TIANJIN UNIVERSITY OF TECHNOLOGY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

WED-AM-242

Distilling Cross-Temporal Contexts for Continuous Sign Language Recognition

Leming Guo[1] Wanli Xue[1*] Qing Guo[2*] Bo Liu[3] Kaihua Zhang[4]
Tiantian Yuan[5] Shengyong Chen[1]

[1] School of Computer Science and Engineering, Tianjin University of Technology,
[2] Centre for Frontier AI Research (CFAR), A*STAR, Singapore, [3] Walmart Global Tech, Sunnyvale, CA, USA,
[4] School of Computer and Software, Nanjing University of Information Science and Technology,
[5] Technical College for the Deaf, Tianjin University of Technology

➢ Motivation and Contribution

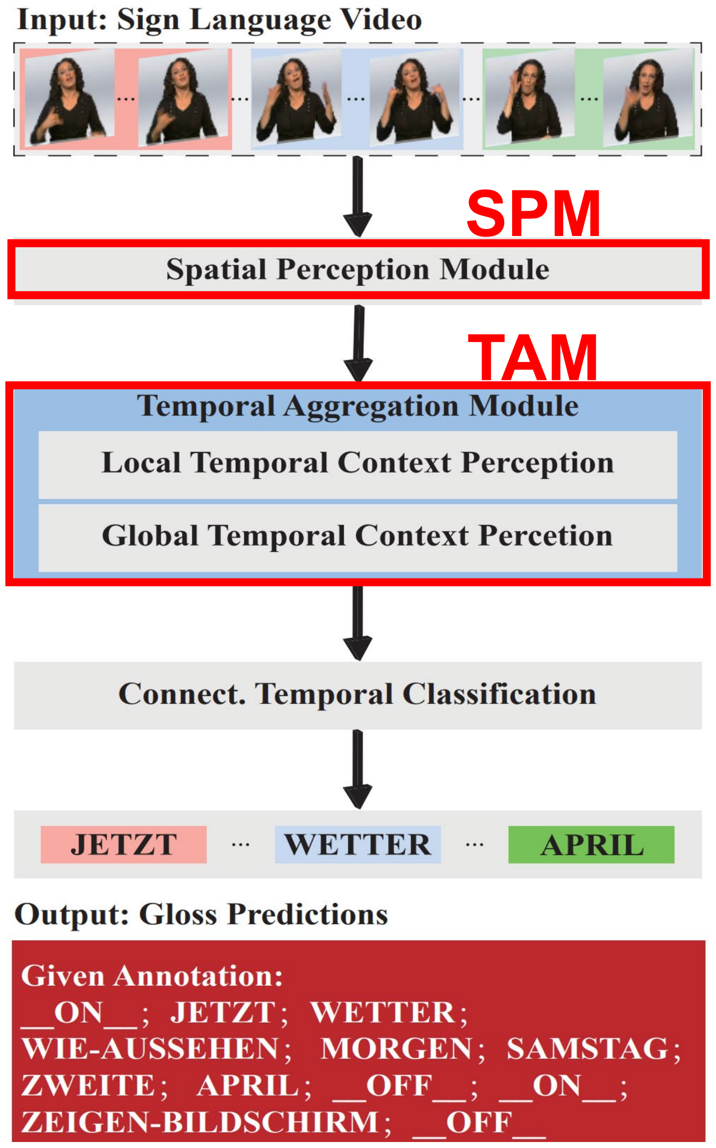
- **Motivation:**
  - The spatial perception module tends to be undertrained.
  - However, we have no idea about the desired temporal aggregation module.

- **Contribution:**
  - We extensively study the limitation and desirable properties of the temporal aggregation module and find it should be a shallow one and have high temporal aggregation capability.
  - We propose the cross-temporal context aggregation (CTCA) that a shallow temporal aggregation module has capable of incorporating local-global temporal contexts and the linguistic prior.

➢ The SOTA framework of CSLR

- Spatial Perception Module (SPM):
  - Spatial feature extraction.
- Temporal Aggregation Module (TAM):
  - Local-global temporal feature extraction, which is crucial to recognition performance.
  - It includes the local temporal perception module (1D-TCNs), and the global temporal perception module (BLSTM).
- Sequence prediction:
  - Connectionist temporal classification (CTC) function.

