Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition

Junfu Pu, Wengang Zhou, Houqiang Li
CAS Key Laboratory of Technology in Geo-spatial
Information Processing and Application System,
EEIS Department, University of Science and Technology of China
pjh@mail.ustc.edu.cn, zhwg@ustc.edu.cn, lihq@ustc.edu.cn

July 2018

Outline

- Background
- Contribution
- Proposed Architecture
- Iterative Optimization
- Experimental Results
- Conclusions

Outline

- Background
- Contribution
- Proposed Architecture
- Iterative Optimization
- Experimental Results
- Conclusions

Background

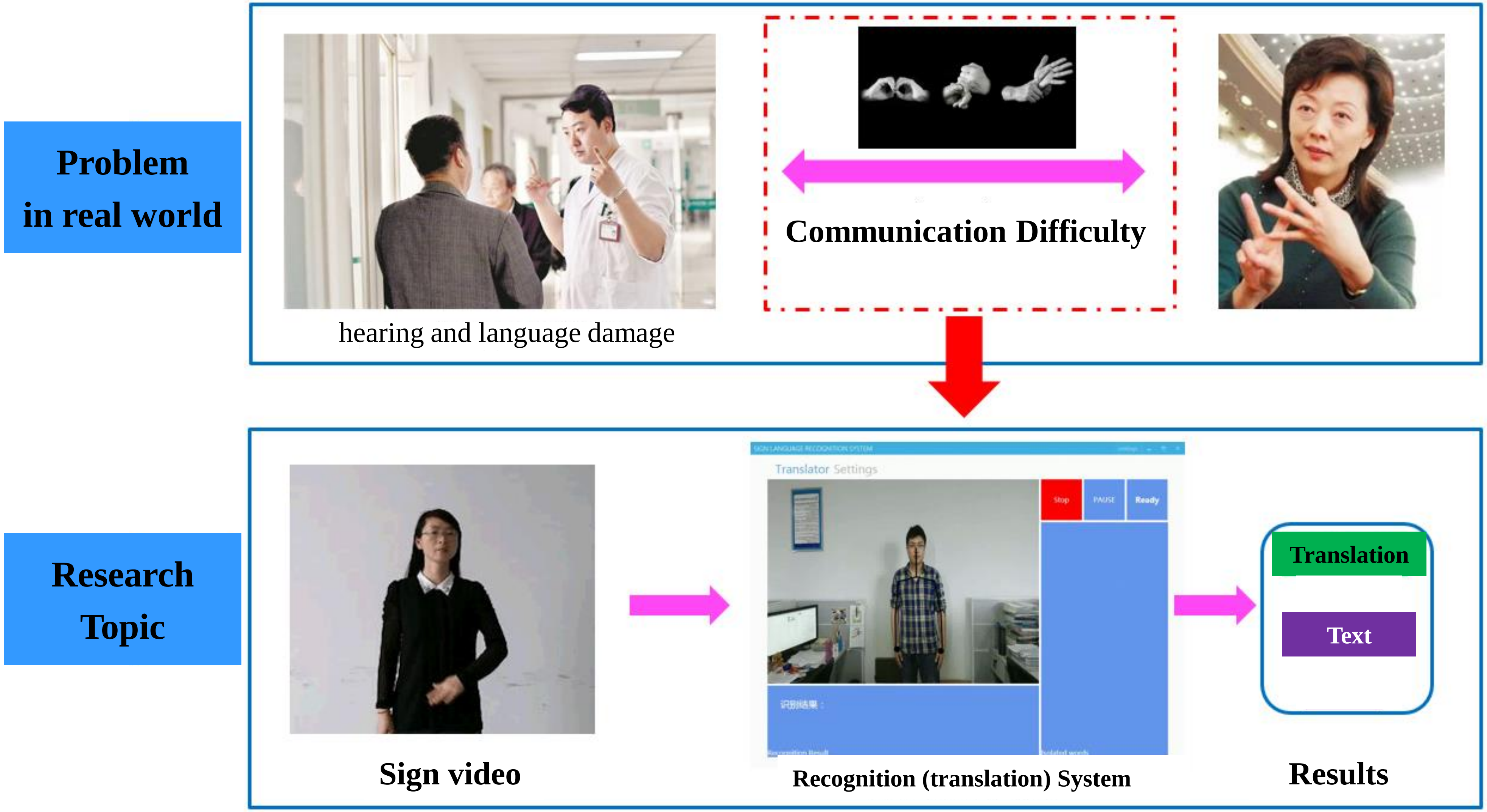
What is Sign Language?

