Hand-Model-Aware Sign Language Recognition

Hezhen Hu,[1] **Wengang Zhou,** [1, 2] **Houqiang Li** [1, 2]
[1] CAS Key Laboratory of GIPAS, EEIS Department, University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
alexhu@mail.ustc.edu.cn, {zhwg, lihq}@ustc.edu.cn

Abstract

Hand gestures play a dominant role in the expression of sign language. Current deep-learning based video sign language recognition (SLR) methods usually follow a data-driven paradigm under the supervision of the category label. However, those methods suffer limited interpretability and may encounter the overfitting issue due to limited sign data sources. In this paper, we introduce the hand prior and propose a new hand-model-aware framework for isolated SLR with the modeling hand as the intermediate representation. We first transform the cropped hand sequence into the latent semantic feature. Then the hand model introduces the hand prior and provides a mapping from the semantic feature to the compact hand pose representation. Finally, the inference module enhances the spatio-temporal pose representation and performs the final recognition. Due to the lack of annotation on the hand pose under current sign language datasets, we further guide its learning by utilizing multiple weakly-supervised losses to constrain its spatial and temporal consistency. To validate the effectiveness of our method, we perform extensive experiments on four benchmark datasets, including NMFs-CSL, SLR500, MSASL and WLASL. Experimental results demonstrate that our method achieves state-of-the-art performance on all four popular benchmarks with a notable margin.

Introduction

Sign language, as a natural language of the deaf community, has a unique linguistic characteristic. It conveys semantic meaning via hands, including hand motions, shape, orientation, *etc.*, together with non-manual features, including facial expressions. To facilitate the communication between the deaf and the hearing people, automatic sign language recognition (SLR) has been widely studied and attracted increasing attention. It aims at mapping the sign video into the text word or sentence, which corresponds to two subtasks, *i.e.,* isolated SLR and continuous SLR. Isolated SLR is a kind of fine-grained classification task and focuses on the recognition at the word level, while continuous SLR tries to recognize the signs in their presenting order. In this work, we focus on the former task, *i.e.,* isolated SLR.

The hand acts as a dominant role in sign language. As shown in Figure 1, it occupies a relatively small area, exhibiting highly articulated joints and similar appearance with fewer local characteristic features, when compared with the body or face. During the sign, it usually encounters the motion blur and self-occlusion among joints with complex backgrounds. Early works adopt hand-crafted features to describe hand gestures (Starner, Weaver, and Pentland 1998; Buehler, Zisserman, and Everingham 2009). Recently, many works have leveraged the advance of deep convolutional neural networks (CNNs) (Huang et al. 2019; Albanie et al. 2020; Koller et al. 2018; Cui, Liu, and Zhang 2019; Zhou et al. 2020). It is worth mentioning that some methods highlight the importance of hands by utilizing the cropped hands as the extra stream and achieve a notable performance gain (Camgoz et al. 2017; Huang et al. 2018; Koller et al. 2020). These deep-learning based methods work in a data-driven paradigm and learn feature representations adaptively under the supervision of the video-level category label.

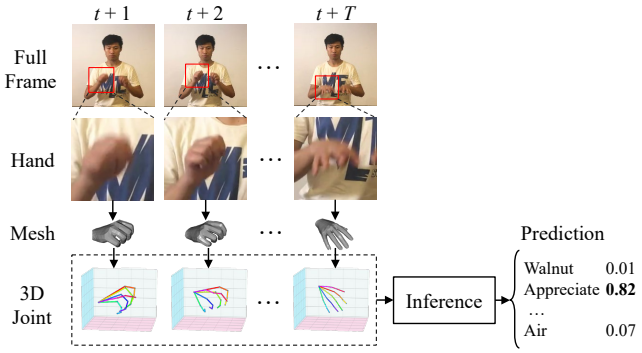

Figure 1: Illustration on the challenge of the hand gestures in sign language recognition and our idea with the modeling hand as the intermediate representation.

However, direct data-driven SLR methods suffer limited interpretability for the learned hand feature and may overfit under limited training data. The limited sign data sources are partially attributed to the fact that there is a strong requirement for expert knowledge during the manual annotation. Consequently, compared with current action recognition datasets (Goyal et al. 2017; Carreira and Zisserman 2017), sign language datasets, *e.g.,* WLASL (Li et al. 2020b), MSASL (Joze and Koller 2019) and NMFs-

CSL (Hu et al. 2020), usually contain much fewer samples per word.

To tackle this issue, we introduce the hand prior and propose a hand-model-aware framework for isolated SLR, with visible hand meshes and poses as the intermediate representation. The framework consists of three modules, *i.e.,* a visual encoder, a hand-model-aware decoder and an inference module. The visual encoder transforms the hand sequence into the latent semantic feature. Then the model-aware decoder provides a mapping from the latent feature to the hand mesh, as well as a compact pose. Specifically, the decoder is a fixed statistical mesh-based model, which stores the knowledge learned from a large variety of high-quality hand scans. In this way, the irrational poses can be effectively filtered out based on the imported hand prior. The inference module enhances the spatio-temporal representation of the hand pose sequence and performs recognition.

Our approach follows a paradigm in line with the insight (Clarke and Tyler 2015) on human cognition, which reveals that the ventral visual pathway in the brain treats the recognition process as a dynamic process of transformation from low-level visual input to specific conceptual knowledge representations. Due to the lack of hand-joint annotations in current sign datasets, we further focus on the spatial and temporal context of the pose representation, and design several weakly-supervised losses to guide its learning.

To our best knowledge, it is the *first* hand-model-aware framework for sign language recognition. Extensive experiments on four benchmark datasets, *i.e.,* NMFs-CSL, SLR500, MSASL and WLASL, validate the effectiveness of our method, achieving new state-of-the-art performance on all these datasets.

Related Work

In this section, we briefly review the related topics, including sign language recognition, hand pose estimation and hand models used for reconstruction.

