Visual Intelligence

RESEARCH Open Access

Alleviating data insufficiency for Chinese sign language recognition

Wanli Xue[1], Jingze Liu[1], Siyi Yan[1], Yuxi Zhou[1], Tiantian Yuan[2*] and Qing Guo[3,4]

Abstract

Continuous Chinese sign language recognition (CCSLR) methods have shown their strong ability to learn excellent model architectures from datasets. However, due to data insufficiency, it is difficult to complete the CCSLR task. In this work, we focus on a simple but important solution to alleviate data insufficiency: how to refine the model architecture of a CCSLR network to improve the robustness of feature processing by using some better-quality non-Chinese sign language datasets. To this end, a simple empirical study was first conducted to verify the feasibility of knowledge transfer in the CCSLR task. Surprisingly, just by pre-training of our recognition model on a foreign sign language dataset, we can refine the model architecture and improve its robustness significantly. To make it more practical, the key issue of how to fine-tune the existing feature processing models for effective guidance should be carefully investigated. Then, we propose a novel scheme for fine-tuning of pre-trained models named FTP, which updates the spatial feature extractor initialized by a pre-trained backbone and freezes the temporal feature extractor implemented by a better shareable transformer encoder. Compared with the baseline method, our FTP method can achieve significant performance improvement on the public dataset USTC-CCSL.

Keywords: Computer vision, Continuous Chinese sign language recognition (CCSLR), Pre-training, Fine-tuning

1 Introduction

Different from hearing people, hearing-impaired people rely primarily on sign language for communication, which is a gesture system with linguistic properties gradually influenced by spoken language and words [1]. Like natural language, different countries have their sign languages, and similarities exist between them [2]. Therefore, research on Chinese sign language recognition is gradually receiving attention. Computer vision and natural language processing researchers process and analyze the gestures shown by different signers in sign language videos and translate a series of signs to the corresponding sign gloss sentences to facilitate the daily communication between the hearing-impaired people and the hearing people [3].

Many recent studies related to continuous Chinese sign language recognition (CCSLR) focused on model training from scratch on a small-scale dataset [4–7]. However, it is a well-known fact that almost all deep neural network models [8] that are trained from scratch are prone to overfitting due to the small-scale of the dataset that has been used [9]. Moreover, the widely used sign language dataset, USTC-CCSL, is insufficient for the CCSLR task [4]. Therefore, it is crucial to investigate how to alleviate data insufficiency for Chinese sign language recognition.

There are many ways to alleviate data insufficiency, such as data augmentation [10, 11] and pre-training [12], and an intuitive solution to the problem of lacking Chinese sign language datasets is to learn from other countries' sign language datasets. Research has found that CCSLR model performance can be promoted by pre-training on other cross-lingual datasets [13–15]. Benefiting from the conclusions in the study of sign language linguistics, it has been proven that similarities exist between sign languages in different countries, such as the same gestures and similar facial expression changes [2]. Hence, we propose using transfer learning methods to complete the tasks of shar-

*Correspondence: yuantt@tjut.edu.cn
[2]Technical College for the Deaf, Tianjin University of Technology, Tianjin, China
Full list of author information is available at the end of the article

Springer

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

ing knowledge between different countries' sign languages. By using some non-Chinese sign language datasets with larger scale and better recognition performance, we propose a novel scheme for fine-tuning of pre-trained models termed FTP for CCSLR. As illustrated in Fig. 1, this method focuses on feature processing in the CCSLR task. It consists of two parts: the spatial feature module for initialization and fine-tuning to extract the frames' spatial features and the temporal feature module for transferring learning to extract the frames' temporal features. To better extract the feature expression of each frame, we use ResNet18 [16] as the backbone. At the same time, considering the differences in the shooting background and appearance of sign language performers in different videos, we adopt the pre-trained ResNet18 model as the initial value of the FTP backbone and then fine-tune it.

To extract the temporal features of frame sequence, we not only notice single-frame information but also focus on contextual information. This is essential for the CCSLR task [17]. The current better transformer encoder [18, 19] is used as the temporal feature module. It has been proven that knowledge of transformer encoder representation is easier to share with other models [20], which is very conducive to transferring knowledge between sign language recognition tasks. Considering the similarities between sign languages in different countries and the size limitation of USTC-CCSL, we choose to freeze the transformer encoder from a better non-Chinese dataset pre-trained in the FTP scheme. In this way, the model can avoid overfitting on the USTC-CCSL dataset. The main contributions of this paper are listed as follows:

1) A novel CCSLR method is proposed, which avoids overfitting when training on a smaller dataset. The ablation study shows that the proposed FTP scheme adopted on the foreign sign language datasets brings a significant improvement in recognition performance compared to training using a single USTC-CCSL.
2) For feature processing, our proposed fine-tuning scheme is designed to initialize the parameters of the backbone from the pre-trained model and update the parameters to obtain a more robust spatial feature module. For the temporal feature module, a frozen pre-trained transformer encoder is used to share more reliable contextual information.

2 Related work

Considering that the focus of this paper is feature processing in CCSLR, we primarily introduce the research progress of CCSLR from two aspects of spatial feature processing and temporal feature processing in recent years. Another essential part of CCSLR, the alignment model, is also introduced in Sect. 2.3.

