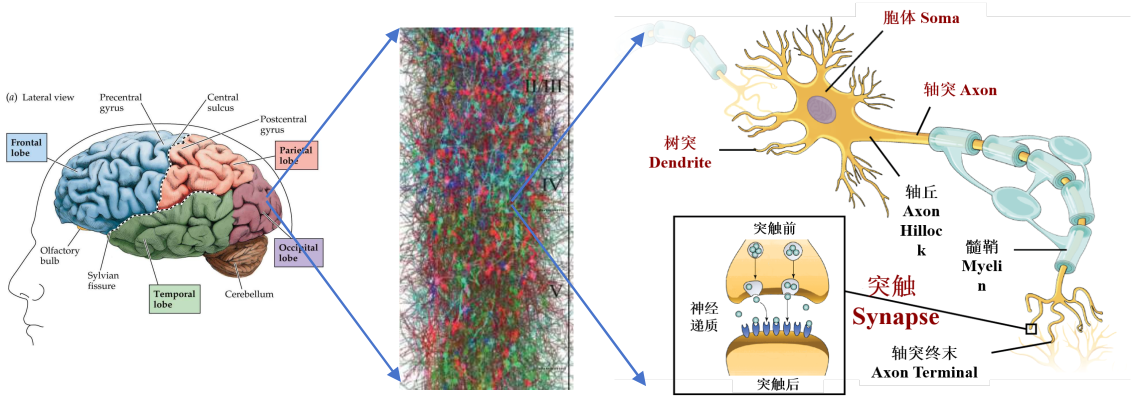
The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)

QKFormer: Hierarchical Spiking Transformer using Q-K Attention

Chenlin Zhou[1], Han Zhang[1,2], Zhaokun Zhou[1,3], Liutao Yu[1], Liwei Huang[1,3], Xiaopeng Fan[1,2], Li Yuan[1,3],
Zhengyu Ma[1], Huihui Zhou[1], Yonghong Tian[1,3]
[1]Pengcheng Laboratory, [2]Harbin Institute of Technology, [3]Peking University

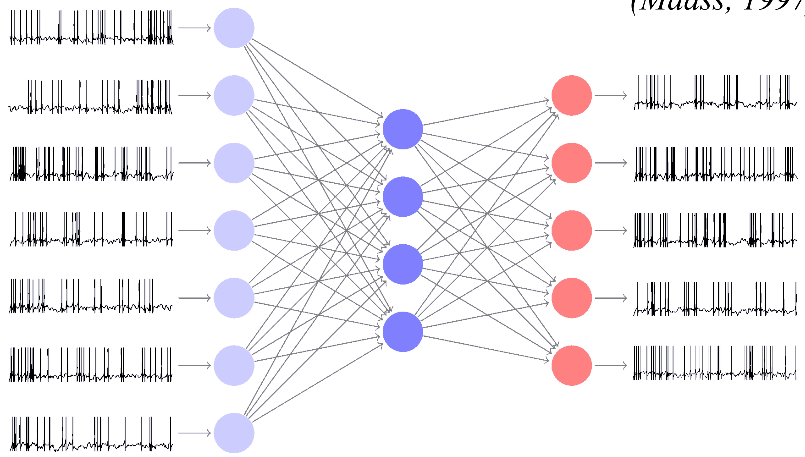
Background and Motivation

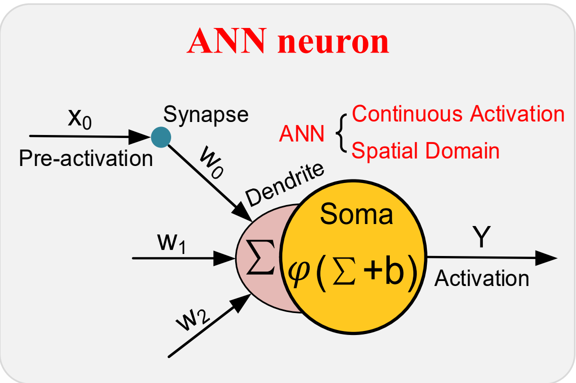


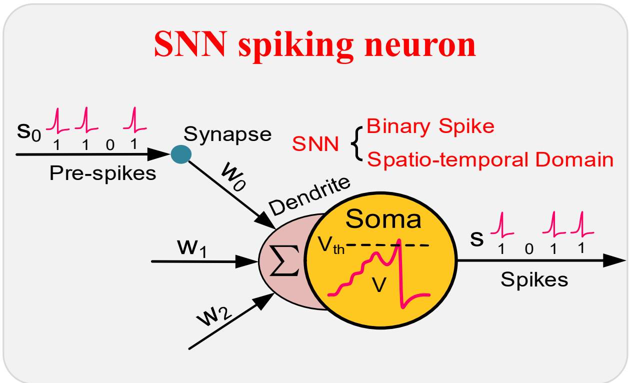
Neuron in the brain

SNNs: the third generation of neural network models

(Maass, 1997)







- ✓ **Biological plausibility,**
- ✓ **Spatiotemporal dynamics,**
- ✓ **Strong robustness,**
- ✓ **High energy-efficient**, spike communication,
- ? **Performance**.