(a) Existing CSLR framework

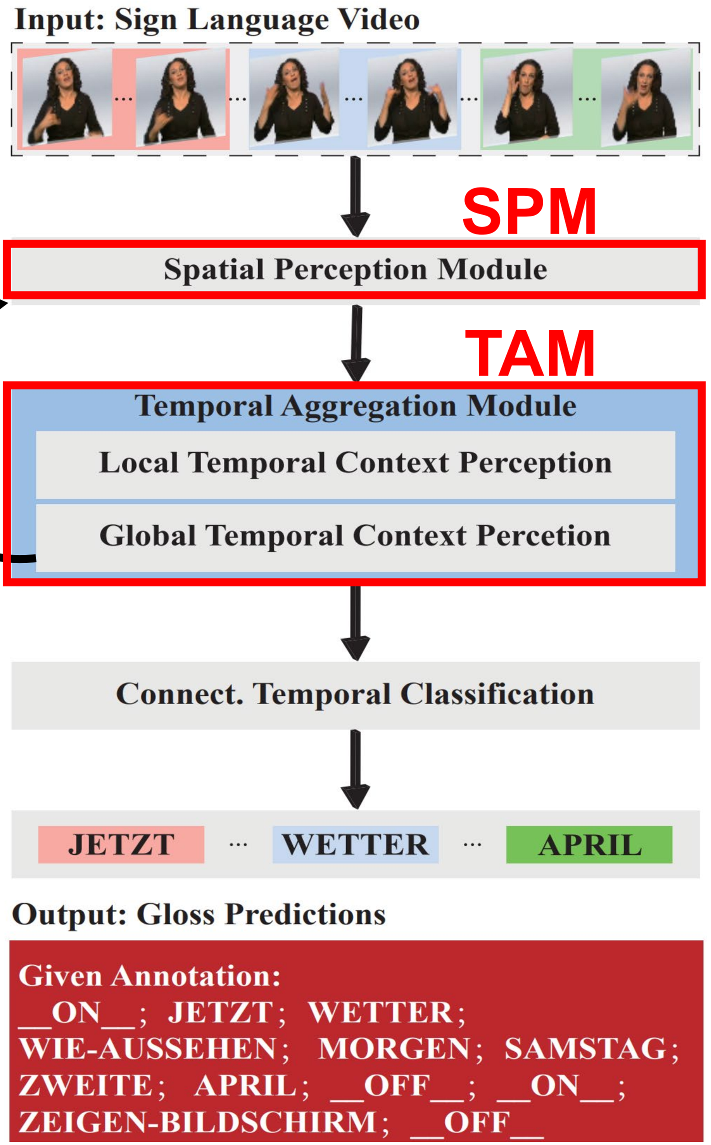


➢ Motivation

- The spatial perception module tends to be undertrained due to the easy overfitting temporal aggregation module and the objective function[1-3].

Insufficient feedback information

(a) Existing CSLR framework



- What are the effects of the TAM on the SPM?
- What are the properties of the desired TAM?

