Fact sheet: CVPR 2021 ChaLearn Looking at People Large Scale Signer Independent Isolated SLR Challenge

This is the fact sheet's template for the CVPR 2021 ChaLearn Looking at People Large Scale Signer Independent Isolated SLR Challenge [1]. Please fill out the following sections carefully in a scientific writing style. Then, send the compressed project (in `.zip` format), i.e., the generated PDF, `.tex`, `.bib` and any additional files to `juliojj@gmail.com`, and put in the Subject of the email "CVPR 2021 SLR Challenge / Fact Sheets", following the schedule and instructions provided in the Challenge webpage [1] "*Wining solutions (post-challenge), Fact Sheets*". Note, if you participated in both track, you will need to send one fact sheet per track.

I. Team details

- **Challenge Track**: RGB
- Team leader name: Hao Zhou
- Username on Codalab: rhythmblue
- Team leader affiliation: University of Science and Technology of China
- Team leader address: No.96, JinZhai Road Baohe District, Hefei, Anhui, 230026, P.R. China.
- Team leader phone number: +86-18861088366
- Team leader email: zhouh156@mail.ustc.edu.cn
- Name of other team members (and affiliation): Hezhen Hu, Weichao Zhao, Haoxin Sun, Wengang Zhou, and Houqiang Li. (University of Science and Technology of China)
- Team website URL (if any):

II. Contribution details

A. Title of the contribution

Sign language incorporates manual and non-manual cues to express the ideas of signers. In this challenge, we construct an ensemble framework composed of multiple neural networks (*e.g., I3D, SGN*) to conduct isolated sign language recognition. With extracted pose data, hand patch and face patch, Our simple framework ranked 3rd in the test phase of RGB track.

B. Introduction and Motivation

To decompose the multi-cue information, we first extract the pose data. Then, hand and face patches are cropped according to the keypoints. For patch sequence of full-frame, hands and face, 3D-CNNs are used to model the spatiotemporal information. For pose data, GCN-based method is selected to capture the skeleton correlation. During ensemble stage, we adopt late fusion for final predictions.

In [2], multi-channels of video streams, including color, depth and body joint positions are concatenated as input to the 3D CNN. Differently, we train networks separately for different cues and acquire predictions by late fusion. The pro is that the performance is improved by ensemble. The cons are summarized as two aspects. First, the computation burden is large both in training and inference stage. Second, the framework is naive for not considering sophisticated design of fusion strategy.

C. Representative image / workflow diagram of the method

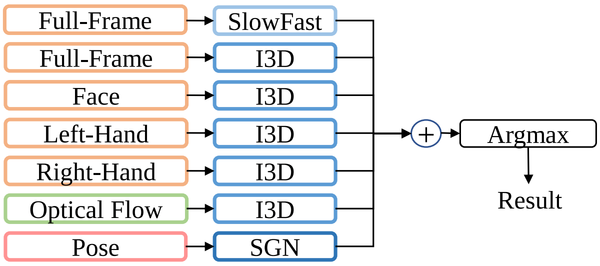

Fig. 1. The workflow diagram of the method.

D. Detailed method description

Preprocessing. In the original video, the body of signers only occupies about $1/3$ area of the image. So, we get the bounding box of signer with the MMDetection [3] and then crop the upper-body patch according to the joint positions estimated by HRNet [4]. Afterwards, we extract the full-body joint positions with MMPose [5] and crop the hand and face patch with the outer rectangle of keypoints. Finally, five types of data are generated, *i.e.*, full-body patch, left-hand patch, right-hand patch, face patch and full-body pose.

To process full-body patch, left-hand patch and right-hand patch, we separately use I3D [6] networks. We also train SlowFast [7] Network for full-body patch. To process full-body pose, we use SGN [8]. To process TV-L1 optical flow estimated from full-body patch, we use I3D. During inference, we sum all outputs before the SoftMax layers of the above networks with weights and select the category with the largest activation as result.

In our experiment, all networks are implemented by PyTorch [9] and trained on NVIDIA RTX 3090. During training, cross-entropy loss is utilized as the loss function for all networks.

For full frames, we use two networks to process. (1) We choose I3D-RGB with the pre-trained model provided

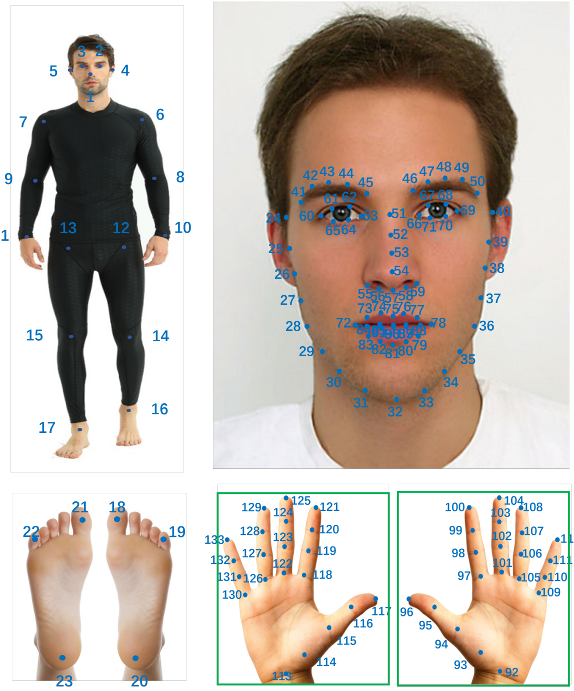

Fig. 2. The view of mmpose skeleton

by BSL-1K [10]. This network is trained with the SGD optimizer. The batch size, weight decay and momentum are set to 4, 1e-4 and 0.9, respectively. We start at the initial learning rate of 0.01 and reduce it by a factor of 0.1 when iterating to [15, 25, 40, 45, 50, 55, 60]. (2) SlowFast is used with the pre-trained model provide by I3D-Kinectics-Flow [6]. This network is trained with the SGD optimizer. The batch size, weight decay and momentum are set to 16, 1e-4 and 0.9, respectively. We start at the initial learning rate of 0.005 and reduce it by a factor of 0.1 when iterating to [25, 40, 45, 50, 55, 60].