Sign Language Recognition

Sign language recognition methods can be divided into two groups based on the input modality, *i.e.,* RGB-based (using the RGB video as input) and pose-based (using the skeleton sequence as input) methods.

RGB-based methods. Early methods rely on hand-crafted features, such as HOG, SIFT, motion trajectories, for hand representation (Buehler, Zisserman, and Everingham 2009; Koller, Forster, and Ney 2015; Yasir et al. 2015; Evangelidis, Singh, and Horaud 2014). Recently, deep convolutional neural networks (CNNs) have shown a high capacity for representation learning and been widely used in many computer vision tasks. Many researchers have explored the design of networks for video representation, *e.g.,* 2D-CNNs, 3D-CNNs or mixture of them (Carreira and Zisserman 2017; Chen et al. 2018; Qiu, Yao, and Mei 2017; Qiu et al. 2019; Simonyan and Zisserman 2014; Wang et al. 2016; Xie et al. 2018). For the task of sign language recognition, Koller *et al.* adopt 2D-CNNs for spatial representation, followed by HMM to model temporal dependencies (Koller et al. 2018). Some other works utilize 3D-CNNs for spatio-temporal representation modeling (Huang et al. 2019; Joze and Koller 2019; Li et al. 2020b,a; Albanie et al. 2020).

Pose-based methods. Besides the above mentioned RGB-based methods, many works study the pose-based methods. Pose is a type of well-structured data, a high-level semantic representation with a low dimension, which also enables the computation efficiency. Recurrent neural networks, *e.g.,* GRU (Cho et al. 2014) and LSTM (Hochreiter and Schmidhuber 1997), have been used to model the temporal information of the keypoint sequence (Du, Wang, and Wang 2015; Song et al. 2017; Zhu et al. 2016). Some CNN-based works attempt to transform the input keypoint sequence into the feature map and use the popular CNNs to capture spatio-temporal dynamics (Li et al. 2018; Cao et al. 2018). Considering the well-structured characteristic of the pose, more and more works adopt graph convolutional networks (GCNs) (Yan, Xiong, and Lin 2018; Shi et al. 2019; Zhang et al. 2020). Yan *et al.* (Yan, Xiong, and Lin 2018) make the first attempt to propose a spatial-temporal GCN for action recognition. Specifically, it builds a graph with nodes and edges pre-defined by human keypoints and their physical connections, respectively. These GCN-based methods are able to process pose data more efficiently and show promising results.

Hand Pose Estimation

There have been several works predicting hand poses from the RGB images. The 2D hand pose estimation has been greatly improved by multiview bootstrapping (Simon et al. 2017). Further improvement is achieved on the inference speed (Wang, Zhang, and Peng 2019). There also exist some works estimating 3D pose representations, *e.g.,* estimating 3D poses from 2D counterparts (Cai et al. 2019), constraining intermediate reconstructed depth (Iqbal et al. 2018), *etc.* Recent works learn 3D hand shape and pose jointly (Boukhayma, Bem, and Torr 2019; Ge et al. 2019; Zhang et al. 2019). These methods are all trained under the supervision of the hand-joint annotations and focus on the precise predictions of the joint positions. Different from them, our proposed recognition framework utilizes the hand poses as the intermediate representation and learn them without hand-joint annotations.

Hand Model Learning

To model the hand, many works have been proposed using various techniques, including shape primitives (Oikonomidis, Lourakis, and Argyros 2014; Qian et al. 2014), sum-of-Gaussians (Sridhar, Oulasvirta, and Theobalt 2013) and a more generalized sphere-meshes method (Tkach, Pauly, and Tagliasacchi 2016). To model the hand shape more precisely, some works (Ballan et al. 2012; Tzionas et al. 2016) propose to adopt a triangulated mesh with Linear Blend Skinning (LBS) (Lewis, Cordner, and Fong 2000). Da La Gorce *et al.* (de La Gorce, Fleet, and Paragios 2011) further introduce the scaling terms for each bone to change hand shape. MANO (Romero, Tzionas, and Black 2017) is the most popular fully-differentiable statistical model, which learns from a large variety of hand scans. It deforms the

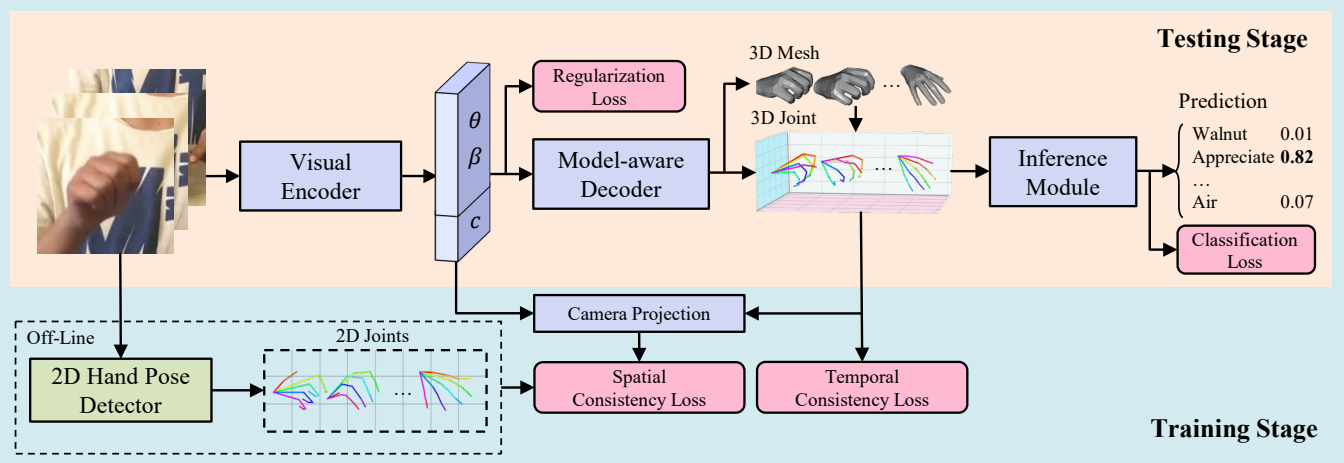

Figure 2: Overview of our proposed framework. The framework consists of a visual encoder, a hand-model-aware decoder and an inference module. Jointly with the video-level supervision, we further constrain the spatial and temporal consistency of intermediate 3D pose representations for further performance improvement. The modules utilized in training and testing stages are highlighted in light blue and orange, respectively.

mean mesh and factors the geometric changes into the shape and pose. In this work, we adopt MANO hand model into our framework to import the hand prior.