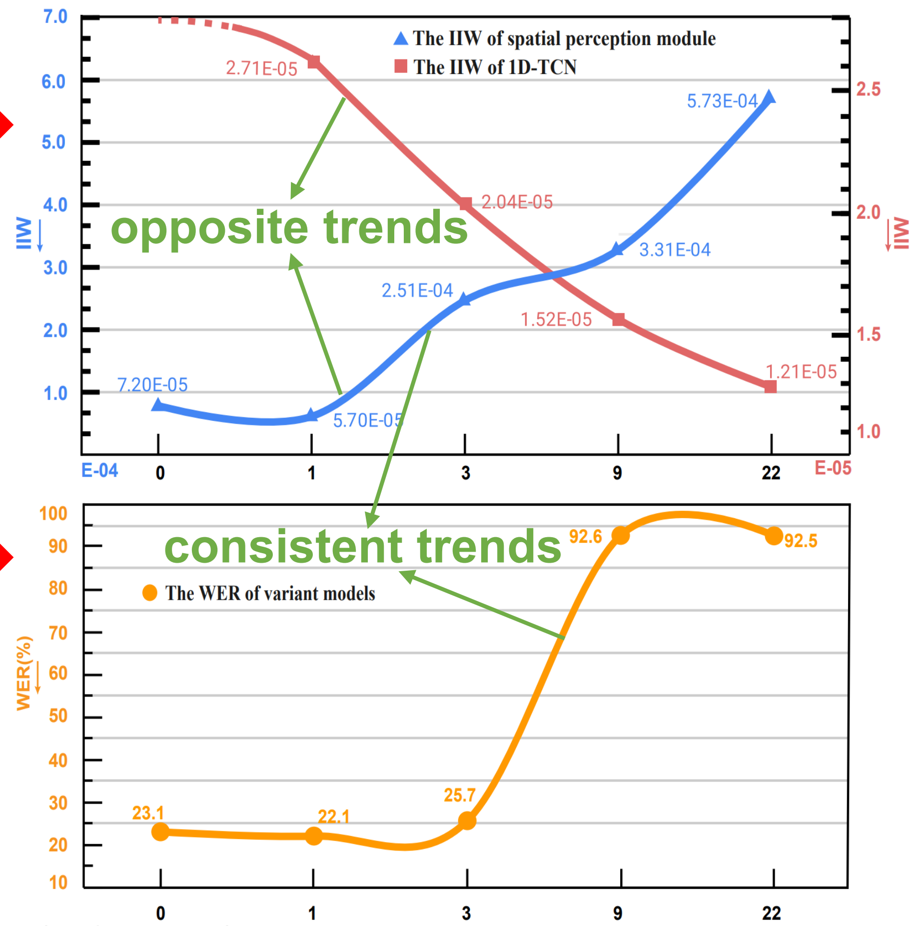
1. Ronglai Zuo and Brian Mak. C^{2}SLR: Consistency-enhanced continuous sign language recognition. In CVPR, 2022.
2. Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In ICCV, 2021.
3. Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In CVPR, 2019.

TIANJIN UNIVERSITY OF TECHNOLOGY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

➢ Empirical Studies and Analysis

- Model Generalizability Metric:
  - IIW (the compression of information stored in weights)[1].
- Observations:
  - The effects of chain depth on the capability of SPM and TAM have completely opposite trends.
  - SPM: has higher effects on the final prediction.
- TAM desired properties:
  - SPM desires a shallow TAM.
  - TAM desires a deeper architecture.



1. Zifeng Wang, Shao-Lun Huang, Ercan Engin Kuruoglu, Ji-meng Sun, Xi Chen, and Yefeng Zheng. Pac-bayes information bottleneck. In ICLR, 2022.

➢ The conflict caused by shallow TAM

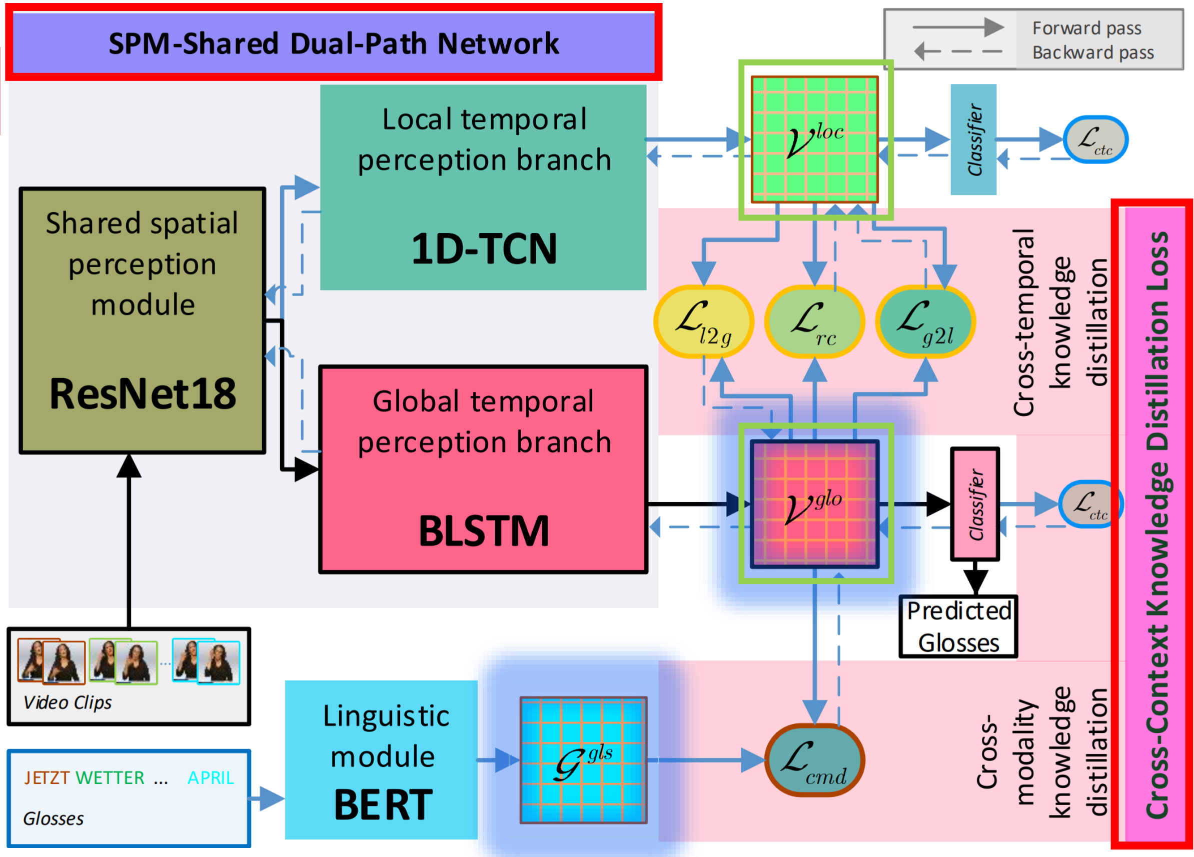
- Advantage:
  - Shallow TAM allows more thorough training of the spatial perception module.