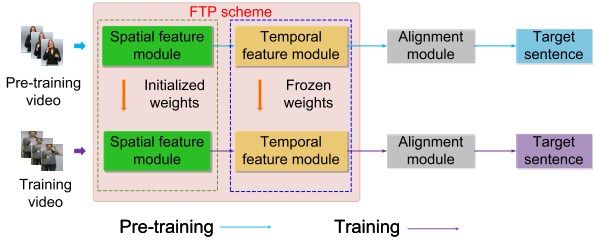


Figure 1 The pipeline of the proposed scheme for fine-tuning of pre-trained models termed FTP to enhance the spatial feature module and temporal feature module for addressing data insufficiency for CCSLR. Our main idea is to share useful information from other non-Chinese sign language datasets through transfer learning

2.1 Spatial feature processing in CCSLR

Feature extraction can be divided into single-cue processing and multi-cue fusion. The former focuses on a single visual factor, such as hands and body. The latter chooses a combination of multiple visual factors.

Some researchers [21, 22] utilized hand-crafted features for single thread feature extraction. For example, Zheng and Liang [23] extended the histogram of oriented gradient (HOG) descriptor to the 3D motion map based pyramid histogram of oriented gradient (M-PHOG) by adding a pyramid representation for CCSLR. Pu et al. [24] developed a curve feature descriptor depending on the shape of the trajectory based on the movement of the hands. At present, convolutional neural networks (CNNs) are applied to such image recognition tasks as CCSLR [25]. Li et al. [26] employed a ResNet-152 network that removes the final fully connected layer to extract the features of each frame.

Considering the coordination of visual factors in sign language, the idea of implicit cooperation of multiple cues has been gradually adopted [3]. Wang et al. [27] proposed a representation model based on the covariance matrix to fuse information from multimodal sources. Benefiting from the conclusions in the study of sign language linguistics, it has been proven that similarities exist between sign languages in different countries, such as the same gestures and similar facial expression changes [28]. At the same time, considering the differences in the shooting background and appearance of sign language performers in different videos, we developed the pre-trained ResNet18 model as the initial value of the FTP backbone and then fine-tuned it.

Guo et al. [29] proposed a new hierarchical deep recursive fusion (HRF) model for automatic identification of the red-green-blue (RGB) images and skeleton symbols. Zhang et al. [30] introduced a method of fusing trajectory probability and hand shape probability, specifically using a new feature of enhanced shape context (eSC) to represent the spatial and temporal information of the trajectory.

Zhou et al. [3] interpreted multi-semantics as parts of the body that extract and merge different meanings in their work. Their work adopted the VGG-11 model as the backbone network to generate multi-cue features of full frames, hands, faces, and poses.

Hu et al. [12] investigated cross-language continuous sign language recognition (CSLR) by introducing an additional shared sequential module to learn shared knowledge of different languages, and in contrast to their work, our scheme can achieve significant performance improvement without introducing an additional module.

2.2 Temporal feature processing in CCSLR

Temporal features are primarily used to model context information, which is important for CCSLR [3, 29, 31]. Cheng et al. [32] established a fully convolutional network (FCN) for online sign language recognition to simultaneously learn the spatial and temporal characteristics of sequences. The CSLR framework solution composed of a pre-trained visual module followed by a contextual module such as recurrent neural networks (RNNs) and a connectionist temporal classification (CTC) based alignment module for gloss generation has been used in CCSLR tasks. RNN is used to learn long-term information of spatial-temporal feature sequence and the CTC algorithm is used for mapping video sequence to ordered sign gloss sequence. For instance, Wei et al. [7] introduced a new multi-scale perception (MSP) strategy based on the CNN-RNN-CTC framework to learn the discriminative representation of video clips. In addition, Liao et al. [33] proposed a feature extraction module based on B3D ResNet, which extracts spatial-temporal features by inputting segmented video frames. To improve the recognition accuracy, Zhao et al. [34] put forward a 3D-CNN method combined with optical flow processing. Yang et al. [35] adopted the combination of LSTM and 3D convolution at the frame level to create an effective temporal modeling architecture. Pu et al. [36] used soft-DTW to constrain the long short-term memory (LSTM) results to obtain better annotations, and to fine-tune 3D-ResNet to extract better temporal features. Niu and Mak [28] developed a transformer encoder to extract the temporal information between video frames. The advantage of the transformer encoder is that the residual connections between layers are beneficial for better backpropagation.

2.3 Feature sequence alignment in CCSLR

In CCSLR, the alignment module is used to find the correspondence between the video feature sequence and the gloss annotation sequence [37]. At present, the method based on the CTC algorithm is adopted in most sign language recognition methods [38] for alignment [6, 32, 36, 39–42]. Aiming at the problem of inconsistency between the CTC objective and evaluation metric, Pu et al. [39] proposed cross-modality augmentation for their architecture. Zhou et al. [6] introduced a dynamic pseudo-label decoding method. The rational alignment path is found through dynamic programming. The model introduces blank classes through the CTC algorithm to filter out wrong labels and generate false labels that conform to the natural word order of sign languages.