Our Approach

In this section, we first give a brief overview of our framework. Then we elaborate each component of our framework and the optimization objective functions of the framework.

Overview

As shown in Figure 2, given a cropped RGB hand sequence, the visual encoder first transforms it into the latent semantic embedding and predicts the camera parameters. Then the decoder works in model-aware and provides the mapping from the latent semantic feature to the refined 3D hand mesh and pose. The compact 3D pose representation is fed into the lightweight inference module. It enhances the representation of each joint and performs the final classification. The framework is optimized with a video-level cross-entropy loss, together with several weakly-supervised loss terms based on the spatial and temporal relationships of the intermediate poses.

Framework Design

The framework contains three key modules, *i.e.,* a visual encoder, a hand-model-aware decoder and an inference module. We will discuss these modules in the following.

Visual encoder. Given a RGB hand sequence $\mathbf{V} = \{\mathbf{v}_t\}_{t=1}^{T}$ with T frames from a sign video, the visual encoder $E(\cdot)$ transforms the RGB hand sequence into the latent semantic feature describing the hand status and the camera parameters, which is formulated as follows,

$$\mathbf{F_{la}} = \{\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{c}_r, \mathbf{c}_o, c_s\}_{t=1}^{T} = E(\mathbf{V}), \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^6$ and $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are the pose and shape embedding for the following decoder, while $\mathbf{c}_r \in \mathbb{R}^{3\times3}$, $\mathbf{c}_o \in \mathbb{R}^2$, and $c_s \in \mathbb{R}$ are the camera parameters, indicating the rotation, translation and scale, respectively. In our implementation, the encoder contains a ResNet34 (without the classifier) (He et al. 2016) to generate the high-dimensional feature, followed by a fully-connected layer to derive the low-dimensional semantic feature.

Hand-model-aware decoder. This module attempts to derive a compact pose representation from the latent semantic embeddings with a hand-model-aware method. With the encoded hand prior, the decoder constrains the distribution of possible poses and implicitly filters out the irrational predicted poses during its mapping. Finally, it produces a more compact and reliable hand pose, which will alleviate the optimization difficulty of the following inference module.

In this work, we utilize the fully differentiable MANO hand model (Romero, Tzionas, and Black 2017) as the decoder. MANO is a statistical model similar to the SMPL model (Loper et al. 2015), which is learned from a large variety of high-quality registered hand scans. In this way, the hand prior is encoded and a compact mapping can be established to describe the hand, *i.e.,* from the low-dimensional semantic embedding to the triangulated hand mesh $\mathbf{M} \in \mathbb{R}^{N\times3}$ of N=778 vertices and 1538 faces. More precisely, to generate a physically plausible mesh, the input pose and shape represent the coefficients of PCA components calculated from the collected hand scan data. The model is formulated as follows,

$$\mathbf{M}(\boldsymbol{\beta}, \boldsymbol{\theta}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}), \tag{2}$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}) = \bar{\mathbf{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}), \tag{3}$$

where $B_S(\cdot)$ and $B_P(\cdot)$ are blend functions, and $\mathbf{W}$ is a set of blend weights. The hand template $\bar{\mathbf{T}}$ is posed and skinned with the pose and shape corrective blend shapes, *i.e.,* $B_P(\boldsymbol{\theta})$ and $B_S(\boldsymbol{\beta})$. Further, the final mesh is generated by rotating each part around joints $J(\boldsymbol{\beta})$ using the linear skinning function $W(\cdot)$ (Kavan and Žára 2005).

With the hand model, the 3D joint locations $\widetilde{J}_{3D}$, as a more compact representation, can also be derived by the linear interpolation of relevant vertices in the mesh. It is notable that the original MANO model only provides 16 hand keypoints. To keep consistent with the 2D keypoints directly detected in the image plane, we select 5 extra vertices from the mesh with the index of 734, 333, 443, 555, 678 and add them as the fingertips. As a result, the hand is represented with 21 3D joints.

Inference module. The predicted pose sequence from the decoder may contain some unsatisfactory results. The inference module is utilized to refine its spatio-temporal representation. With the further calculation of adaptive attention, the inference module captures informative cues and performs the video-level classification.

The hand pose sequence is a well-structured data with the physical connections between joints, which makes it naturally to be organized as a spatio-temporal graph. In this work, we adopt a popular GCN (Yan, Xiong, and Lin 2018), which has proven effective to process pose data. Given a hand pose sequence $\widetilde{J}_{3D}$ representing 3D locations (x, y, z coordinates) of each joint in each frame, an undirected spatio-temporal graph $G(V, E)$ is first defined by V and E as the node and edge set, respectively. The node set V contains all the corresponding hand joints, while the edge set E includes the intra-frame and inter-frame set, *i.e.,* the physical connection of hand joints and connection of the same joint along the time, respectively. The adjacency matrix $\widetilde{\mathbf{A}}$ derived from the defined edge set will be adopted in GCN with the identity matrix $\mathbf{I}$. The graph convolution is formulated as follows,

$$Z = \sum_k \mathbf{D}_k^{-\frac{1}{2}} (\mathbf{A}_k \circ \mathbf{M}) \mathbf{D}_k^{\frac{1}{2}} \widetilde{J}_{3D} \mathbf{W}_k, \tag{4}$$

where Z is the output feature, k is the index of neighbour types (for each node, its neighbouring nodes are divided into several types), $\mathbf{W}_k$ is the convolution weight, $\widetilde{\mathbf{A}} + \mathbf{I}$ is dismantled into k sub-matrices, *i.e.,* $\widetilde{\mathbf{A}} + \mathbf{I} = \sum_k \mathbf{A}_k$, $\mathbf{T}_k = \mathbf{A}_k \circ \mathbf{M}$ and $\mathbf{D}_k^{ii} = \sum_j \mathbf{T}_k^{ij}$. The message is transferred among edges to enhance the representation of each joint. Further, the Hadamard product is performed between the learnable attention weight $\mathbf{M}$ initialized as all-one matrix and $\mathbf{A}_k$ to capture the discriminative cues. With several stacked GCN layers, a global pooling is adopted to merge the information contained in the enhanced node features, which is followed by a fully-connected layer to perform the final recognition.