- Disadvantage:
  - However, a shallow TAM cannot well capture both local and global temporal context information.

➤ Cross-Temporal Context Aggregation (CTCA)

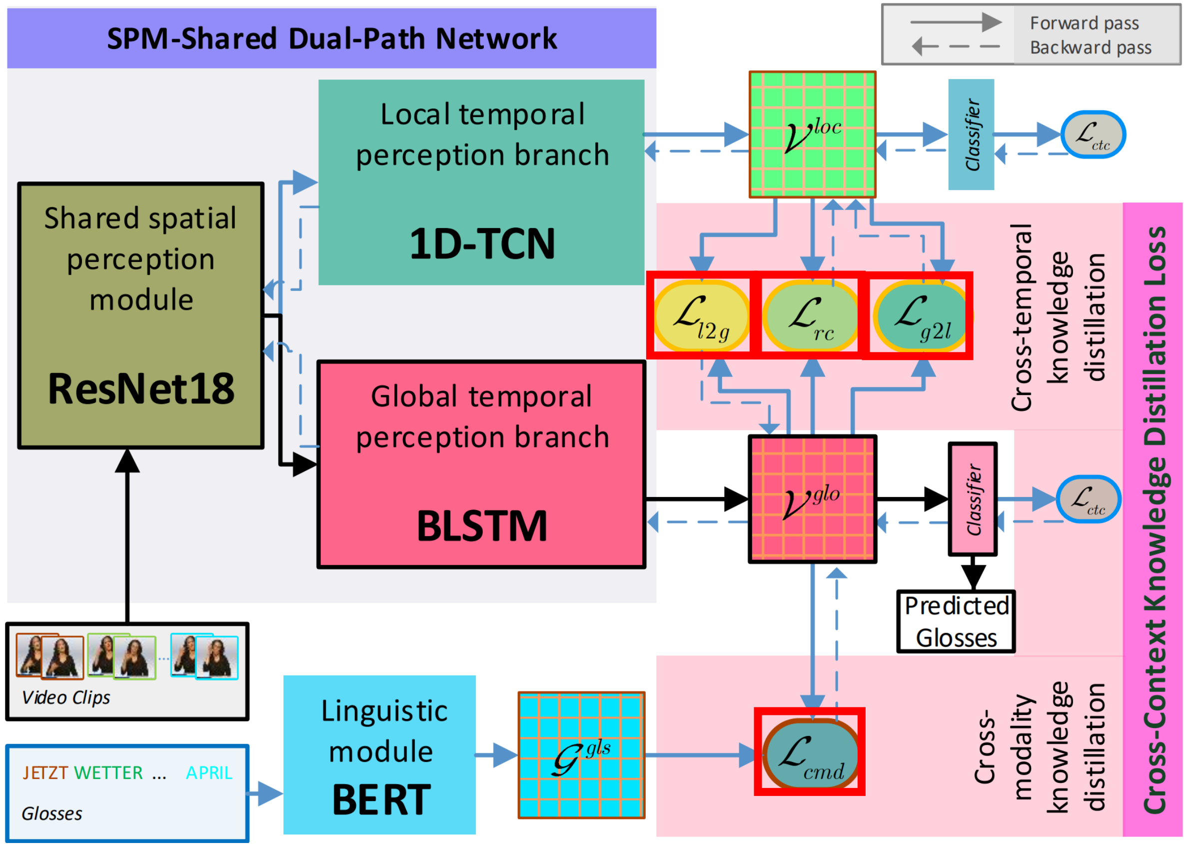
- SPM-Shared Dual-Path Network (SDPN): designed for a shallow TAM allows more thorough training of the SPM.