Background and Motivation

Substantial gap in performance!!!

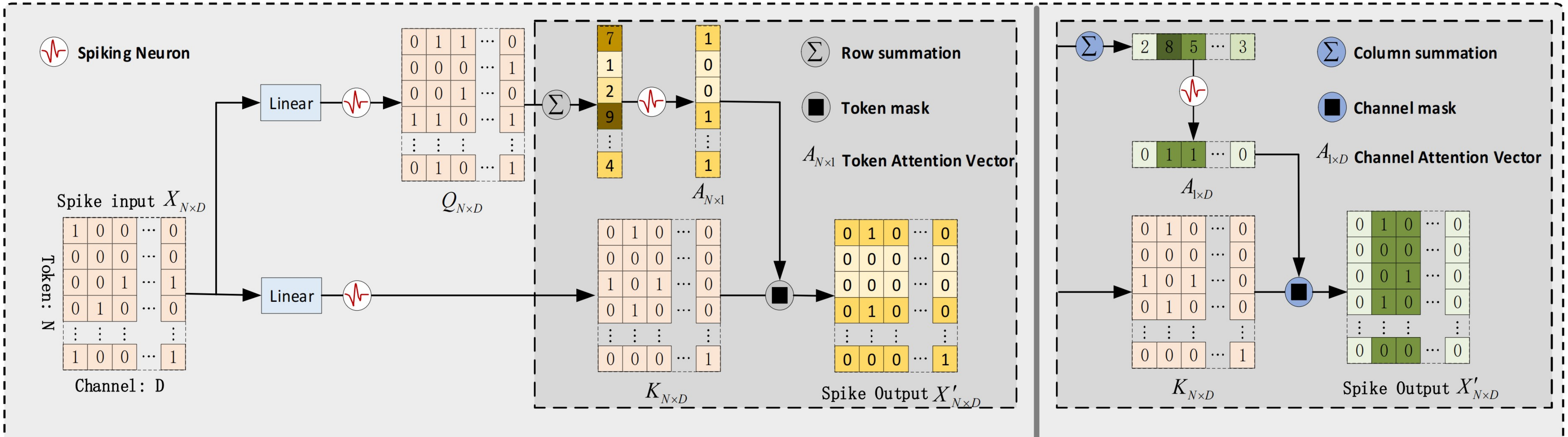
There remains a substantial gap in performance between SNNs and ANNs on large-scale datasets.

Methods	Type	Param.	ImageNet Acc
Spikformer	SNN	66.3M	74.8
Swin Transformer	ANN	87.7M	84.5
Our work	SNN	64.9M	85.6

Our Solutions:

- **Q-K Attention:** A new efficient spike-based attention module that allows the construction of larger models.
- **Multi-scale spiking transformer respresentation.**
- **Novel spiking patch embedding.**

Method: Q-K Attention



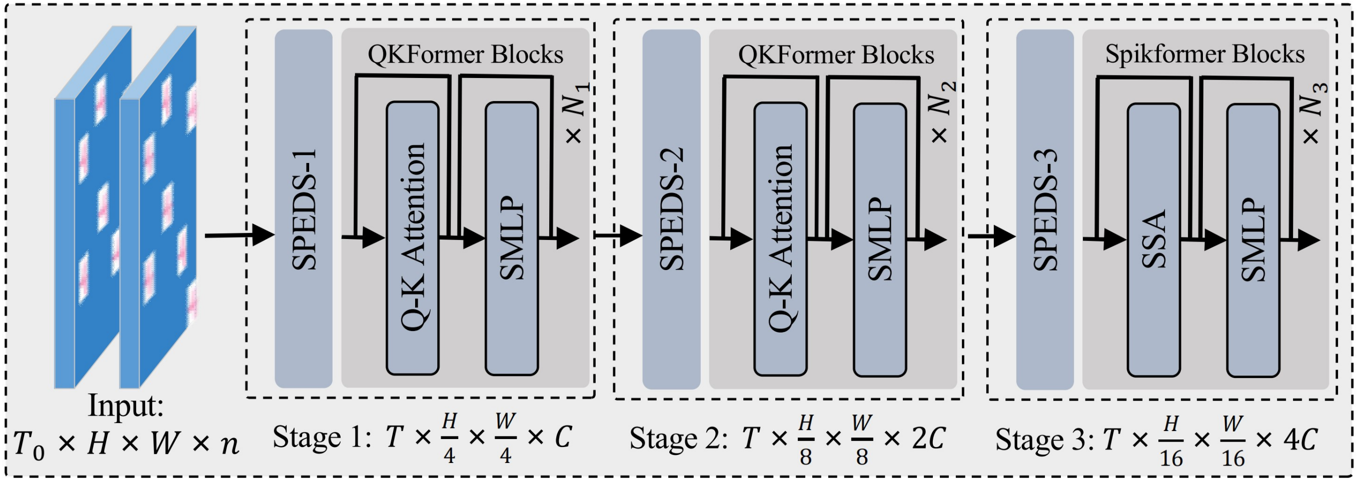
(a) Q-K Token Attention (QKTA)