Objective Function & Inference

Since current sign language datasets have no annotation on the hand pose, besides the cross-entropy classification loss $\mathcal{L}_{cla}$, we elaborately design several loss terms to guide the learning of intermediate pose representations.

Spatial consistency loss. First, we utilize the consistency between our predicted 3D and pre-extracted 2D joints from OpenPose (Cao et al. 2019; Simon et al. 2017). Specifically, we first project the predicted 3D joints to its 2D counterparts based on the weak-perspective camera model. The projection process can be formulated as follows,

$$\widetilde{J}_{2D} = c_s \prod (\mathbf{c}_r \widetilde{J}_{3D}) + \mathbf{c}_o, \tag{5}$$

where $\prod(\cdot)$ denotes the orthographic projection. Then we utilize the pre-extracted 2D hand joints J_{2D} as the pseudo label, and constrain the consistency between our projected one $\widetilde{J}_{2D}$ and J_{2D}. The spatial consistency loss is then calculated as follows,

$$\mathcal{L}_{spa} = \sum_{t=1}^{T} \sum_{j=1}^{21} \mathbb{1}(c(t,j) >= \epsilon) \left\| \widetilde{J}_{2D}(t,j) - J_{2D}(t,j) \right\|_1, \tag{6}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, and $c(t, j)$ denotes the confidence of the pre-extracted J_{2D} with the joint j at time t. To align the 2D hand joints predicted by different methods, we utilize the root-relative representation for these joints, *i.e.,* the root joint (palm) is set as the origin. It is notable that the joints in J_{2D} with the confidence $c(t, j)$ lower than the threshold ϵ will be ignored.

Temporal consistency loss. To avoid the jittering predictions, we further enforce the temporal consistency on the 3D hand pose. Different hand joints usually have different moving speeds during the sign, *e.g.,* joints closer to the palm usually have a lower speed. Thus we manually divide the hand joints into three groups, $\{S_i | i = 0, 1, 2\}$, *i.e.,* palm, middle and terminal joints, respectively. The temporal consistency loss is implemented by a derivative regularization, which is formulated as follows,

$$\mathcal{L}_{tem} = \sum_i \sum_{j \in S_i} \sum_{t=2}^{T} \alpha_i \left\| \widetilde{J}_{3D}(t,j) - \widetilde{J}_{3D}(t-1,j) \right\|_2^2, \tag{7}$$

where α_i denotes the pre-defined weight for S_i and we penalize more for the group having the lower speed.

Regularization loss. To ensure the hand model work in a proper way and generate the hand mesh plausibly, the regularization loss is added by constraining the magnitude of the partially latent feature, which is defined as follows,

$$\mathcal{L}_{reg} = \|\theta\|_2^2 + w_\beta \|\beta\|_2^2, \tag{8}$$

where w_β denotes the weighting factor.

The final objective loss function is defined as follows,

$$\mathcal{L} = \mathcal{L}_{cla} + \lambda_{spa}\mathcal{L}_{spa} + \lambda_{tem}\mathcal{L}_{tem} + \lambda_{reg}\mathcal{L}_{reg}, \tag{9}$$

where $\lambda_{spa}, \lambda_{tem}$ and λ_{reg} denote the weighting factor for spatial, temporal consistency loss and regularization loss, respectively. During training, the above loss function is utilized to optimize the full framework. Notably, both hands are involved and fused for the final recognition.

Inference. Considering only the cropped hands are insufficient to convey the full meaning of sign language, it is necessary to fuse recognition results based on hands with that on the full frame, which can be represented by either full keypoints or full RGB data. To this end, we use the results based on hand modeling, full keypoints and full RGB data. Those results can be assembled with late fusion by directly summing their prediction results (Karpathy et al. 2014). Specifically, for the recognition with the full keypoints, we utilize

ST-GCN as the backbone and all the 137 2D joints as input, while for the full RGB input, we sample a fixed number of frames and use a common CNN, *e.g.*, 3D-ResNet50, as the classifier. In the following, we refer our method with only hands, fusion of hands and the full keypoints, fusion of hands and the full RGB as **Ours (Hand)**, **Ours (Hand + Pose)** and **Ours (Hand + RGB)**, respectively.

Experiments

Datasets and Evaluation

Datasets. We evaluate our proposed method on four publicly available datasets, including NMFs-CSL (Hu et al. 2020), SLR500 (Huang et al. 2019), MSASL (Joze and Koller 2019) and WLASL (Li et al. 2020b).

NMFs-CSL is the most challenging Chinese sign language (CSL) dataset due to a large variety of confusing words caused by fine-grained cues. It contains a total of 1,067 words with 610 confusing words and 457 normal words. This dataset is recorded by a RGB camera at 30 FPS with a resolution of 1280×720. Specifically, 25,608 and 6,402 samples are used for training and testing, respectively.

SLR500 is another CSL dataset, which contains 500 daily words with 12,5000 recording samples performed by 50 signers. It is recorded by Kinect and provides RGB and depth modalities. There are 90,000 and 35,000 samples for training and testing, respectively.

MSASL is an American sign language dataset (ASL) with a vocabulary size of 1,000. It is collected from Web videos. It contains 25,513 samples in total with 16,054, 5,287 and 4,172 for training, validation and testing, respectively. Besides, in this dataset, the top-100 and top-200 most frequent words are selected as two subsets for training and testing, referred to as MSASL100 and MSASL200.