- Communicating language used primarily by deaf people
- Use different medium such as hands, face, etc. for communication purpose

Why Sign Language?

- > 20 million people with hearing damage
- Algorithm applied for human-machine interaction
- Social impact: AI techniques improve the life quality for people with disabilities

Background


Problem
in real world
Communication Difficulty
hearing and language damage
Research
Topic
SIGN LANGUAGE RECOGNITION SYSTEM
Translator Settings
Stop
PAUSE
Ready
识别结果：
Translation
Text
Sign video
Recognition (translation) System
Results

Background

☐ Problem Formulation

Input

➢ Continuous SLR

$$\boldsymbol{s} = \{s_i\}_{t=1}^{T}$$
$$s_i \in \mathcal{V} | i = 1,2,\dots,K\}$$

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s} \in \boldsymbol{s}^*} p(\boldsymbol{s}|\boldsymbol{V})$$

MOEGLICH HEUTE NACHT FROST GLATT VORSICHT FLUSS MOEGLICH PLUS ACHT

➢ Isolated SLR

$$\hat{c} = \arg\max_{i} p(c_i|\boldsymbol{V})$$
$$i = 1,2,\dots,K$$

Democracy

Output

Outline

- Background
- Contribution
- Proposed Architecture
- Iterative Optimization
- Experimental Results
- Conclusions

Contribution

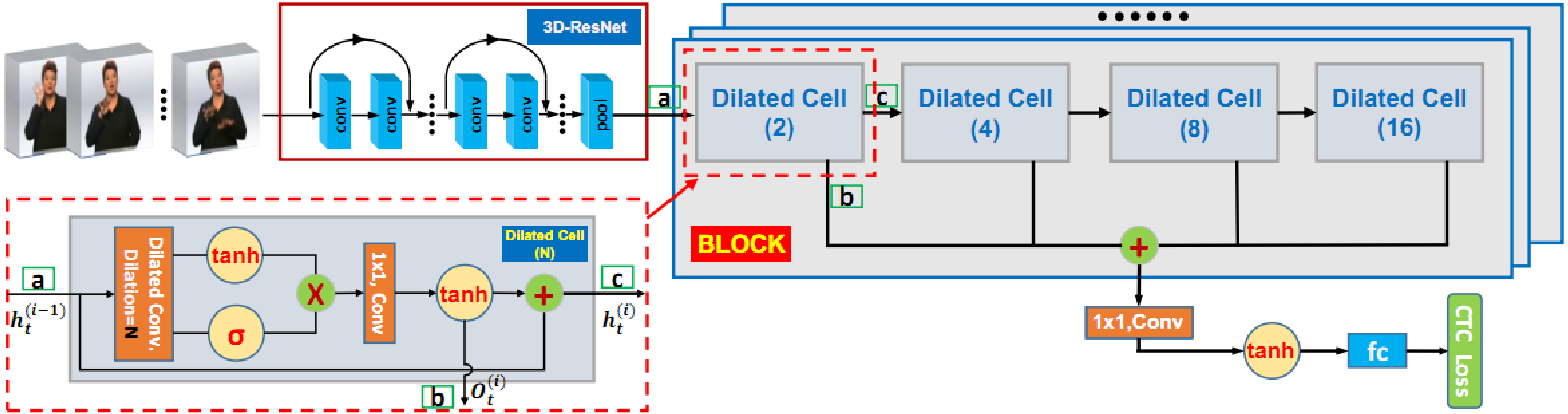
- Develop a new framework based on 3D residual network and dilated convolutions for continuous sign language recognition
- Propose an iterative optimization strategy with Connectionist Temporal Classification (CTC) for our sign language recognition system
- Outperform the state-of-the-art methods on RWTH-PHOENIX-Weather dataset

Outline

- Background
- Contribution
- Proposed Architecture
- Iterative Optimization
- Experimental Results
- Conclusions