Different from the above research, our FTP scheme focuses on mitigating the model overfitting problem caused by the small-scale dataset and the training method. A FTP scheme is proposed in this paper for CCSLR migration to achieve better feature processing.

3 Methodology

The pipeline of the proposed framework is illustrated in Fig. 2.

3.1 Overview of the proposed framework

Consider a processed sign language video $\mathcal{X} = \{\boldsymbol{x}_t\}_{t=1}^{T}$, where $\boldsymbol{x}$ and T represent the video frame and the number of frames in the video. First of all, in the spatial feature module, the pre-trained and updated fine-tuned ResNet18 network is adopted to extract the corresponding spatial features $\mathcal{Z} = \{\boldsymbol{z}_t\}_{t=1}^{T}$. Then, in the temporal feature module, the pre-trained and frozen fine-tuned transformer encoder network is used to extract the corresponding spatial-temporal features $\mathcal{C} = \{\boldsymbol{c}_t\}_{t=1}^{T}$. We have developed an alignment model from the stochastic fine-grained label (SFL) method [28]. The gloss sequence $\mathbf{y}$ corresponding to the video $\mathcal{X}$ is first input into the SFL model [28], and $\mathbf{y}$ is adjusted to the multi-state $\tilde{\mathbf{y}}$ by SFL's reinforcement learning algorithm. Next, in the alignment module, the gloss corresponding to the video frame $\boldsymbol{x}$ is determined by the probability $p(\tilde{\mathbf{y}}|\boldsymbol{x})$, which is calculated by the CTC algorithm [38]. Because our basic framework is developed based on SFL; therefore, we use SFL as our baseline method in this work.

3.2 Fine-tuning pre-trained scheme

To perform pre-training from non-Chinese datasets, we process the relevant parameters of the convergence model, including spatial and temporal features, and transfer them to the CCSLR task. In addition, we use the transfer learning method of Ref. [13] for reference and make some adjustments on this basis to verify the effectiveness of our method.

3.2.1 Spatial feature extractor module

The goal of CCSLR is to predict the corresponding gloss sequence $\mathbf{y} = \{\boldsymbol{y}_i\}_{i=1}^{K}$ $(K \leq M)$ based on the video sequence $\mathcal{X} = \{\boldsymbol{x}_t\}_{t=1}^{T}$. When transferring the parameters of the spatial feature model, we transfer the pre-trained w and b (see Eq. (1) for details) on the PHOENIX-2014 dataset to FTP's backbone for initialization and update it in the subsequent CCSLR training phase.

$$\boldsymbol{z} = \sum_{\text{chan}} \sum_{\delta u, \delta v} \boldsymbol{x}(u + \delta u, v + \delta v, c) \times \boldsymbol{w}(\delta u, \delta v) + \boldsymbol{b}, \tag{1}$$

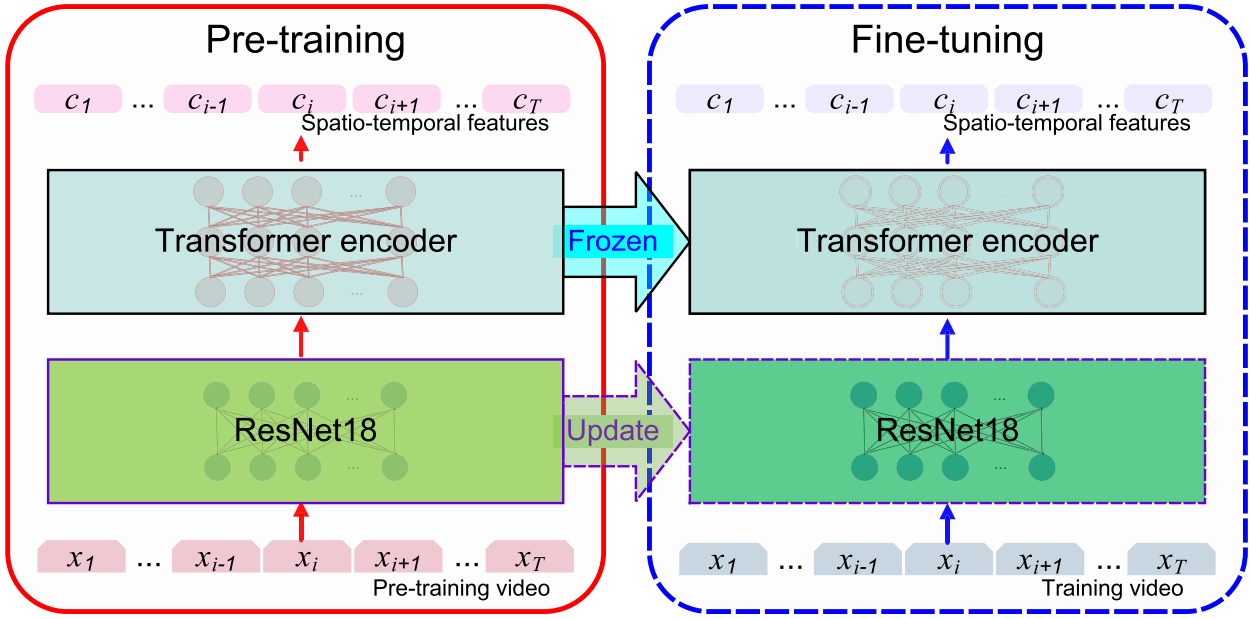


Figure 2 An illustration of our fine-tuning pre-trained scheme with two modules

where u and v denote the pixel coordinates of the video frame $\boldsymbol{x}$ with chan channels, respectively, and $\boldsymbol{b}$ is the bias term. The spatial neighbourhood, δ_u and δ_v is defined by the convolution layer's kernel size [43], and $\boldsymbol{w}$ is the weight of the convolution kernel.

