Results

We compare our framework with recent CSLR methods on both PHOENIX-2014 [12] and PHOENIX-2014-T [11] benchmarks. Tables 1 and 2 show the WER scores, while we specify the type of either extra annotations or modalities used during training for each method.

PHOENIX-2014. Table 1 summarises the results on Dev and Test splits from PHOENIX-2014 for several CSLR baselines. First, Ours achieves the state-of-the-art performances on Test split among RGB-based approaches. In particular, Ours outperforms the recently proposed FCN [1], fine-grained labeling [22], VAC [14] with alignment supervision to visual features, and CMA [33] with both gloss and video augmentation. Moreover, Ours shows superior performance over several recent methods that explicitly require extra annotations for training [2, 3, 6], and comparable performances to SMKD [24] with algorithmic gloss segmentations and STMC [10] using pose annotations. Note that the proposed method does not require either extra annotations for acquiring the benefit to detect spatially important regions or additional networks for the refinement of pseudo-labels.

PHOENIX-2014-T. Table 2 shows the results on Dev and Test splits of PHOENIX-2014-T. Ours surpasses cnn-lstm-hmm [3] which is trained with both mouth and hand annotations, and even outperforms SLRT [34] that jointly learns sign recognition and translation task from both sign glosses and sentences. Ours also outperforms SMKD [24], a competing baseline using RGB modality, and shows comparable results to STMC [10].

4.2. Ablation Study

Component Analysis. In Table 3, we ablate each component of our method to investigate its effectiveness. In the first row of the table, we show the result of the baseline model with VGG-11 [35] architecture followed by three

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

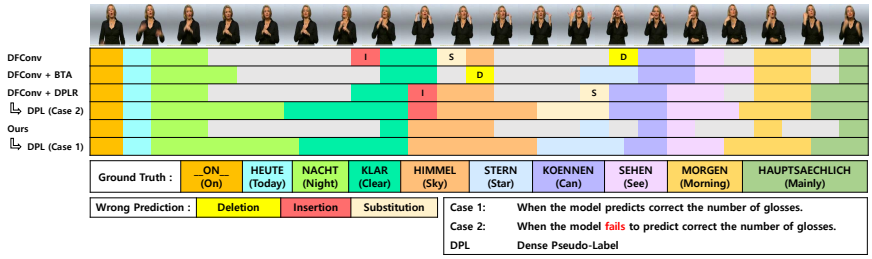

Fig. 4. Gloss predictions in a single sentence sign video from different network architectures (D: deletion, I: insertion, S: substitution). In the fourth and sixth rows, we further visualise two cases of Dense Pseudo-Labels (DPL). Applying the DPLR on the prediction greatly reduces the deletion phenomenon.

			WER (%)	
DFConv	DPLR	BTA	Dev	Test
			26.1	26.7
✓			24.5	24.5
✓	✓		22.4	22.5
✓		✓	24.2	24.1
✓	✓	✓	**20.9**	**20.8**

Table 3. Ablation study of DFConv, DPLR, and BTA. All the proposed components of our method gradually improve the performance.

			WER (%)	
w/ L_{refine}	Densify	Refine	Dev	Test
			24.2	24.1
✓			23.5	23.8
✓		✓	23.3	23.8
✓	✓		22.4	22.5
✓	✓	✓	**20.9**	**20.8**

Table 4. Ablation study on the design choice of DPLR. Both 'Densify' and 'Refine' processes are key in improving performance.

1D convolution layers. All components of our method consistently improve the performance altogether. In particular, when BTA is combined with DFConv, the performance improvement is marginal, but when combined with DPLR, it shows a large performance improvement. From this, we conclude that DPLR and BTA are complementary modules to each other. We ablate qualitatively the gloss predictions of each component in Fig. 4.

Design Choice of DPLR. The baseline in the first row of Table 4 is the same baseline in the fourth row in Table 3, which is the model trained with DFConv and BTA. 'Densify' and 'Refine' indicate whether the prediction from the model is filled by the nearest gloss prediction and whether glosses are replaced with ground truth glosses, respectively, as shown in Fig. 3. We show from the second and third row (without 'Densify') that directly leveraging the output of the model brings marginal improvements to the baseline. 'Densify,' which provides direct alignment supervision on the *frame-level* to the latent features is the key component for improving the model performance. Finally, the proposed Dense Pseudo-Labels (DPL), which is the combination of both 'Densify' and 'Refine' processes, shows the best performance by the correction mechanism with ground truth labels in 'Refine' to reduce the noise in *dense* pseudo-labels.