WLASL is an ASL dataset similar to MSASL, which is also collected from the Web. The size of the vocabulary is 2,000, and there are 21,083 samples divided into the training, validation and testing splits. MSASL and WLASL both bring new challenges due to the unconstrained recording conditions and limited samples for each word.

Notably, all these datasets adopt the signer-independent setting, *i.e.,* signers in the training set will not occur during testing. Besides, all the benchmark datasets only have category labels without any annotations on hand poses.

Evaluation. We evaluate the datasets using the accuracy metrics, including the per-instance and per-class metrics, denoting the average accuracy over each instance and each class, respectively. Since NMFs-CSL and SLR500 datasets have the same number of samples for each class, we only report the per-instance accuracy. Following the original settings in their corresponding works (Hu et al. 2020; Huang et al. 2019), we report top-1, top-2, top-5 accuracy for NMFs-CSL, and top-1 accuracy for SLR500. For MSASL and WLASL, we report the top-1 and top-5 accuracy under both per-instance and per-class metrics.

Implementation Details

In our experiment, all the models are implemented in PyTorch (Paszke et al. 2019) platform and trained on NVIDIA RTX-TITAN. Temporally, we extract 32 frames using random and center sampling during training and testing, respectively. During training, the input frames are randomly cropped to 256×256 at the same spatial position. Then the frames are randomly horizontally flipped with a probability of 0.5. During testing, the input video is center cropped to 256×256 and fed into the model. The model is trained with Stochastic Gradient Descent (SGD) optimizer. The weight decay and momentum are set to 1e-4 and 0.9, respectively. We set the initial learning rate as 5e-3 and reduce it by a factor of 0.1 when the validation loss is saturated. In all experiments, the hyper parameters ϵ, w_β, λ_{spa}, λ_{tem}, λ_{reg}, α_0, α_1 and α_2 is set to 0.4, 10, 0.1, 0.1, 0.1, 1, 2.5 and 4, respectively. We use OpenPose (Cao et al. 2019; Simon et al. 2017) to extract the full keypoints, *i.e.,* the 137 2D joints of body, face and hands. The extracted hand and shoulder keypoints are further utilized to crop the hand from the full frame. Besides, for the training of the RGB and pose baseline, we follow the original settings in their works (Carreira and Zisserman 2017; Yan, Xiong, and Lin 2018).

Cla.	Reg.	Spa.	Tem.	Top-1	Top-2	Top-5
✓				61.5	80.3	90.8
✓	✓			62.0	78.8	88.9
✓	✓	✓		64.0	81.6	90.7
✓	✓	✓	✓	**64.7**	**81.8**	**91.0**

Table 1: Ablation studies on the effect of each loss term on NMFs-CSL dataset. Cla., Reg., Spa. and Tem. denote the classification, regularization, spatial and temporal consistency loss, respectively.

Hand		Full frame		Accuracy		
OP	Ours	Keypoints	RGB	Top-1	Top-2	Top-5
✓				54.6	72.2	85.2
	✓			**64.7**	**81.8**	**91.0**
		✓		59.9	71.3	83.7
✓		✓		67.3	83.0	93.0
	✓	✓		**71.7**	**88.6**	**95.7**
			✓	62.1	73.2	83.7
✓			✓	71.7	84.3	92.3
	✓		✓	**75.6**	**88.4**	**95.3**

Table 2: Experimental results based on the hand modeling, full keypoints and full RGB data. For the hand-based method, we compare the results between our generated 3D hand pose and the 2D OpenPose-detected one (OP), which is utilized as the pseudo label in our framework.

Ablation Study

We perform ablation studies on the effectiveness of loss terms and the complementary effect of our method.

Effectiveness of loss terms. From Table 1, it can be observed that the top-1 accuracy is improved gradually by adding each loss term. Although the regularization loss brings relatively less improvement, it is crucial for the hand model to generate plausible meshes. It is notable that consistency losses contribute a lot to boosting the performance.

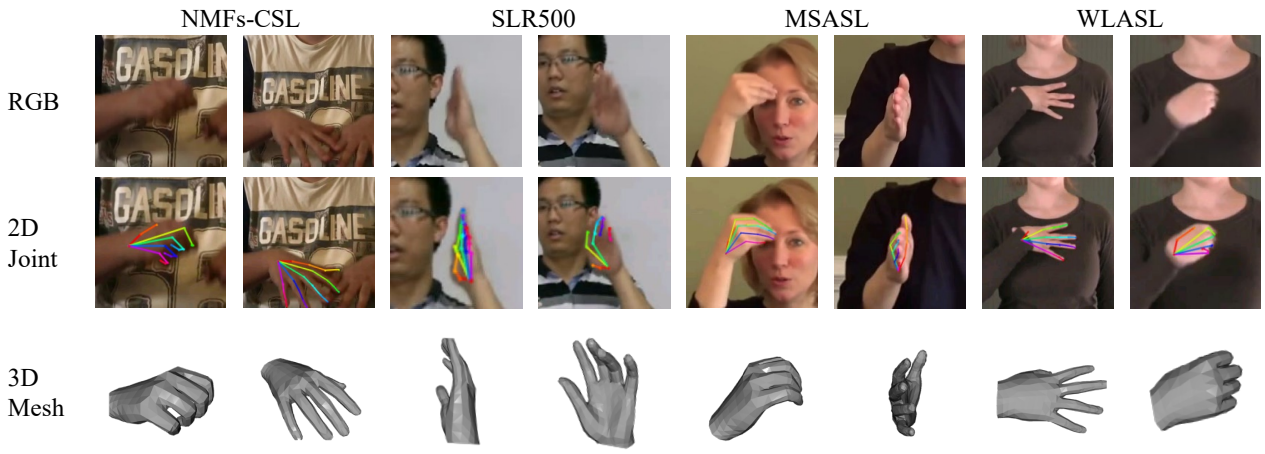

Figure 3: Visualization of the intermediate mesh representation. From the first to the third row, we present the RGB hand, 2D joint detected by OpenPose and the 3D mesh generated by our method. We visualize one sample in the test set for each benchmark dataset, including NMFs-CSL, SLR500, MSASL ans WLASL. For each sample, we visualize two key frames.

