Continuous Sign Language Recognition with Iterative Spatiotemporal Fine-tuning

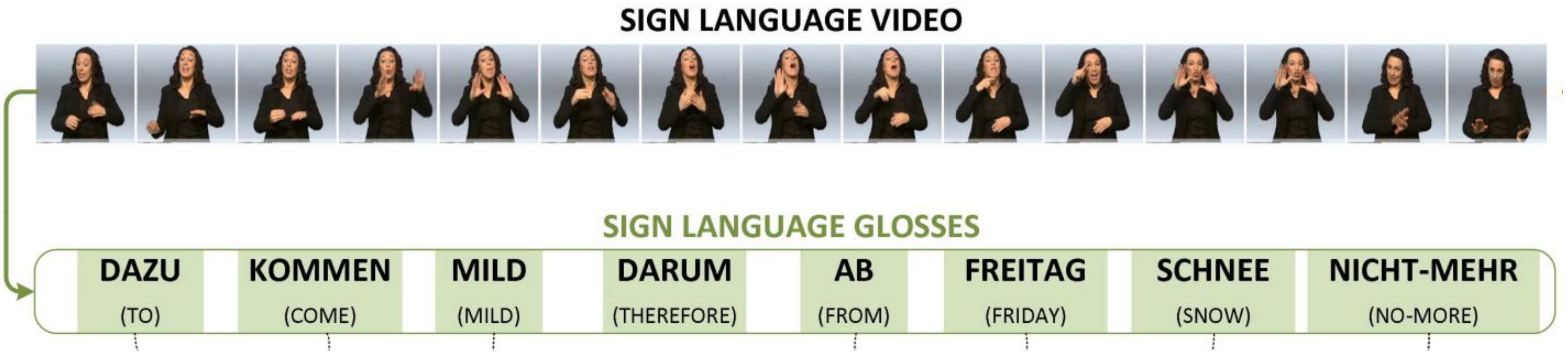
Kenessary Koishybay
kenessary.koishybay@nu.edu.kz

Medet Mukushev
mmukushev@nu.edu.kz

Anara Sandygulova
anara.sandygulova@nu.edu.kz

NAZARBAYEV
UNIVERSITY

Continuous Sign Language Recognition (CSLR)



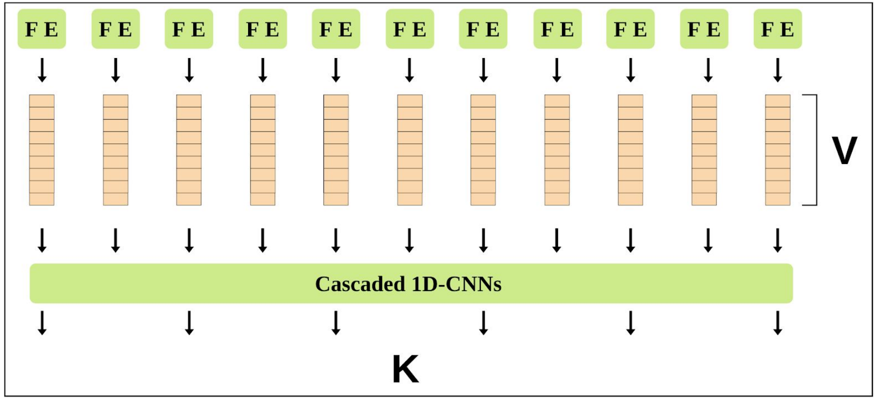
SIGN LANGUAGE VIDEO

SIGN LANGUAGE GLOSSES

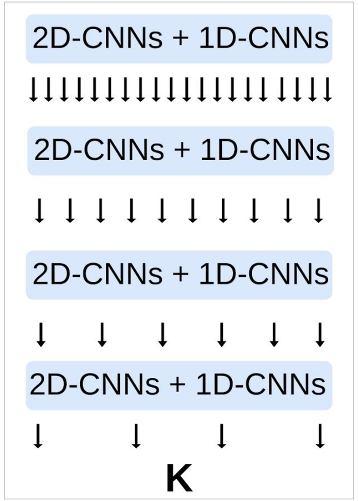
DAZU (TO) · KOMMEN (COME) · MILD (MILD) · DARUM (THEREFORE) · AB (FROM) · FREITAG (FRIDAY) · SCHNEE (SNOW) · NICHT-MEHR (NO-MORE)

Proposed Approach:
Spatio Temporal Fusion / Gloss Recognition Model

- Feature extraction:
  - GoogleNet, DenseNet (V=1024)
  - OpenPose (V=411)
- Temporal Fusion:
  - Cascaded series of 1D-CNNs