- Cross-Context Knowledge Distillation (CCKD): enables the global perception module to achieve local-global temporal perception and be more discriminative.



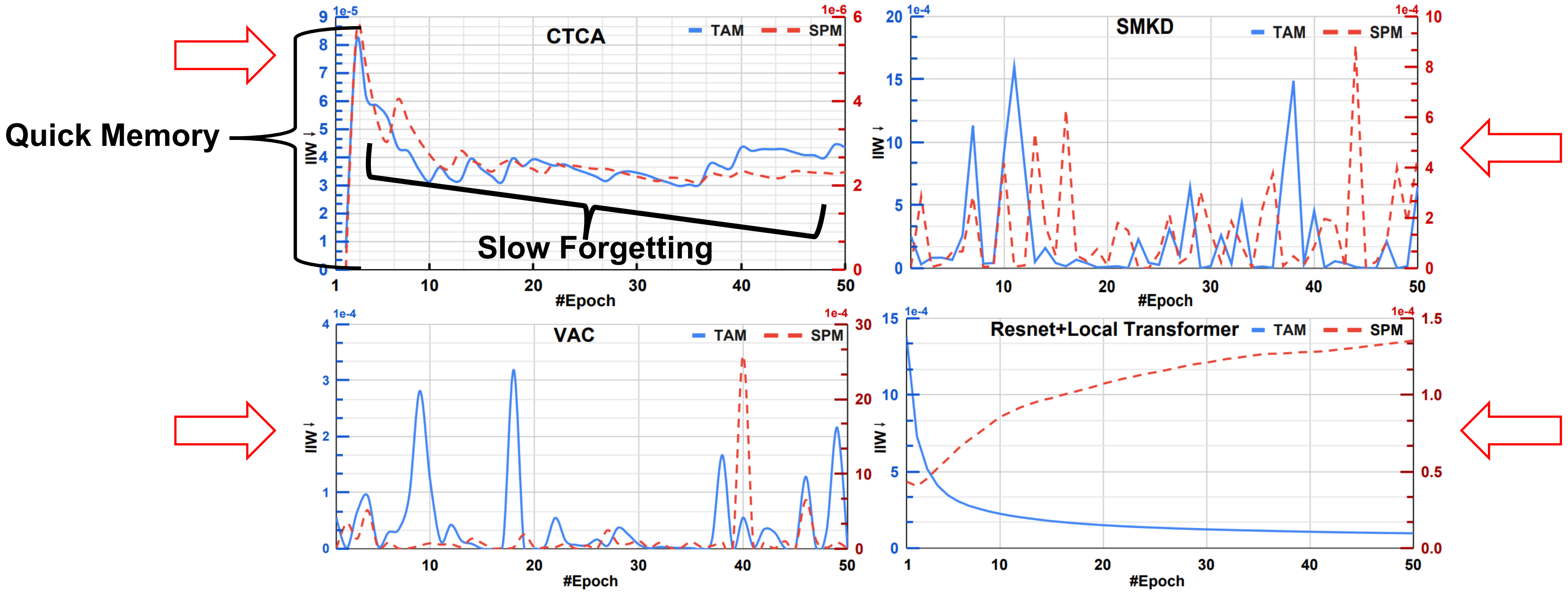
➤ Cross-Temporal Context Aggregation (CTCA)

- Cross-Context Knowledge Distillation:
- Cross-temporal knowledge distillation:
  - Local temporal context guidance loss: encourages V^{glo} to learn sign-wise context maintained in V^{loc}.
  - Global temporal context guidance loss: evolves distilling correlation among co-occurring signs to V^{loc}.
  - Reconstruction loss: reinforces the above cross-temporal context distillations.
- Cross-modality knowledge distillation: encourages V^{glo} to learn the inter-gloss discrimination indirectly.