Method	Total			Confusing			Normal		
	Top-1	Top-2	Top-5	Top-1	Top-2	Top-5	Top-1	Top-2	Top-5
ST-GCN (Yan, Xiong, and Lin 2018)	59.9	74.7	86.8	42.2	62.3	79.4	83.4	91.3	96.7
3D-R50 (Qiu, Yao, and Mei 2017)	62.1	73.2	82.9	43.1	57.9	72.4	87.4	93.4	97.0
DNF (Cui, Liu, and Zhang 2019)	55.8	69.5	82.4	33.1	51.9	71.4	86.3	93.1	97.0
I3D (Carreira and Zisserman 2017)	64.4	77.9	88.0	47.3	65.7	81.8	87.1	94.3	97.3
TSM (Lin, Gan, and Han 2019)	64.5	79.5	88.7	42.9	66.0	81.0	93.3	97.5	99.0
Slowfast (Feichtenhofer et al. 2019)	66.3	77.8	86.6	47.0	63.7	77.4	92.0	96.7	98.9
GLE-Net (Hu et al. 2020)	69.0	79.9	88.1	50.6	66.7	79.6	93.6	97.6	99.3
Ours (Hand)	64.7	81.8	91.0	42.3	69.4	84.8	94.6	98.4	99.3
Ours (Hand + Pose)	71.7	**88.6**	**95.7**	54.2	**81.2**	**92.8**	95.0	98.5	99.5
Ours (Hand + RGB)	**75.6**	88.4	95.3	**59.7**	80.2	91.8	**96.9**	**99.4**	**99.9**

Table 3: Accuracy comparison on NMFs-CSL dataset.

The spatial consistency loss brings the largest accuracy gain, *i.e.,* from 62.0% to 64.0% top-1 accuracy. With the temporal consistency loss further added, the top-1 accuracy is improved to 64.7%. All the above results demonstrate the effectiveness of the proposed loss terms.

Complementarity between hand and full frame. The first part in Table 2 shows the classification results using hand keypoints as input based on the ST-GCN backbone. The first row denotes using the 2D hand keypoints detected by OpenPose, while the second row denotes our generated 3D ones. It can be observed that the accuracy using 3D hand keypoints as input largely outperforms that using 2D ones.

As indicated in Table 2, the top-1 accuracy increases from 59.9% to 71.7% when fusing recognition results of our hand joints and full keypoints. In contrast, when combined with the full-RGB based method, the accuracy improvement is 13.5%, which is larger than that combined with the full-keypoints based method.

Further, we also perform the qualitative visualization on the reconstructed hand mesh in Figure 3. The mesh also improves the interpretability of the whole framework. It can be observed that the video samples from different datasets vary a lot in their backgrounds and signer's clothing. The detection of 2D hand joints usually fails when the motion blur or self-occlusion occurs. In contrast, with the hand prior encoded, the generated mesh has more stability with all the fingers occurring and mostly reproduces the hand motion. It somewhat deals with some hard situations, *e.g.,* motion blur, mutually occurring of the hand and face, and self-occlusion.

Comparison with State-of-the-art Methods

We perform extensive experiments and compare with state-of-the-art methods on four benchmark datasets, *i.e.,* NMFs-CSL, SLR500, MSASL and WLASL.

Evaluation on NMFs-CSL. As shown in Table 3, the first two rows represent the baseline methods. DNF (Cui, Liu, and Zhang 2019) is a state-of-the-art method in continuous SLR and we utilize its visual encoder followed by a fully-connected layer as the backbone for comparison. GLE-Net (Hu et al. 2020) enhances discriminative cues from global and local views and achieves state-of-the-art performance. Compared with these competitors, our method (only cropped hands) achieves comparable performance with a majority of them. Our method ((Hand + Pose), (Hand +

Method	MSASL100				MSASL200				MSASL1000			
	Per-instance		Per-class		Per-instance		Per-class		Per-instance		Per-class	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
(Yan, Xiong, and Lin 2018)	59.84	82.03	60.79	82.96	52.91	76.67	54.20	77.62	36.03	59.92	32.32	57.15
(Joze and Koller 2019)[1]	-	-	81.76	95.16	-	-	81.97	93.79	-	-	57.69	81.05
(Li et al. 2020a)	83.04	93.46	83.91	93.52	80.31	91.82	81.14	92.24	-	-	-	-
(Albanie et al. 2020)	-	-	-	-	-	-	-	-	64.71	85.59	61.55	84.43
Ours (Hand)	73.45	89.70	74.59	89.70	66.30	84.03	67.47	84.03	49.16	69.75	46.27	68.60
Ours (Hand + Pose)	78.57	91.41	79.48	91.62	72.19	88.15	73.52	88.46	56.02	76.51	52.98	74.90
Ours (Hand + RGB)	**87.45**	**96.30**	**88.14**	**96.53**	**85.21**	**94.41**	**86.09**	**94.42**	**69.39**	**87.42**	**66.54**	**86.56**

[1] (Joze and Koller 2019) denotes the RGB baseline.

Table 4: Accuracy comparison on MSASL dataset.

Method	Accuracy
STIP (Laptev 2005)	61.8
GMM-HMM (Tang et al. 2015)	56.3
C3D (Tran et al. 2015)	74.7
Atten (Huang et al. 2019)	88.7
ST-GCN (Yan, Xiong, and Lin 2018)	90.0
3D-R50 (Qiu, Yao, and Mei 2017)	95.1
GLE-Net (Hu et al. 2020)	96.8
Ours (Hand)	95.9
Ours (Hand + Pose)	97.5
Ours (Hand + RGB)	**98.3**

Table 5: Accuracy comparison on SLR500 dataset.