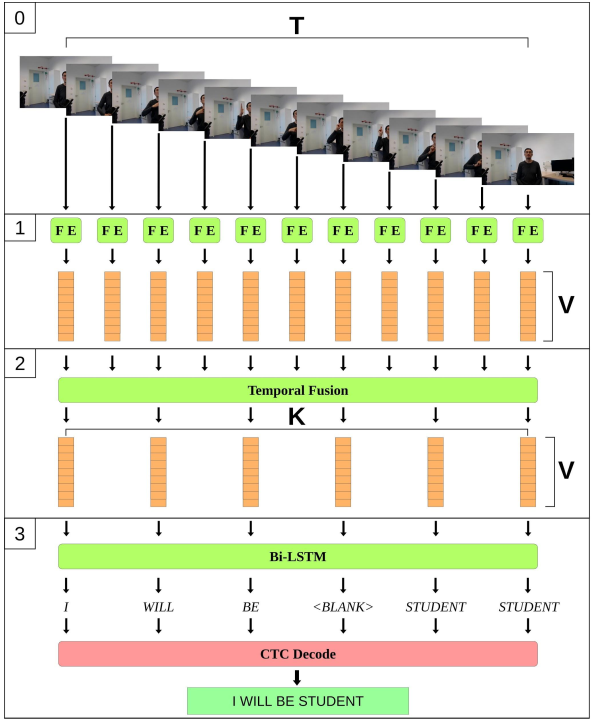


- ResNet(2+1)D
  first 4 layers (V = 1024)



NAZARBAYEV UNIVERSITY

Model Architecture
End2End Model Pipeline



Preprocessing => (T, x, y, 3) => (T, 224, 224, 3),
where T is video length

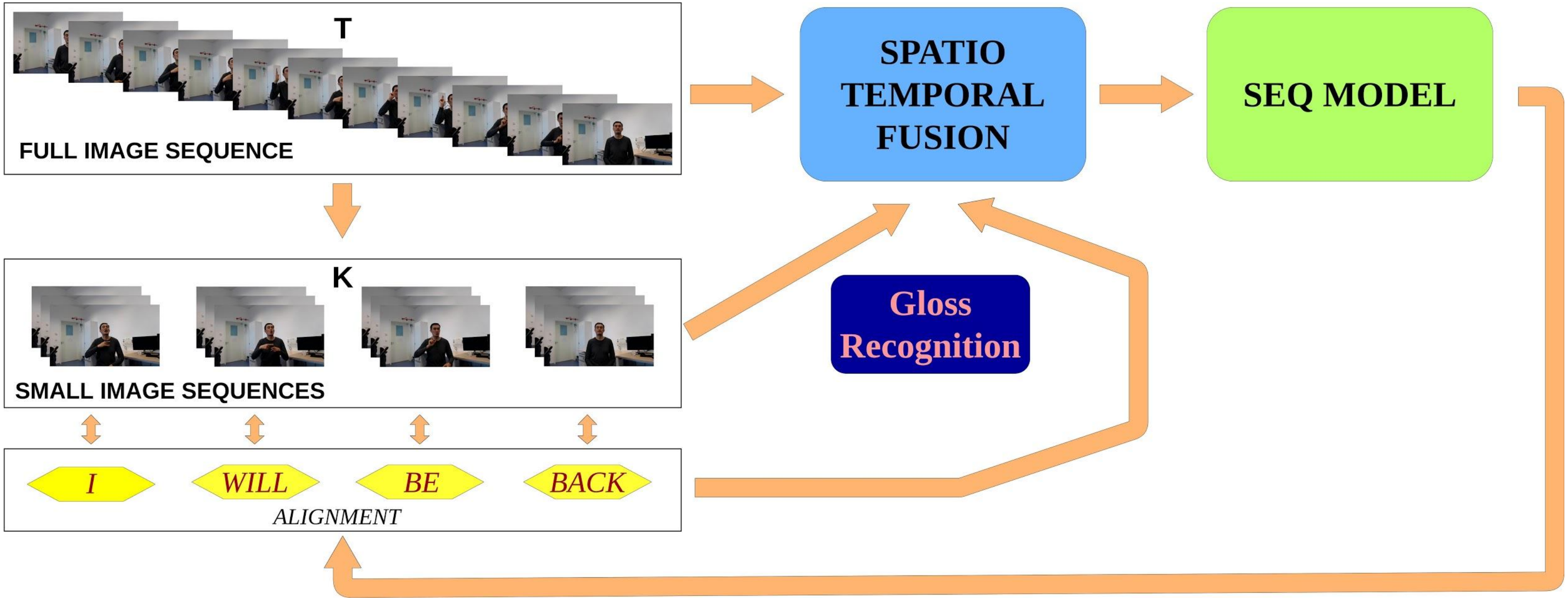
Spatial Representation: (T, 224, 224, 3) = > (T, V),
where V is the number of features