Proposed Architecture

☐ Overall Framework



- **Visual Feature Extractor: 3D-ResNet**

$$\mathbf{X} = \{x_t\}_{t=1}^{T} \Rightarrow \mathbf{V}^N = \{v_t\}_{t=1}^{N} \Rightarrow \mathbf{F}^N = \{\Phi_\Theta(v_t)\}_{t=1}^{N}$$

- **Sequence Learning Model: Dilated Conv. Net with CTC**

$$z = \tanh\left(\mathcal{C}_d\left(h_t^{(i-1)}\right)\right) \odot \sigma(\mathcal{C}_d(h_t^{(i-1)}))$$

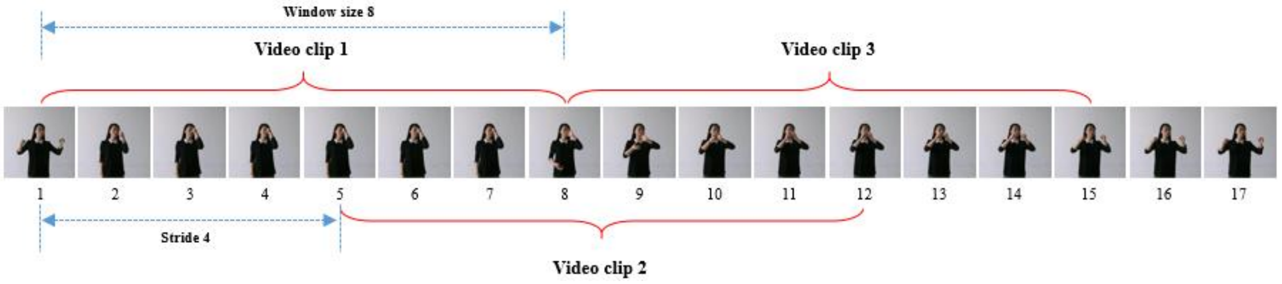
$$o_t^{(i)} = \tanh(\mathcal{C}_{1*1}(z))$$

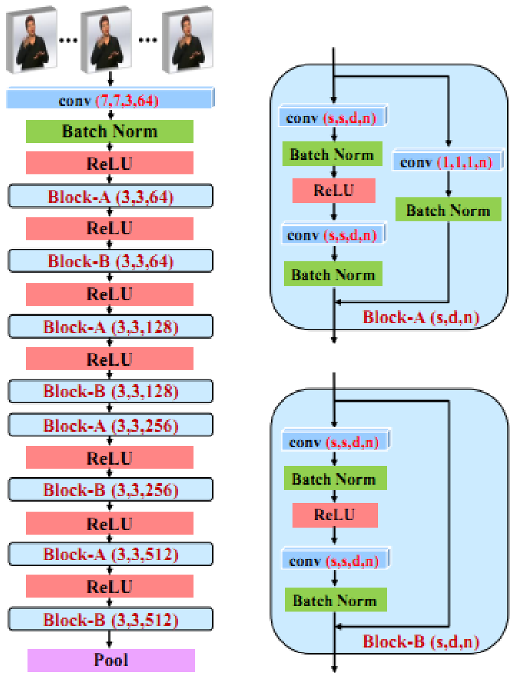
$$h_t^{(i)} = h_t^{(i-1)} + o_t^{(i)}$$

$$o_t = \sum_{all-blocks} \sum_{i} o_t^{(i)}$$

Proposed Architecture

3D ResNet



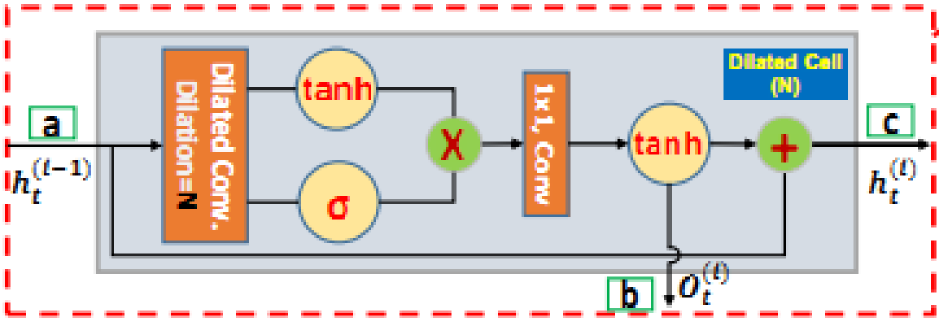


$$\mathbf{X} = \{x_t\}_{t=1}^{T}$$

$$\mathbf{V}^N = \{v_t\}_{t=1}^{N}$$

$$\mathbf{F}^N = \{\mathbf{\Phi}_\Theta(v_t)\}_{t=1}^{N}$$

Dilated Cell



$$z = \tanh\left(\mathcal{C}_d\left(h_t^{(i-1)}\right)\right) \odot \sigma(\mathcal{C}_d(h_t^{(i-1)}))$$