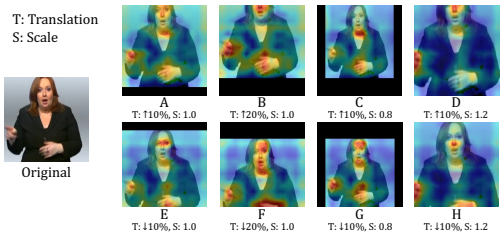

Fig. 5. Activation maps from Ours on *test-time* novel transformations.

Robustness of DFConv. *Our method is more robust where the signer is not bounded to a specific region at inference time than the state-of-the-art methods in practical cases.* To simulate such a scenario, we make a set of transformed data from PHOENIX-2014 Dev and Test splits, each of which includes a different degree of vertical translation (T) and scale operation (S). In Table 5, we list RGB based state-of-the-art method (VAC), the pose-based method (STMC), our model tested with the original division ratio ($r = 0.35$)

	VAC		STMC		Ours		Ours†	
Transform(T,S)	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Original	21.2	22.3	21.0	**20.7**	20.9	20.8	**20.9**	20.8
A: (↑10%, 1.0)	23.9	24.1	22.9	21.3	22.9	23.0	**21.1**	**21.2**
B: (↑20%, 1.0)	27.1	27.1	33.6	32.9	24.2	24.9	**22.7**	**22.5**
C: (↑10%, 0.8)	38.5	38.0	42.9	41.1	32.0	30.9	**28.5**	**28.1**
D: (↑10%, 1.2)	28.9	29.8	31.4	32.0	26.9	27.1	**24.8**	**24.5**
E: (↓10%, 1.0)	30.9	30.4	**24.4**	**24.1**	27.2	26.8	24.9	24.6
F: (↓20%, 1.0)	35.5	34.7	31.4	30.0	31.1	30.9	**29.7**	**29.9**
G: (↓10%, 0.8)	56.5	52.3	46.7	42.1	46.5	43.9	**38.1**	**37.9**
H: (↓10%, 1.2)	28.5	27.5	26.9	28.0	26.2	**26.2**	**26.1**	26.3
Average	33.7	33.0	32.5	31.4	29.6	29.2	**27.0**	**26.9**

Table 5. Robustness comparison with state-of-the-art methods in simulated real world scenario. We compare the WER on a model that has been trained on a train set without these transformations. Ours denotes a model with r set to 0.35, and Ours† denotes a model where r is moved along with the transformations during inference (T: vertical translation, S: scale). We note that STMC is our reproduction. We reimplement it as faithfully as possible.

(Ours), and our model where the r is changed along the corresponding transformation (Ours†). Although shifting the r gives the best performance (Ours†), in the real world, where we are not able to adjust on the fly r (Ours), the average performance of Ours still surpasses VAC and STMC.

We also present the activation maps of DFConv with the static division ratio ($r = 0.35$) in Figure 5. DFConv steadily captures non-manual and manual features even when half of the signer's face is out of focus (B, D) or when the signer's hands are partially out (D, F, H) with failure cases of pose-detectors shown in our project page. This shows that pose-based sign recognition methods are heavily reliant on the performance of the pose-detector.

5. CONCLUSION

In this paper, we propose two novel methods, DFConv and DPLR, that complement missing annotations in the existing weakly-labeled sign language datasets. To the best of our knowledge, we are the first to propose a method to extract manual and non-manual features individually by designing a task-specific convolution without any additional networks or annotations. In addition, we introduce DPLR module that does not require additional networks during the pseudo-labeling process and demonstrate its effectiveness through various experiments. The experimental results show that our framework achieves state-of-the-art performance on two large-scale benchmarks among RGB-based methods, and also outperforms or is comparable to methods based on multi-modality.

6. ACKNOWLEDGEMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT, 2022-0-00989, Development of Artificial Intelligence Technology for Multi-speaker Dialog Modeling for KAIST and 2020-0-01373, Artificial Intelligence Graduate School Program for Hanyang Univ.).