3.2.2 Temporal feature extractor module

After obtaining the spatial features $\mathcal{Z} = \{\boldsymbol{z}_t\}_{t=1}^{T}$ of the video sequence $\mathcal{X} = \{\boldsymbol{x}_t\}_{t=1}^{T}$, we need to obtain further temporal features $\mathcal{C} = \{\boldsymbol{c}_t\}_{t=1}^{T}$ based on $\mathcal{Z} = \{\boldsymbol{z}_t\}_{t=1}^{T}$. In this work, a transformer encoder is adopted as the temporal feature extractor. It achieves the effect of updating features by calculating the weights of features at different times. First of all, three types of features $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ are mapped to $\mathcal{Z} = \{\boldsymbol{z}_t\}_{t=1}^{T}$ through three linear transformation matrices $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$.

We obtain Attention$(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ according to Eq. (2):

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{2}$$

where d_k represents a scaling factor.

After the model outputs $\mathcal{L}$, $\mathcal{L}$ enters a fully connected feed-forward network, and the output of the network is added to $\mathcal{L}$ and normalized, as explained in Eq. (3). Then, the spatial-temporal feature $\mathcal{C}$ is obtained:

$$\mathcal{L} = \text{LayerNorm}\big(\mathcal{Z} + \text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})\big), \tag{3}$$

$$\mathcal{C} = \text{LayerNorm}\big(\mathcal{L} + \max(0, \mathcal{L}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2\big). \tag{4}$$

We transfer the seven parameters, $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, $\boldsymbol{W}_V$, $\boldsymbol{W}_1$, $\boldsymbol{b}_1$, $\boldsymbol{W}_2$, and $\boldsymbol{b}_2$, pre-trained on the PHOENIX-2014 dataset to the CCSLR task. In particular, the fine-tuning method is frozen.

3.3 Discussion

At present, few datasets can be selected for the CCSLR task. Training a deep neural network on a small-scale dataset is prone to overfitting. To alleviate data insufficiency for CCSLR at the method level, we learn from the mature fine-tuning pre-trained scheme in transfer learning, which focuses on feature processing in CCSLR. It alleviates the problem of overfitting by learning "knowledge" from better non-Chinese datasets, including spatial feature description and temporal feature description. Taking into account the differences in the image level of sign language speakers in different datasets (such as clothing and background), we choose to use the initialized pre-trained backbone model and further update it on the Chinese sign language dataset; at the same time, based on the premise that the temporal feature has a certain generalization ability, we freeze the transformer encoder parameters from Pheonix-2014 pre-trained on the CCSLR model.

4 Experiments

4.1 Implementation details

In the pre-training stage, for the pre-training of ResNet18, we remove its fully connected layer and set the output size to 512. For the pre-training of the transformer encoder, we adopt a 2-layer transformer encoder with four heads

and set the dimension parameter of the position-wise feed-forward layer to 2048 and the dimension parameter of the model to 512.

In the model optimization stage, the Adam optimizer is used. The learning rate of the optimizer is set to 0.0001, and the weight decay is 1×10^{-4}. In addition, we set the batch size to 4. We let the model iterate 100 times and save the model once after each epoch. After testing, the model with the best test result will be chosen as our pre-trained model.

In the fine-tuning stage, we transfer the seven parameters of the pre-trained model (see above text for the specific meaning of the parameters) to the CCSLR task. In particular, due to the frozen fine-tuning method, we set the corresponding frozen parameter learning rate to 0 in the CCSLR model in the transformer encoder. The model stops training after the 100th iteration. After the training is over, we test the saved model and select the model with the best performance as our FTP model.

The proposed CCSLR method is implemented using the PyTorch framework on an NVIDIA RTX2080Ti GPU.

4.1.1 Datasets and evaluation

The PHOENIX-Weather-2014 dataset [44] is a German sign language dataset, which was captured from sign language interpretations of weather forecasts. The resolution of the video is 210×260 pixels. The size of the vocabulary is 1081. It contains a total of 6841 videos. There are a total of nine sign language speakers for demonstration in this dataset.

The PHOENIX-Weather-2014-T dataset [45] is an extension of the PHOENIX-2014 dataset. It is designed for sign language translation, but it can also be used to evaluate CSLR tasks. Similar to the PHOENIX-2014 dataset, the video resolution of this dataset is 210×260 pixels and there are also nine signers in the video sample of PHOENIX-2014-T. The vocabulary size of this dataset is 1,066 words. There are 8257 video samples in total.