$$o_t^{(i)} = \tanh(\mathcal{C}_{1*1}(z))$$

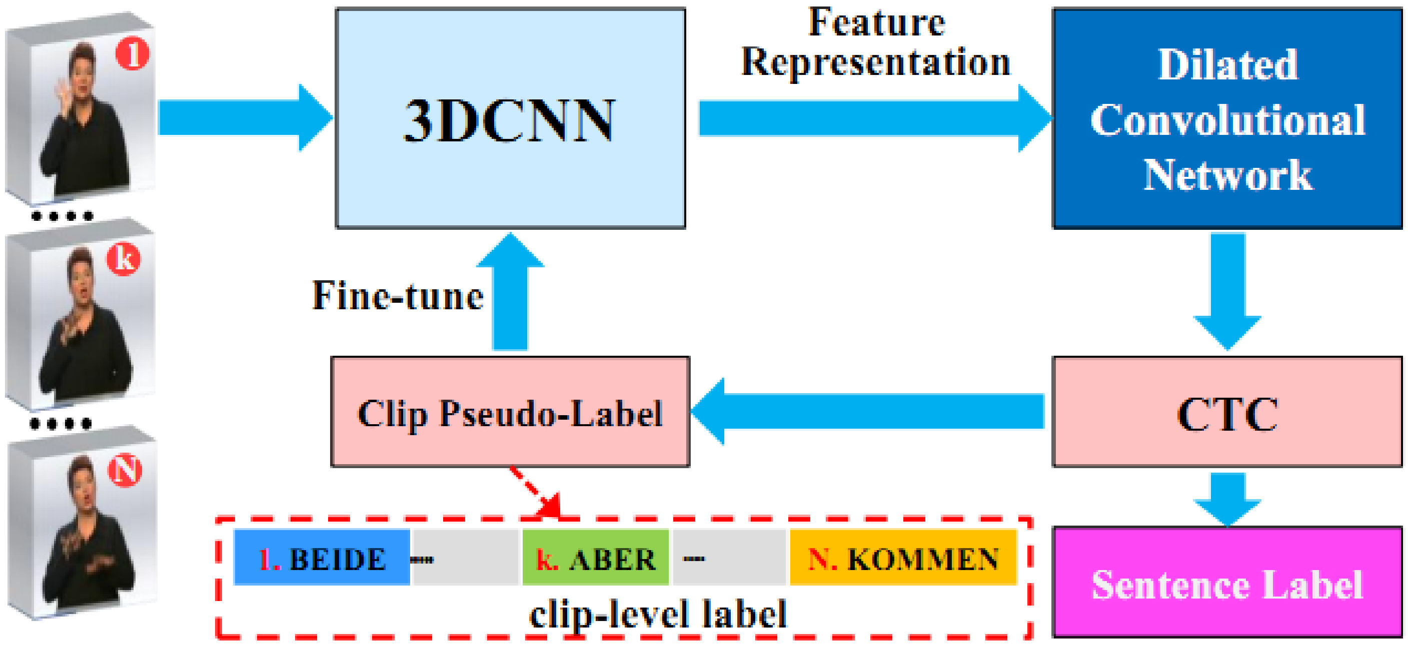
$$h_t^{(i)} = h_t^{(i-1)} + o_t^{(i)}$$

$$o_t = \sum_{all-blocks} \sum_{i} o_t^{(i)}$$

Outline

- Background
- Contribution
- Proposed Architecture
- **Iterative Optimization**
- Experimental Results
- Conclusions