Temporal Fusion: (T, V) = > (K, V),
where K = T / 4 , 4 is temporal stride

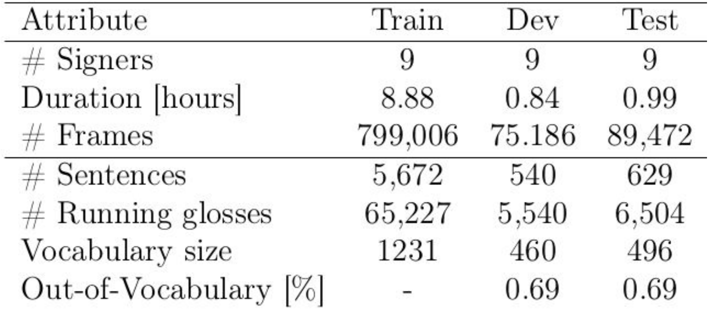
Alignment Prediction: (K, V) = > (K, M),
where M => Vocabulary Size

NAZARBAYEV
UNIVERSITY

Training process
Iterative Fine Tuning

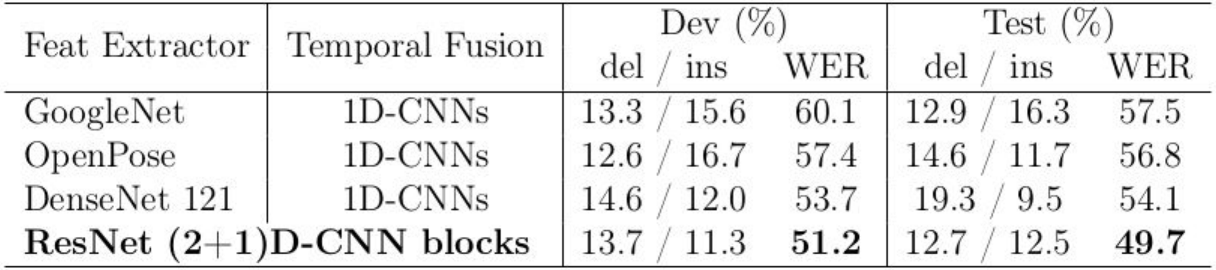


Evaluated on Phoenix2014T



Attribute	Train	Dev	Test
# Signers	9	9	9
Duration [hours]	8.88	0.84	0.99
# Frames	799,006	75.186	89,472
# Sentences	5,672	540	629
# Running glosses	65,227	5,540	6,504
Vocabulary size	1231	460	496
Out-of-Vocabulary [%]	-	0.69	0.69

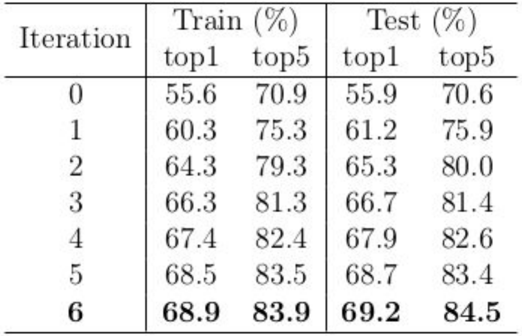
Results
Spatiotemporal Fusion selection



Feat Extractor	Temporal Fusion	Dev (%)		Test (%)	
		del / ins	WER	del / ins	WER
GoogleNet	1D-CNNs	13.3 / 15.6	60.1	12.9 / 16.3	57.5
OpenPose	1D-CNNs	12.6 / 16.7	57.4	14.6 / 11.7	56.8
DenseNet 121	1D-CNNs	14.6 / 12.0	53.7	19.3 / 9.5	54.1
ResNet (2+1)D-CNN blocks		13.7 / 11.3	**51.2**	12.7 / 12.5	**49.7**

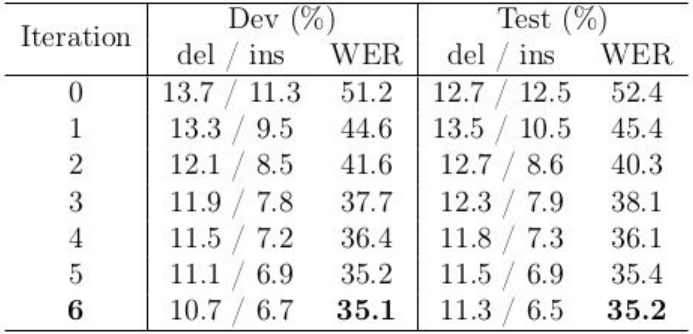
Table 1. Spatiotemporal Fusion Comparison