Method	Per-instance		Per-class	
	Top-1	Top-5	Top-1	Top-5
(Yan, Xiong, and Lin 2018)	34.40	66.57	32.53	65.45
(Li et al. 2020b)[1]	32.48	57.31	-	-
(Albanie et al. 2020)	46.82	79.36	44.72	78.47
Ours (Hand)	37.91	71.26	35.90	70.00
Ours (Hand + Pose)	46.32	81.90	44.09	81.08
Ours (Hand + RGB)	**51.39**	**86.34**	**48.75**	**85.74**

[1] (Li et al. 2020b) denotes the RGB baseline.

Table 6: Accuracy comparison on WLASL dataset.

RGB)) outperforms the most challenging competitor GLE-Net, *i.e.,* 2.7% and 6.6% top-1 accuracy gain, respectively.

Evaluation on SLR500. As illustrated in Table 5, STIP (Laptev 2005) and GMM-HMM (Tang et al. 2015) denote the methods based on the hand-crafted features. Atten (Huang et al. 2019) utilizes multiple data modalities as input, including RGB, optical flow, depth, *etc.* The aforementioned GLE-Net (Hu et al. 2020) still achieves the best performance on this dataset. Even compared with GLE-Net, our method still achieves comparable performance. For our method ((Hand + Pose), (Hand + RGB)), the top-1 accuracy reaches 97.5% and 98.3%, which is new state-of-the-art performance on this dataset.

Evaluation on MSASL. MSASL contains limited samples for each word. The samples vary a lot in the resolution and unconstrained backgrounds, which makes MSASL more challenging. As shown in Table 4, we also release ST-GCN method as the pose baseline (Yan, Xiong, and Lin 2018). Compared with the RGB baseline, it shows inferior performance under both per-instance and per-class accuracy metrics. It may be caused by the failure of the pose detection, due to the partially occluded upper body of the signer, low-quality video, and noisy backgrounds. Albanie *et al.* (Albanie et al. 2020) and Li *et al.* (Li et al. 2020a) both use external sign videos to boost the performance and achieve state-of-the-art performance on MSASL or its subset, respectively. It is worth mentioning that our method outperforms the most challenging competitor by a notable margin, *i.e.,* 4.41%, 4.90% and 4.68% per-instance top-1 accuracy improvement on MSASL100, MSASL200 and MSASL1000 dataset, respectively. Besides, the complementary effects of our method are also validated on this dataset.

Evaluation on WLASL. Compared with MSASL dataset, WLASL has a vocabulary with doubled size but fewer samples. As shown in Table 6, when fused with the RGB baseline, our method achieves 51.39% top-1 per-instance accuracy, which brings 18.91% top-1 per-instance accuracy improvement over the RGB baseline. It also validates the effectiveness of our model-aware method under the dataset with limited samples. Compared with the most challenging competitor (Albanie et al. 2020), our method outperforms it by 4.57% and 4.03% top-1 per-instance and per-class accuracy improvement.

Conclusion

In this work, we introduce the hand prior and present the *first* hand-model-aware end-to-end framework for isolated sign language recognition. Our framework consists of three components, *i.e.,* a visual encoder, a hand-model-aware decoder and an inference module. The hand sequence is first transformed to the latent semantic feature, which is then processed by the hand-model-aware decoder to derive compact pose representations. Then the inference module refines the pose representations and performs recognition. Besides the video-level supervision, we guide the learning of the intermediate pose representation on its spatial and temporal consistency in a weakly-supervised way. Extensive experiments demonstrate the superiority of our method, achieving new state-of-the-art performance on all four benchmark datasets.

Acknowledgements

The work of Wengang Zhou was supported in part by the National Natural Science Foundation of China (NSFC) under Contract U20A20183 and 61632019, and in part by the Youth Innovation Promotion Association CAS under Grant 2018497. The work of Houqiang Li was supported by NSFC under Contract 61836011 and 62021001. The work is supported by MCC Lab of Information Science and Technology Institution, USTC.

References

Albanie, S.; Varol, G.; Momeni, L.; Afouras, T.; Chung, J. S.; Fox, N.; and Zisserman, A. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*.

Ballan, L.; Taneja, A.; Gall, J.; Van Gool, L.; and Pollefeys, M. 2012. Motion capture of hands in action using discriminative salient points. In *ECCV*, 640–653.

Boukhayma, A.; Bem, R. d.; and Torr, P. H. 2019. 3D hand shape and pose from images in the wild. In *CVPR*, 10843–10852.

Buehler, P.; Zisserman, A.; and Everingham, M. 2009. Learning sign language by watching TV (using weakly aligned subtitles). In *CVPR*, 2961–2968.

Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.-J.; Yuan, J.; and Thalmann, N. M. 2019. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In *ICCV*, 2272–2281.

Camgoz, N. C.; Hadfield, S.; Koller, O.; and Bowden, R. 2017. SubUNets: End-to-end hand shape and continuous sign language recognition. In *ICCV*, 3075–3084.

Cao, C.; Lan, C.; Zhang, Y.; Zeng, W.; Lu, H.; and Zhang, Y. 2018. Skeleton-based action recognition with gated convolutional neural networks. *TCSVT* 29(11): 3247–3257.

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *TPAMI* .

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? A new model and the Kinetics dataset. In *CVPR*, 6299–6308.

Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; and Feng, J. 2018. Multi-fiber networks for video recognition. In *ECCV*, 352–367.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, 1724–1734.

Clarke, A.; and Tyler, L. K. 2015. Understanding what we see: how we derive meaning from vision. *Trends in Cognitive Sciences* 19(11): 677–687.

Cui, R.; Liu, H.; and Zhang, C. 2019. A deep neural framework for continuous sign language recognition by iterative training. *TMM* 21(7): 1880–1891.

de La Gorce, M.; Fleet, D. J.; and Paragios, N. 2011. Model-based 3D hand pose estimation from monocular video. *TPAMI* 33(9): 1793–1805.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 1110–1118.