Iterative Optimization



- **Step 1: Optimize dilated convolutional network with CTC loss, generate pseudo labels.**

$$\mathcal{L}_{\mathrm{CTC}} = -\ln p(\boldsymbol{s}|\mathbf{V})$$

$$\ell_i = \arg\max_j P_{i*}$$

- **Step 2: Fine-tune 3D-ResNet with category loss using pseudo labels.**
- **Step 3: Extract improved C3D features for sequence learning. Alternately run Step 1 and Step 2 until converge.**

Outline

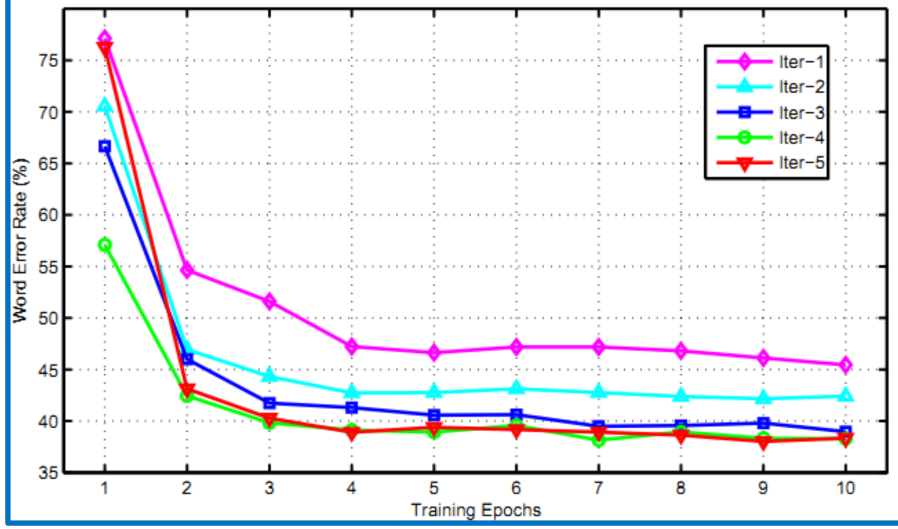
- Background
- Contribution
- Proposed Architecture
- Iterative Optimization
- **Experimental Results**
- Conclusions

Experiments

- Dataset and Evaluation
  - Continuous SLR Dataset: RWTH-PHOENIX-Weather
  - Evaluation Metric: Word Error Rate (WER)
- 3D-ResNet Setups and Initialization
  - Image crops: 224x224
  - Sliding window: length 8, step 4 (50% overlap)
  - Pre-trained on an isolated Chinese SLR dataset
  - Batch size 5, learning rate 0.001, weight decay 5×10^{-5}
  - Pooling-5b activations for clip representation
- Dilated Convolutional Network Setups
  - Dilations for each layer: 1, 2, 4, 8, 16
  - Size of blocks: 5

Experimental Results

- Iterative Results



Iterations	Dev del / ins	Dev WER	Test del / ins	Test WER
Iter-0	18.5 / 2.6	60.3	18.1 / 2.8	59.7
Iter-1	11.0 / 3.9	45.4	11.7 / 3.5	45.5
Iter-2	9.7 / 4.3	41.2	9.1 / 4.5	41.5
Iter-3	9.7 / 4.0	39.0	8.8 / 4.1	39.0
Iter-4	8.7 / 4.5	38.3	7.8 / 4.4	37.7
Iter-5	8.3 / 4.8	38.0	7.6 / 4.8	37.3

- Comparison

Methods	Dev del / ins	Dev WER	Test del / ins	Test WER
1-Mio-Hands	16.3 / 4.6	47.1	15.2 / 4.6	45.1
SubUNet	14.6 / 4.0	40.8	14.3 / 4.0	40.7
CNN-Hybrid	12.6 / 5.1	38.3	11.1 / 5.7	38.8
Staged-Opt	13.7 / 7.3	39.4	12.2 / 7.5	38.7
Ours	8.3 / 4.8	**38.0**	7.6 / 4.8	**37.3**

Experimental Results

An example for iterative optimization

(a) Pseudo-labels generated from each iteration.

(b) CTC beam search decoded results for each iterations.

Conclusions

- A novel framework with dilated convolutions for continuous sign language recognition.
- An iterative optimization strategy to train the proposed architecture by generating pseudo labels.
- Performs well both in accuracy and speed.


Thank You!