The RWTH-Boston-104 dataset [46] was collected by the National Center for Sign Language and Gesture Resources of Boston University. The dataset was captured by three black and white cameras and one color camera in multiple perspectives at the same time. The resolution of the video is 312×242 pixels. The dataset consists of a total of 201 annotated ASL videos. It divides the video into a training set and a test set. The training set includes 161 sentences, and the test set includes 40 sentences.

The GSL dataset [47] has a total of 40,785 video samples and a total of seven sign language speakers to demonstrate. The video resolution of GSL is 848×480. There are 310 words in total.

The USTC-CCSL dataset [4] was recorded using Kinect devices. The resolution of the video is 1280×720 pixels. A total of 50 signers participated in the video recording. The dataset has 25,000 videos in total.

4.1.2 Evaluation indicator

Similiar to most CCSLR studies in the literature, we choose word error rate (WER) as the evaluation index. The minimum sum of the replacement, insertion, and deletion operations is performed to convert the recognized sequence into the corresponding sign language annotation, which is described as follows:

$$\mathrm{WER} = \frac{\#\mathrm{sub} + \#\mathrm{ins} + \#\mathrm{del}}{\#\mathrm{reference}}, \tag{5}$$

where #sub, #ins, and #del represent the least operations of substitution, insertion, and deletion, respectively, required to convert the hypothetical sentence to the reference sentence.

It is worth noting that we have used SPLIT-II [48] for the identification task, which is more realistic. The specific reasons are as follows:

1) SPLIT-I task pays more attention to distinguishing signers, which reflects the robustness of the model's backbone. When the dataset is small, it is prone to overfitting. It is difficult to make a difference in WER indicators of different recognition methods.

2) In the SPLIT-II task, sentences in the test set do not appear in the training set, but the gestures in the test set are included in the training set. This feature is similar to the most influential PHOENIX-2014 dataset.

4.2 Comparisons with the state-of-the-art CCSLR methods

We evaluate our FTP method on the commonly used public dataset USTC-CCSL and report the comparison results with a large number of state-of-the-art CCSLR methods, such as CMA [39], STMC [3], and SFL [28]. At the same time, for fairness, we directly use the experimental results in these papers, except for the baseline (among the methods listed in Table 1, only baseline (SFL) has open source code; however, it has not been tested on the USTC-CCSL dataset).

For baseline (SFL) [3] and our FTP method, we fix the parameters of these two methods on the USTC-CCSL dataset

Table 1 Comparison of our FTP method with the state-of-the-art CCSL methods on the USTC-CCSL SPLIT-II task. WER represents word error rate

CCSLR method	WER on USTC-CCSL SPLIT-II task
HLSTM-attn [29]	64.1
CTM [5]	61.9
DenseTCN [48]	44.7
IAN [36]	32.7
STMC [3]	28.6
CMA [39]	24.5
Hu et al. [12]	18.1
Zhao et al. [49]	13.7
SFL [28](Baseline)	28.3
FTP(Ours)	**24.5**

and report the results in terms of the official evaluation measure recommended by USTC-CCSL. Table 1 lists the comparison results, where the evaluated CCSLR methods are ranked in descending order according to their WER. We can observe that our FTP method achieves the same performance as the CMA [39] method.

To vividly illustrate the superiority of our FTP method, we have selected three representative videos from the USTC-CCSL test dataset. By analyzing the challenging video sequences processed by FTP, as demonstrated in Fig. 3, we found that our FTP method performs well in the following three types of typical problems:

In the 96th video, FTP handles better recognition of auxiliary words, because in Chinese sign language, the gestures of auxiliary words are usually simple and short. The 7th video shows that FTP's backbone is more robust. Although the baseline captured this gesture, it generated an incorrect gloss. The gestures of "policeman" and "nurse" are indeed similar. In the 15th video, our FTP and baseline did not obtain full marks, but we only deviated from the two gestures of "ruler" and "country" (these two words have a common gesture of raising both hands in Chinese sign language), and the baseline showed more errors.

It can be observed that on the USTC-CCSL dataset, our method has achieved significant performance improvement on the SPLIT-II task. This means that our FTP method is effective. What needs to be explained is that our FTP method listed in the Table 1 has been pre-trained and fine-tuned on the PHOENIX-2014 dataset.

4.3 Comparison with different pre-trained datsets

To further analyze the strength and the weakness of the proposed FTP method on different pre-trained datasets, we further evaluate it on different non-Chinese sign language datasets. We fine-tune different pre-trained models on the USTC-CCSL SPLIT-II task. The results are displayed in Table 2. We can observe that the PHOENIX-2014 dataset performs best, while PHOENIX-2014-T, which fo-



Figure 3 Recognition results of the baseline method and our method. The red rectangle indicates an error in the recognition. It can be seen from the 96th video (top) that our method corrects an insertion error, from the 7th video (middle) that our method corrects a substitution error, and from the 15th video (bottom) that our method corrects two errors even though it also makes an incorrect prediction

Table 2 Comparison of the transfer effects of different non-Chinese sign language datasets. WER represents word error rate

Pre-trained dataset	WER on USTC-CCSL SPLIT-II task
PHOENIX-2014-T	30.40
RWTH-BOSTON-104	28.34
GSL	26.08
PHOENIX-2014	**24.55**

cuses on translation tasks, has the least performance improvement. From the results of the experiment, it can be found that the recognition task requires special recognition datasets, and PHOENIX-2014 is a more suitable dataset for transferring knowledge to Chinese sign language.