➢ Generalizability of SPM and TAM

- "Quick Memory - Slow Forgetting" [1,2]



1. N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," 2015 IEEE Information Theory Workshop (ITW), Jerusalem, Israel, 2015
2. Zifeng Wang, Shao-Lun Huang, Ercan Engin Kuruoglu, Ji-meng Sun, Xi Chen, and Yefeng Zheng. Pac-bayes information bottleneck. In ICLR, 2022.

➢ Comparison with state-of-the-arts

Table 1. Comparison with state-of-the-art methods on the RWTH-2014 dataset. (WER (%) the lower is the better).

Methods	Dev		Test	
	del/ins	WER	del/ins	WER
DNF	7.8/3.5	23.8	7.8/3.4	24.4
FCN	-	23.7	-	23.9
VAC	7.9/2.5	21.2	8.4/2.6	22.3
CMA	7.3/2.7	21.3	7.3/2.4	21.9
SMKD	6.8/2.5	20.8	6.3/2.3	21.0
C^2SLR	-	20.5	-	20.4
TLP	6.3/2.8	19.7	6.1/2.9	20.8
RadialCTC	6.5/2.7	19.4	6.1/2.6	20.2
CTCA(Ours)	6.2/2.9	19.5	6.1/2.6	20.1

Table 2. Comparison with state-of-the-art methods on the RWTH-2014T dataset. (WER (%) the lower is the better).

Methods	WER	
	Dev	Test
SLT	24.6	24.5
CNN+LSTM+HMM	22.1	24.1
BN-TIN+Transf	22.7	23.9
V-L Mapper	21.9	22.5
SMKD	20.8	22.4
C^2SLR	20.2	20.4
TLP	19.4	21.2
CTCA(Ours)	19.3	20.3

Table 3. Comparison with state-of-the-art methods on the CSL-Daily dataset. (WER (%) the lower is the better).

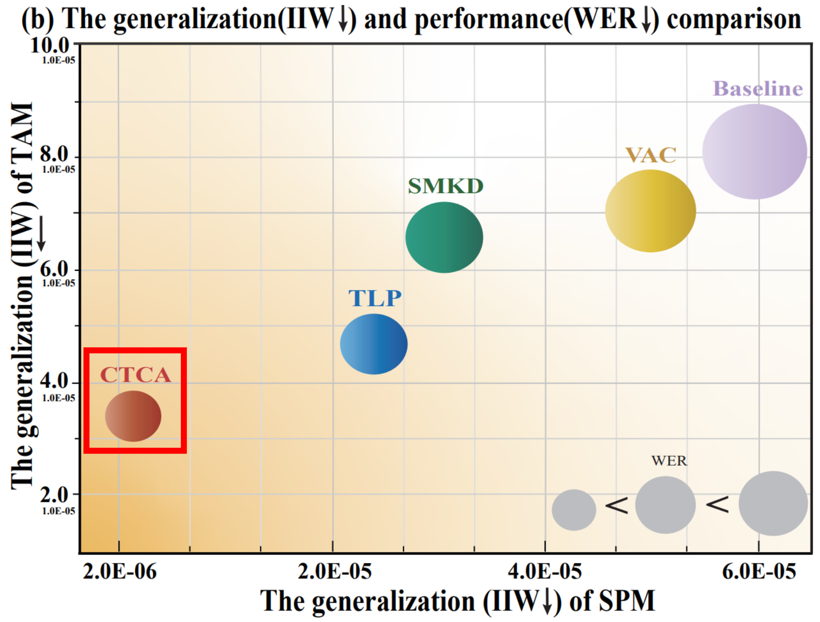
Methods	Dev		Test	
	del/ins	WER	del/ins	WER
LS-HAN	14.6/5.7	39.0	14.8/5.0	39.4
SLT(Gloss+Text)	10.3/4.4	33.1	9.6/4.1	32.0
FCN	12.8/4.0	33.2	12.6/3.7	32.5
BN-TIN+Transf	13.9/3.4	33.6	13.5/3.0	33.1
TIN+Iterative	12.8/3.3	32.8	12.5/2.7	32.4
CTCA(Ours)	9.2/2.5	31.3	8.1/2.3	29.4