7. REFERENCES

[1] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai, "Fully convolutional networks for continuous sign language recognition," in *ECCV*, 2020, pp. 697–714. 1, 2, 3

[2] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li, "Video-based sign language recognition without temporal segmentation," in *AAAI*, 2018, vol. 32. 1, 2, 3

[3] 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," *TPAMI*, 2019. 1, 2, 3

[4] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network," in *CVPR*, 2016, pp. 4207–4215. 1

[5] Zhipeng Liu, Xiujuan Chai, Zhuang Liu, and Xilin Chen, "Continuous gesture recognition with hand-oriented spatiotemporal feature," in *ICCVW*, 2017, pp. 3056–3064. 1

[6] Runpeng Cui, Hu Liu, and Changshui Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1880–1891, 2019. 1, 3

[7] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," in *ICCV*, 2017, pp. 3075–3084. 1, 2, 3

[8] Dong-Jin Kim, Tae-Hyun Oh, Jinsoo Choi, and In So Kweon, "Dense relational image captioning via multi-task triple-stream networks," *TPAMI*, 2021. 1, 2

[9] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon, "Acp++: Action co-occurrence priors for human-object interaction detection," *TIP*, vol. 30, pp. 9150–9163, 2021. 1

[10] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li, "Spatial-temporal multi-cue network for continuous sign language recognition.," in *AAAI*, 2020, pp. 13009–13016. 1, 2, 3

[11] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden, "Neural sign language translation," in *CVPR*, 2018, pp. 7784–7793. 1, 2, 3

[12] Oscar Koller, Jens Forster, and Hermann Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *CVIU*, vol. 141, pp. 108–125, 2015. 1, 2, 3

[13] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376. 1, 2

[14] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen, "Visual alignment constraint for continuous sign language recognition," in *ICCV*, October 2021, pp. 11542–11551. 1, 2, 3

[15] Junfu Pu, Wengang Zhou, and Houqiang Li, "Dilated convolutional network with iterative optimization for continuous sign language recognition.," in *IJCAI*, 2018, vol. 3, p. 7. 1, 2

[16] Ke Sun et al., "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019. 1

[17] 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 *CVPR*, 2016, pp. 3793–3802. 2, 3

[18] Youngjoon Jang, Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, Joon Son Chung, and In So Kweon, "Signing outside the studio: Benchmarking background robustness for continuous sign language recognition," *arXiv preprint arXiv:2211.00448*, 2022. 2

[19] 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 *CVPR*, 2021, pp. 2735–2744. 2

[20] Oscar Koller, Sepehr Zargaran, and Hermann Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms," in *CVPR*, 2017, pp. 4297–4305. 2, 3

[21] Runpeng Cui, Hu Liu, and Changshui Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *CVPR*, 2017, pp. 7361–7369. 2, 3

[22] Zhe Niu and Brian Mak, "Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition," in *ECCV*, 2020, pp. 172–186. 2, 3

[23] Junfu Pu, Wengang Zhou, and Houqiang Li, "Iterative alignment network for continuous sign language recognition," in *CVPR*, 2019, pp. 4165–4174. 2, 3

[24] Aiming Hao, Yuecong Min, and Xilin Chen, "Self-mutual distillation learning for continuous sign language recognition," in *ICCV*, 2021, pp. 11303–11312. 2, 3

[25] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," *TSP*, vol. 45, no. 11, pp. 2673–2681, 1997. 2

[26] 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 *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–8. 2

[27] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He, "Gaitpart: Temporal part-based model for gait recognition," in *CVPR*, 2020, pp. 14225–14233. 2

[28] Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang, "Hierarchical lstm for sign language translation," in *AAAI*, 2018. 2

[29] Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden, "Deep sign: hybrid cnn-hmm for continuous sign language recognition," in *BMVC*, 2016. 3

[30] Zhaoyang Yang, Zhenmei Shi, Xiaoyong Shen, and Yu-Wing Tai, "Sf-net: Structured feature network for continuous sign language recognition," *arXiv:1908.01341*, 2019. 3

[31] Hao Zhou, Wengang Zhou, and Houqiang Li, "Dynamic pseudo label decoding for continuous sign language recognition," in *ICME*, 2019, pp. 1282–1287. 3

[32] Chengcheng Wei, Jian Zhao, Wengang Zhou, and Houqiang Li, "Semantic boundary detection with reinforcement learning for continuous sign language recognition," *TCSVT*, vol. 31, no. 3, pp. 1138–1149, 2020. 3

[33] Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li, "Boosting continuous sign language recognition via cross modality augmentation," in *MM*, 2020, pp. 1497–1505. 3

[34] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *CVPR*, 2020, pp. 10023–10033. 3

[35] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015. 3