Therefore, PHOENIX-2014 is selected as the pre-trained dataset in our proposed FTP method.

4.4 Ablation study of FTP

After determining the pre-trained dataset, the next step is to select the appropriate fine-tuning method to obtain the appropriate spatial and temporal features. Through the display of the ablation experiments in Table 3 and Table 4, we use the experimental data to answer why we should choose the respective specific fine-tuning forms for the backbone and transformer encoder modules of our FTP.

4.4.1 Effect of the spatial feature extractor

For fairness, we first pre-train the backbone on four typical non-Chinese sign language datasets and then transfer their backbones to the baseline method. Afterward, the initialized backbones are fine-tuned on the USTC-CCSL training dataset by freezing or updating, and at this time, the transformer encoder weights are kept updated. As presented in Table 3, it is found that for the spatial feature extractor, updating the transferred backbone parameters is more conducive to the CCSLR task. Because the backbone focuses more on the appearance model of the video frame, and different datasets often bring different appearance representations due to different signers and environments, it needs to be updated in time. It is worth noting that using the pre-trained backbone for initialization can not only alleviate overfitting but also shorten the training time.

In this ablation experiment, PHOENIX-2014-T obtained the best result, which is mainly due to the following factors. Compared with other datasets, PHOENIX-2014-T is not only of better quality but also contains more samples. Although the RWTH-BOSTON-104 is not large in scale, it still achieves a better WER. The main reason is that RWTH-BOSTON-104 is a dataset with multi-view shooting, which makes it easier to obtain the representative frames.

Table 3 Ablation results of our FTP regarding ResNet18's weight state on the USTC-CCSL SPLIT-II task. In each different pre-trained dataset, the best results are highlighted in bold

Pre-trained dataset	ResNet18's weight state	
	Frozen	Update
RWTH-BOSTON-104	54.22	**37.68**
GSL	41.38	**33.79**
PHOENIX-2014	40.01	**33.53**
PHOENIX-2014-T	38.02	**31.32**

Table 4 Ablation results of our FTP regarding the transformer encoder's weight state on the USTC-CCSL SPLIT-II task. For each different pre-trained dataset, the best results are highlighted in bold

Pre-trained dataset	Transformer encoder's weights state	
	Frozen	Update
PHOENIX-2014-T	**30.40**	31.32
RWTH-BOSTON-104	**28.34**	37.68
GSL	**26.08**	33.79
PHOENIX-2014	**24.55**	33.53

4.4.2 Effect of the temporal feature extractor

Table 4 presents that frozen weights can well assist the temporal feature extractor, and at this time, the ResNet18's weights are kept updated. Experiments on four pre-trained datasets have confirmed that the freezing transformer encoder improves the recognition results more significantly. Compared with GSL and RWTH-BOSTON-104, PHOENIX-2014 has more gesture vocabulary and videos, and gloss annotations are more suitable for recognition tasks. PHOENIX-2014-T, which mainly serves the translation task, is different from the recognition task in the form of annotation and dataset division, which leads to the pre-trained temporal feature extractor unable to complete the recognition task.

It is noted that our FTP is primarily composed of the spatial feature module and the temporal feature module. In particular, the FTP method composed of these two modules has been completed on the PHOENIX-2014 dataset. Therefore, the selection of the PHOENIX-2014 dataset as the source dataset for transfer learning confirms the importance of temporal feature processing for the CCSLR task.

5 Conclusion

This paper presents a novel scheme for fine-tuning of pre-trained models named FTP for the CCSLR task with a couple of novel spatial feature module and temporal feature module designs. Our method aims to alleviate the problem of data insufficiency for Chinese sign language recognition. First, we adopt the fine-tuning method of updating the backbone model, which has been initialized by the pre-trained scheme. Second, we propose a frozen method for

the temporal feature extractor module. These two innovations ensure that high-quality features can be obtained.

Acknowledgements
The authors are grateful to all reviewers for constructive comments.

Funding
This work was supported by the National Natural Science Foundation of China (Nos. 62376197, 62020106004, 92048301, and 62202332).

Abbreviations
CCSLR, continuous Chinese sign language recognition; CNNs, convolutional neural networks; CTC, connectionist temporal classification; FCN, fully convolutional network; FTP, fine-tuning pre-trained; HOG, histogram of oriented gradient; LSTM, long short-term memory; WER, word error rate.

Availability of data and materials
The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors have no competing interests to declare that they are relevant to the content of this article.

Author contributions
WX proposed the overall idea of the article and was responsible for writing and revising the paper. JL was responsible for completing the algorithm design and revising and reviewing the paper. SY was responsible for completing the experiments and writing the paper. YZ was responsible for revising and reviewing the paper, and TY and QG were involved in the discussion and review of the idea of the article. All authors read and approved the final manuscript.

Author details
[1]School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China. [2]Technical College for the Deaf, Tianjin University of Technology, Tianjin, China. [3]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [4]Singapore Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore.