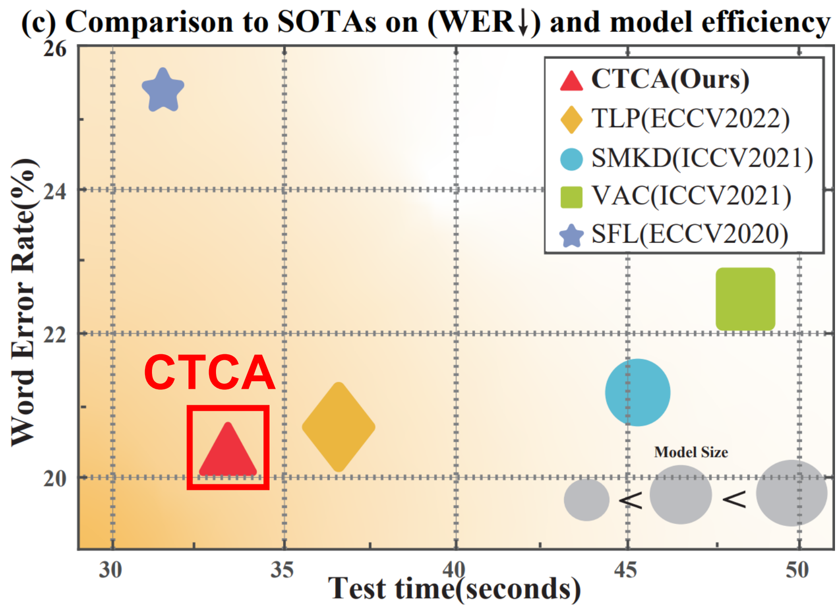
TIANJIN UNIVERSITY OF TECHNOLOGY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

➢ Comparison with state-of-the-arts

- Better Generalizability, Smaller Parameter size
- Higher Performance, Faster Inference





Ablation Study

Table 5. Ablation study on cross-context knowledge distillation loss on the RWTH-2014.

Method	$\mathcal{L}_{ctd}$			$\mathcal{L}_{cmd}$	Dev	Test
	$\mathcal{L}_{l2g}$	$\mathcal{L}_{g2l}$	$\mathcal{L}_{rc}$			
Baseline	-	-	-	-	21.8	22.1
Vanilla	-	-	-	-	21.7	21.9
SDPN A	✓	-	-	-	21.0	21.1
SDPN A- $I(.;)$	✓	-	-	-	21.3	21.5
SDPN B	-	✓	-	-	20.8	20.7
SDPN C	✓	✓	-	-	20.4	20.6
SDPN D	✓	✓	✓	-	20.0	20.4
SDPN D-$\omega(.;)$	✓	✓	✓	-	20.2	20.6
SDPN E	-	-	-	✓	21.3	21.0
CTCA	✓	✓	✓	✓	19.5	20.1

Table 6. Comparison of different knowledge fusion schemes on the RWTH-2014. "Wasserstein" is the Wasserstein distance.

Methods	Knowledge fusion	Dev	Test
SDPN	-	21.7	21.8
	Vanilla distillation	21.6	21.6
	Wasserstein	21.6	21.5
	JMMD	21.3	21.3
	CKD	21.3	21.5
	CTCA ($\mathcal{L}_{l2g}$)	21.0	21.1
	concatenation	22.7	23.6
	point-wise addition	21.2	22.3
	attention	22.2	22.6

➢ Ablation Study

Table 7. Performance comparison of local temporal perception module with distinct temporal window widths on the RWTH-2014. Ft and Ft(d) correspond to the 1D temporal convolution layer with the kernel of t and dilation of d, respectively.

Method	variants	windows	Dev	Test
1D-TCN	F3-F3-F3	7*2	19.5	20.1
	F3(1)-F3(2)	7*2	19.8	20.3
	F5-F5	9*2	20.6	20.6
	F5-F5-F5	13*2	19.9	20.6
	F7-F7	13*2	20.1	20.3

Table 8. Comparison of CTCA with distinct global temporal perception modules (GTPM) on the RWTH-2014.

Method	variants	Dev	Test
GTPM-branch	BLSTM	19.5	20.1
	Dilated blocks	22.2	22.6
	Transformer	28.7	28.9
	Transformer+BLSTM	24.4	24.1

Thanks for your listening!