Results
Iterative Optimization



Iteration	Train (%)		Test (%)	
	top1	top5	top1	top5
0	55.6	70.9	55.9	70.6
1	60.3	75.3	61.2	75.9
2	64.3	79.3	65.3	80.0
3	66.3	81.3	66.7	81.4
4	67.4	82.4	67.9	82.6
5	68.5	83.5	68.7	83.4
6	**68.9**	**83.9**	**69.2**	**84.5**

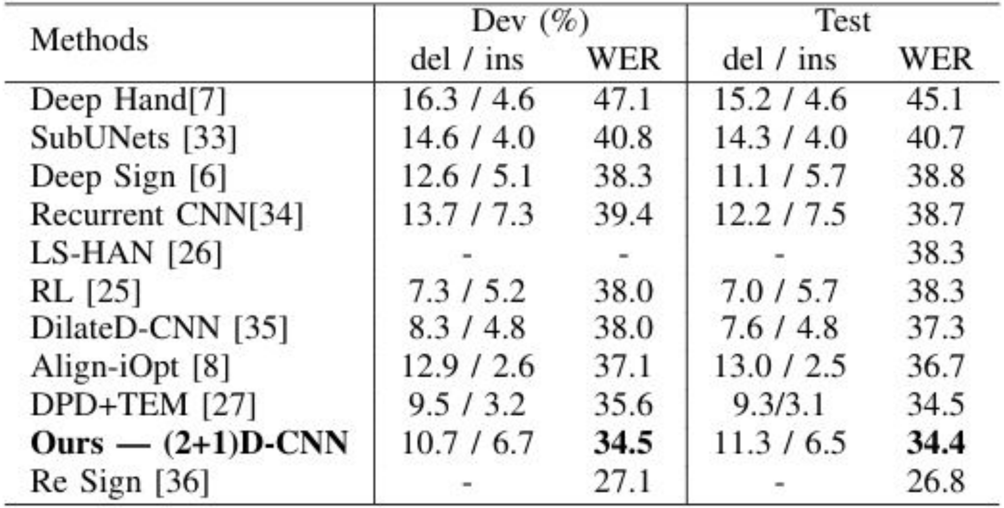
Table 2. Iterative Tuning: Gloss Recognition



Iteration	Dev (%)		Test (%)	
	del / ins	WER	del / ins	WER
0	13.7 / 11.3	51.2	12.7 / 12.5	52.4
1	13.3 / 9.5	44.6	13.5 / 10.5	45.4
2	12.1 / 8.5	41.6	12.7 / 8.6	40.3
3	11.9 / 7.8	37.7	12.3 / 7.9	38.1
4	11.5 / 7.2	36.4	11.8 / 7.3	36.1
5	11.1 / 6.9	35.2	11.5 / 6.9	35.4
6	10.7 / 6.7	**35.1**	11.3 / 6.5	**35.2**

Table 3. Iterative Tuning: End2End Model

Results
Comparison against state-of-the-art



Methods	Dev (%)		Test	
	del / ins	WER	del / ins	WER
Deep Hand[7]	16.3 / 4.6	47.1	15.2 / 4.6	45.1
SubUNets [33]	14.6 / 4.0	40.8	14.3 / 4.0	40.7
Deep Sign [6]	12.6 / 5.1	38.3	11.1 / 5.7	38.8
Recurrent CNN[34]	13.7 / 7.3	39.4	12.2 / 7.5	38.7
LS-HAN [26]	-	-	-	38.3
RL [25]	7.3 / 5.2	38.0	7.0 / 5.7	38.3
DilateD-CNN [35]	8.3 / 4.8	38.0	7.6 / 4.8	37.3
Align-iOpt [8]	12.9 / 2.6	37.1	13.0 / 2.5	36.7
DPD+TEM [27]	9.5 / 3.2	35.6	9.3/3.1	34.5
Ours — (2+1)D-CNN	10.7 / 6.7	**34.5**	11.3 / 6.5	**34.4**
Re Sign [36]	-	27.1	-	26.8

Table 4. Results comparison with state of the art models

Conclusion and Future work

- Outcomes
    - Novel Deep Neural Architecture for CSLR
    - Comparable state-of-the-art Performance
    - Open source code

- Future work
    - More exhaustive design selection
    - Iterative Optimization of Image Feature Extractor and Temporal Fusion model

Thank you

Source code:
https://github.com/I3orn2FLY/CSLR-ISTF