Evangelidis, G. D.; Singh, G.; and Horaud, R. 2014. Continuous gesture recognition from articulated poses. In *ECCV*, 595–607.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *ICCV*, 6202–6211.

Ge, L.; Ren, Z.; Li, Y.; Xue, Z.; Wang, Y.; Cai, J.; and Yuan, J. 2019. 3D hand shape and pose estimation from a single RGB image. In *CVPR*, 10833–10842.

Goyal, R.; Kahou, S. E.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The" Something Something" video database for learning and evaluating visual common sense. In *ICCV*, 5843–5851.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8): 1735–1780.

Hu, H.; Zhou, W.; Pu, J.; and Li, H. 2020. Global-local enhancement network for NMFs-aware sign language recognition. *TOMM* .

Huang, J.; Zhou, W.; Li, H.; and Li, W. 2019. Attention based 3D-CNNs for large-vocabulary sign language recognition. *TCSVT* 29(9): 2822–2832.

Huang, J.; Zhou, W.; Zhang, Q.; Li, H.; and Li, W. 2018. Video-based sign language recognition without temporal segmentation. In *AAAI*, 2257–2264.

Iqbal, U.; Molchanov, P.; Breuel Juergen Gall, T.; and Kautz, J. 2018. Hand pose estimation via latent 2.5D heatmap regression. In *ECCV*, 118–134.

Joze, H. R. V.; and Koller, O. 2019. MS-ASL: A large-scale data set and benchmark for understanding american sign language. *BMVC* .

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*, 1725–1732.

Kavan, L.; and Žára, J. 2005. Spherical blend skinning: a real-time deformation of articulated models. In *ACM I3D*, 9–16.

Koller, O.; Camgoz, C.; Ney, H.; and Bowden, R. 2020. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. *TPAMI* 42(9): 2306–2320.

Koller, O.; Forster, J.; and Ney, H. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *CVIU* 141: 108–125.

Koller, O.; Zargaran, S.; Ney, H.; and Bowden, R. 2018. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. *IJCV* 126(12): 1311–1325.

Laptev, I. 2005. On space-time interest points. *IJCV* 64(2-3): 107–123.

Lewis, J. P.; Cordner, M.; and Fong, N. 2000. Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In *SIGGRAPH*, 165–172.

Li, C.; Zhong, Q.; Xie, D.; and Pu, S. 2018. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *IJCAI*, 786–792.

Li, D.; Rodriguez, C.; Yu, X.; and Li, H. 2020a. Transferring cross-domain knowledge for video sign language recognition. In *CVPR*, 6205–6214.

Li, D.; Rodriguez, C.; Yu, X.; and Li, H. 2020b. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *WACV*, 1459–1469.

Lin, J.; Gan, C.; and Han, S. 2019. TSM: Temporal shift module for efficient video understanding. In *ICCV*, 7083–7093.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ToG* 34(6): 1–16.

Oikonomidis, I.; Lourakis, M. I.; and Argyros, A. A. 2014. Evolutionary quasi-random search for hand articulations tracking. In *CVPR*, 3422–3429.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 8026–8037.

Qian, C.; Sun, X.; Wei, Y.; Tang, X.; and Sun, J. 2014. Realtime and robust hand tracking from depth. In *CVPR*, 1106–1113.

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning spatio-temporal representation with pseudo-3D residual networks. In *ICCV*, 5533–5541.

Qiu, Z.; Yao, T.; Ngo, C.-W.; Tian, X.; and Mei, T. 2019. Learning spatio-temporal representation with local and global diffusion. In *CVPR*, 12056–12065.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ToG* 36(6): 245.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 12026–12035.

Simon, T.; Joo, H.; Matthews, I.; and Sheikh, Y. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 1145–1153.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 568–576.

Song, S.; Lan, C.; Xing, J.; Zeng, W.; and Liu, J. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 4263–4270.

Sridhar, S.; Oulasvirta, A.; and Theobalt, C. 2013. Interactive markerless articulated hand motion tracking using RGB and depth data. In *ICCV*, 2456–2463.

Starner, T.; Weaver, J.; and Pentland, A. 1998. Real-time American sign language recognition using desk and wearable computer based video. *TPAMI* 20(12): 1371–1375.

Tang, A.; Lu, K.; Wang, Y.; Huang, J.; and Li, H. 2015. A real-time hand posture recognition system using deep neural networks. *ACM TIST* 6(2): 1–23.

Tkach, A.; Pauly, M.; and Tagliasacchi, A. 2016. Sphere-meshes for real-time hand modeling and tracking. *ToG* 35(6): 1–11.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatio-temporal features with 3D convolutional networks. In *ICCV*, 4489–4497.

Tzionas, D.; Ballan, L.; Srikantha, A.; Aponte, P.; Pollefeys, M.; and Gall, J. 2016. Capturing hands in action using discriminative salient points and physics simulation. *IJCV* 118(2): 172–193.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 20–36.

Wang, Y.; Zhang, B.; and Peng, C. 2019. SRhandnet: Real-time 2D hand pose estimation with simultaneous region localization. *TIP* 29: 2977–2986.

Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking spatio-temporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 305–321.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 7444–7452.

Yasir, F.; Prasad, P. C.; Alsadoon, A.; and Elchouemi, A. 2015. SIFT-based approach on Bangla sign language recognition. In *ICCIA*, 35–39.

Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; and Zheng, N. 2020. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, 1112–1121.

Zhang, X.; Li, Q.; Mo, H.; Zhang, W.; and Zheng, W. 2019. End-to-end hand mesh recovery from a monocular RGB image. In *ICCV*, 2354–2364.

Zhou, H.; Zhou, W.; Zhou, Y.; and Li, H. 2020. Spatial-Temporal Multi-Cue Network for Continuous Sign Language Recognition. In *AAAI*, 13009–13016.

Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; and Xie, X. 2016. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In *AAAI*, 3697–3703.