Received: 31 March 2023 Revised: 17 October 2023
Accepted: 17 October 2023 Published online: 06 November 2023

References

1. Lv, H. (2019). *Chinese sign language linguistics*. Beijing: Intellectual Property Publishing House.
2. Zhang, X., Dai, S., & Pan, L. (2011). Chinese sign language and American sign language similarity. In *Proceedings of the 3rd international conference on information technology and scientific management* (pp. 1705–1707). Irvine: Scientific Research Publishing.
3. Zhou, H., Zhou, W., Zhou, Y., & Li, H. (2020). Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the 34th AAAI conference on artificial intelligence* (pp. 13009–13016). Palo Alto: AAAI Press.
4. Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-based sign language recognition without temporal segmentation. In *Proceedings of the 32nd AAAI conference on artificial intelligence* (pp. 2257–2264). Palo Alto: AAAI Press.
5. Guo, D., Tang, S., & Wang, M. (2019). Connectionist temporal modeling of video and language: a joint model for translation and sign labeling. In S. Kraus (Ed.), *Proceedings of the 28th international joint conference on artificial intelligence* (pp. 751–757). Cham: Springer.
6. Zhou, H., Zhou, W., & Li, H. (2019). Dynamic pseudo label decoding for continuous sign language recognition. In *Proceedings of the IEEE international conference on multimedia and expo* (pp. 1282–1287). Piscataway: IEEE.
7. Wei, C., Zhao, J., Zhou, W., & Li, H. (2020). Semantic boundary detection with reinforcement learning for continuous sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(3), 1138–1149.
8. Qi, H., Wang, Z., Guo, Q., Chen, J., Juefei-Xu, F., Ma, L., et al. (2021). Archrepair: block-level architecture-oriented repairing for deep neural networks. arXiv preprint. arXiv:2111.13330.
9. Boyd, A., Czajka, A., & Bowyer, K. (2019). Deep learning-based feature extraction in iris recognition: use existing models, fine-tune or train from scratch?. In *Proceedings of the 10th IEEE international conference on biometrics theory, applications and systems* (pp. 1–9). Piscataway: IEEE.
10. Rui, H., Ruofei, W., Qing, G., Wei, J., Yuxiang, Z., Wei, F., et al. (2023). Background-mixed augmentation for weakly supervised change detection. In *Proceedings of the 37th AAAI conference on artificial intelligence* (pp. 7919–7927). Palo Alto: AAAI Press.
11. Bing, Y., Hua, Q., Qing, G., Felix, J.-X., Xiaofei, X., Lei, M., et al. (2022). Deeprepair: style-guided repairing for deep neural networks in the real-world operational environment. *IEEE Transactions on Reliability*, *71*(4), 1401–1416.
12. Hu, H., Pu, J., Zhou, W., & Li, H. (2022). Collaborative multilingual continuous sign language recognition: A unified framework. *IEEE Transactions on Multimedia*. Advance online publication. https://doi.org/10.1109/TMM.2022.3223260.
13. Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X.-L., & Huang, H. (2020). Cross-lingual natural language generation via pre-training. In *Proceedings of the 34th AAAI conference on artificial intelligence* (pp. 7570–7577). Palo Alto: AAAI Press.
14. Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, et al. (Eds.), *Proceedings of the 33th international conference on neural information processing system* (pp. 7059–7069). Red Hook: Curran Associates.
15. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., et al. (2021). Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12299–12310). Piscataway: IEEE.
16. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). Piscataway: IEEE.
17. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805.
18. Camgoz, N.C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10023–10033). Piscataway: IEEE.
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, et al. (Eds.), *Proceedings of the 31st international conference on neural information processing systems* (pp. 5998–6008). Red Hook: Curran Associates.
20. Raganato, A., & Tiedemann, J. (2018). An analysis of encoder representations in transformer-based machine translation. In T. Linzen, G. Chrupala, & A. Alishahi (Eds.), *Proceedings of the workshop: analyzing and interpreting neural networks for NLP* (pp. 287–297). Stroudsburg: ACL.
21. Yin, F., Chai, X., & Chen, X. (2016). Iterative reference driven metric learning for signer independent isolated sign language recognition. In B. Leibe, J. Matas, N. Sebe, et al. (Eds.), *Proceedings of 14th European conference on computer vision* (pp. 434–450). Cham: Springer.
22. Wang, H., Chai, X., & Chen, X. (2019). A novel sign language recognition framework using hierarchical Grassmann covariance matrix. *IEEE Transactions on Multimedia*, *21*(11), 2806–2814.
23. Zheng, L., & Liang, B. (2016). Sign language recognition using depth images. In *Proceedings of the 14th international conference on control, automation, robotics and vision* (pp. 1–6). Piscataway: IEEE.
24. Pu, J., Zhou, W., & Li, H. (2016). Sign language recognition with multi-modal features. In E. Chen, Y. Gong, & Y. Tie (Eds.), *Proceedings of the 17th Pacific-Rim conference on multimedia* (pp. 252–261). Cham: Springer.
25. Hu, H., Zhou, W., Pu, J., & Li, H. (2021). Global-local enhancement network for NMF-aware sign language recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *17*(3), 1–19.
26. Li, H., Gao, L., Han, R., Wan, L., & Feng, W. (2020). Key action and joint CTC-attention based sign language recognition. In *Proceedings of the IEEE*