For three significant parts of human body, *i.e.* face, left and right hands, we still use I3D-RGB with the pre-trained model provided by BSL-1K [10]. Except for the batch size set to 1, other parameter settings and optimizer selection are the same as I3D-RGB for full frames.

For the optical flow, we utilize I3D-Flow with the pre-trained model provided by I3D-Kinectics-Flow [6]. The network is trained with the SGD optimizer, in which the weight decay and momentum are set to 1e-4 and 0.9, respectively. We start at the initial learning rate of 0.01 and reduce it by a factor of 0.1 when iterating to [25, 40, 45, 50, 55, 60]. The batch size is set to 4.

For the pose sequence, we use SGN [8] with the pre-trained model on SLR500 [11]. The network is trained with the Adam optimizer. The batch size, weight decay and momentum are set to 64, 1e-4 and 0.9, respectively. We start at the initial learning rate of 0.001 and reduce it by a factor of 0.1 when iterating to [60, 80, 100]. In particular, as shown in Fig. 2, the skeleton points are made of 104 points, including the uppper body numbered from 1 to 11, and three detailed components numbered from 41 to 133, *i.e.*, face, left and right hands.

Due to aforementioned networks outputing the same dimension features, we use the simply late fusion merging various information for classification.

E. Challenge results and final remarks

Fill Table I with your obtained results, shown in the leaderboard of the challenge associated to the **Challenge Track** you defined in Section I (RGB[1] or RGB+D[2]). Note that if you joined the challenge in the test phase, keep the "development" row blank.

TABLE I
LEADERBOARD: RESULTS OBTAINED BY THE PROPOSED APPROACH.

Phase	**Track**	**Rank position**	**Rec. Rate**
Development			
Test	RGB	3	0.976200

III. ADDITIONAL METHOD DETAILS

Please reply if your challenge entry considered (or not) the following strategies and provide a brief explanation.

- **Did you use any kind of depth information (directly, such as RGBD data, or indirectly such as 3D pose estimation trained on RGBD data), either if during training or testing stage?** ( ) Yes, (✓) No
  If yes, please detail:
- **Did you use pre-trained models?** (✓) Yes, ( ) No
  If yes, please detail:
  I3D (full-body, left-hand, right-hand and face): `https://www.robots.ox.ac.uk/~vgg/research/bsl1k/data/experiments/bsl1k_i3d_m5_l20_kws8_ppose/model.pth.tar`
- **Did you use external data?** ( ) Yes, (✓) No
  If yes, please detail:
- **Did you use other regularization strategies/terms?** ( ) Yes, (✓) No
  If yes, please detail:
- **Did you use handcrafted features?** ( ) Yes, (✓) No
  If yes, please detail:
- **Did you use any face / hand / body detection, alignment or segmentation strategy?** (✓) Yes, ( ) No
  If yes, please detail:
  We use MMDetection to localize the bounding boxes of signers and MMPose to crop the hand and face patches.
- **Did you use any pose estimation method?** (✓) Yes, ( ) No
  If yes, please detail:

[1] RGB: `https://competitions.codalab.org/competitions/27901#results`

[2] RGB+D: `https://competitions.codalab.org/competitions/27902#results`

We use HRNet to help further localize the upper-body position of signers. We extract full-body joints of signers with MMPose.

- **Did you use any fusion strategy of modalities?** (✓) Yes, ( ) No
  If yes, please detail:
  We use late-fusion for networks of different visual cues.
- **Did you use ensemble models?** (✓) Yes, ( ) No
  If yes, please detail:
  Each model above are trained twice.
- **Did you use any spatio-temporal feature extraction strategy?** (✓) Yes, ( ) No
  If yes, please detail:
  We I3D and SlowFast to capture spatiotemporal information.
- **Did you explicitly classify any attribute (e.g. gender)?** ( ) Yes, (✓) No
  If yes, please detail:
- **Did you use any bias mitigation technique (e.g. rebalancing training data)?**
  ( ) Yes, (✓) No
  If yes, please detail:

IV. Code repository

Link to a code repository with complete and detailed instructions so that the results obtained on Codalab can be reproduced locally. This includes a list of requirements, pre-trained models, and so on. Note, training code with instructions is also required. This is recommended for all participants and mandatory for winners to claim their prize. **Organizers strongly encourage the use of docker to facilitate reproducibility**.

Code repository: `https://github.com/ustc-slr/ChaLearn-2021-ISLR-Challenge`

References

[1] ChaLearnLAP. CVPR 2021 ChaLearn Looking at People Large Scale Signer Independent Isolated SLR Challenge. [Online]. Available: http://chalearnlap.cvc.uab.es/challenge/43/description/

[2] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3d convolutional neural networks," in *2015 IEEE international conference on multimedia and expo (ICME)*, 2015, pp. 1–6.

[3] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[4] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[5] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[7] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6202–6211.

[8] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019, pp. 1–12.

[10] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, "Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues," in *European Conference on Computer Vision*, 2020, pp. 35–53.

[11] J. Huang, W. Zhou, H. Li, and W. Li, "Attention based 3D-CNNs for large-vocabulary sign language recognition," *TCSVT*, vol. 29, no. 9, pp. 2822–2832, 2019.