international conference on acoustics, speech and signal processing (pp. 2348–2352). Piscataway: IEEE.
27. Wang, H., Chai, X., Hong, X., Zhao, G., & Chen, X. (2016). Isolated sign language recognition with Grassmann covariance matrices. *ACM Transactions on Accessible Computing*, *8*(4), 1–21.
28. Niu, Z., & Mak, B. (2020). Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In A. Vedaldi, H. Bischof, T. Brox, et al. (Eds.), *Proceedings of the 16th European conference on computer vision* (pp. 172–186). Cham: Springer.
29. Guo, D., Zhou, W., Li, H., & Wang, M. (2018). Hierarchical LSTM for sign language translation. In S. A. McIlraith & K. Q. Weinberger (Eds.), *Proceedings of the 32nd AAAI conference on artificial intelligence* (pp. 6845–6852). Palo Alto: AAAI Press.
30. Zhang, J., Zhou, W., Xie, C., Pu, J., & Li, H. (2016). Chinese sign language recognition with adaptive HMM. In *Proceedings of the IEEE international conference on multimedia and expo* (pp. 1–6). Piscataway: IEEE.
31. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint. arXiv:1406.1078.
32. Cheng, K.L., Yang, Z., Chen, Q., & Tai, Y.-W. (2020). Fully convolutional networks for continuous sign language recognition. In A. Vedaldi, H. Bischof, T. Brox, et al. (Eds.), *Proceedings of 16th European conference on computer vision* (pp. 697–714). Cham: Springer.
33. Liao, Y., Xiong, P., Min, W., Min, W., & Lu, J. (2019). Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks. *IEEE Access*, *7*, 38044–38054.
34. Zhao, K., Zhang, K., Zhai, Y., Wang, D., & Su, J. (2021). Real-time sign language recognition based on video stream. *International Journal of Systems Control & Communications*, *12*(2), 158–174.
35. Yang, Z., Shi, Z., Shen, X., & Tai, Y.-W. (2019). SF-Net: structured feature network for continuous sign language recognition. arXiv preprint. arXiv:1908.01341.
36. Pu, J., Zhou, W., & Li, H. (2019). Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4165–4174). Piscataway: IEEE.
37. Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.
38. Graves, A., Fernández, S., Gomez, F. J., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In W. W. Cohen & A. W. Moore (Eds.), *Proceedings of the 23th international conference on machine learning* (pp. 369–376). New York: ACM.
39. Pu, J., Zhou, W., Hu, H., & Li, H. (2020). Boosting continuous sign language recognition via cross modality augmentation. In C. W. Chen, R. Cucchiara, X.-S. Hua, et al. (Eds.), *Proceedings of the 28th ACM international conference on multimedia* (pp. 1497–1505). New York: ACM.
40. Guo, L., Xue, W., Guo, Q., Liu, B., Zhang, K., Yuan, T., et al. (2023). Distilling cross-temporal contexts for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10771–10780). Piscataway: IEEE.
41. Guo, J., Xue, W., Guo, L., Yuan, T., & Chen, S. (2022). Multi-level temporal relation graph for continuous sign language recognition. In S. Yu, Z. Zhang, P. C. Yuen, et al. (Eds.), *Proceedings of the 5th Chinese conference on pattern recognition and computer vision* (pp. 408–419). Cham: Springer.
42. Liu, J., Du, B., Xue, W., & Yuan, T. (2022). Expanding intra-class difference and boosting frame-level classification for continuous sign language recognition. In *Proceedings of the international conference on high performance big data and intelligent systems* (pp. 85–89). Piscataway: IEEE.
43. Camgoz, N.C., Hadfield, S., Koller, O., & Subunets, R. B. (2017). End-to-end hand shape and continuous sign language recognition. In *Proceedings of IEEE international conference on computer vision* (pp. 3056–3065). Piscataway: IEEE.
44. Forster, J., Schmidt, C., Koller, O., Bellgardt, M., & Ney, H. (2014). Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather. In N. Calzolari, T K. Choukri, T. Declerck, et al. (Eds.), *Proceedings of the 9th international conference on language resources and evaluation* (pp. 1911–1916). Paris: ELRA.
45. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7784–7793). Piscataway: IEEE.
46. Dreuw, P., Forster, J., Deselaers, T., & Ney, H. (2008). Efficient approximations to model-based joint tracking and recognition of continuous sign language. In *Proceedings of the 8th IEEE international conference on automatic face & gesture recognition* (pp. 1–6). Piscataway: IEEE.
47. Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G.T., Zacharopoulou, V., et al. (2020). A comprehensive study on sign language recognition methods. arXiv preprint. arXiv:2007.12530.
48. Guo, D., Wang, S., Tian, Q., & Wang, M. (2019). Dense temporal convolution network for sign language translation. In S. Kraus (Ed.), *Proceedings of the 28th international joint conference on artificial intelligence* (pp. 744–750). Cham: Springer.
49. Zhao, J., Qi, W., Zhou, W., Duan, N., Zhou, M., & Li, H. (2022). Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Transactions on Multimedia*, *24*, 2662–2672.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com