(b) Q-K Channel Attention (QKCA)

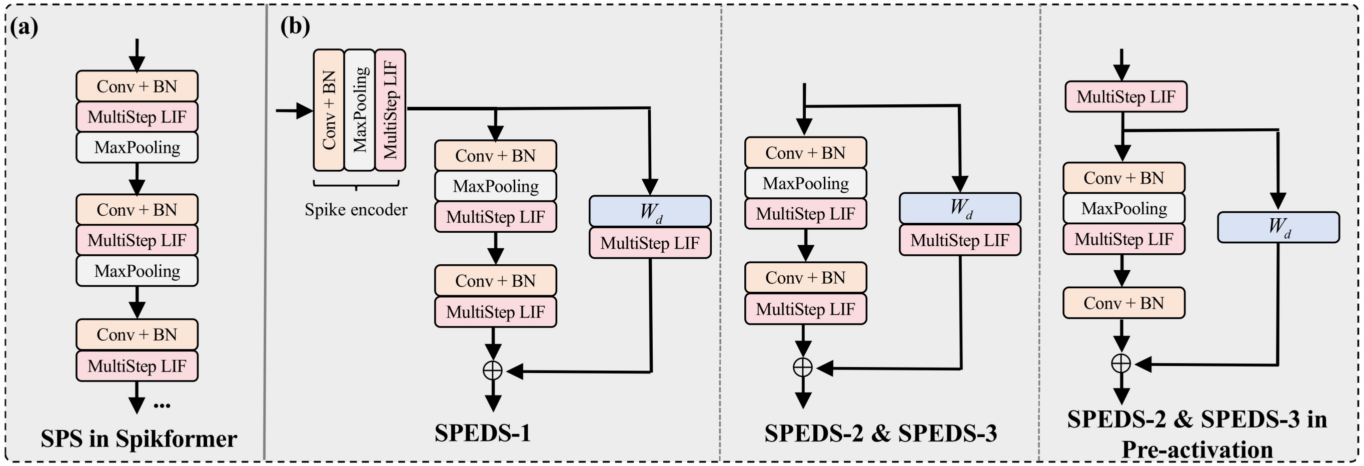
$$
\begin{cases}
\mathbf{Q} = \mathrm{SN}_{\mathrm{Q}}(\mathrm{BN}(\mathbf{X}\mathbf{W}_{\mathrm{Q}})), \ \ \mathbf{K} = \mathrm{SN}_{\mathrm{K}}(\mathrm{BN}(\mathbf{X}\mathbf{W}_{\mathrm{K}})), \\
\mathbf{A}_t = \mathrm{SN}(\sum_{i=0}^{D} \mathbf{Q}_{i,j}), \mathbf{X}' = \mathbf{A}_t \otimes \mathbf{K}, \\
\mathbf{X}'' = \mathrm{SN}(\mathrm{BN}(\mathrm{Linear}(\mathbf{X}'))).
\end{cases}
$$

- **Linear Computational Complexity.**
- **Spike-driven, High Energy Efficiency.**
- **No Scaling Operation.**

Method: QKFormer



- **Multi-scale spiking representation**
- **Mixed Spiking Attention Integration**



- **Identity mapping cross downsampling blocks in spiking patch embedding**
- **SNN-optimized Downsampling**

Reuslts: ImageNet-1K

Methods	Type	Architecture	Input Size	Param (M)	Power (mJ)	Time Step	Top-1 Acc (%)
RMP[21]	A2S	VGG-16	224^2	39.90	-	2048	73.09
QCFS[22]	A2S	ResNet-18	224^2	11.70	-	1024	74.32
MST[23]	A2S	Swin Transformer-T	224^2	28.50	-	512	78.51
SEW ResNet[28]	SNN	SEW-ResNet-34	224^2	21.79	4.89	4	67.04
	SNN	SEW-ResNet-101	224^2	44.55	8.91	4	68.76
	SNN	SEW-ResNet-152	224^2	60.19	12.89	4	69.26
Spikformer[11]	SNN	Spikformer-8-384	224^2	16.81	7.73	4	70.24
	SNN	Spikformer-8-512	224^2	29.68	11.58	4	73.38
	SNN	Spikformer-8-768	224^2	66.34	21.48	4	74.81
Spikingformer[12]	SNN	Spikingformer-8-384	224^2	16.81	4.69	4	72.45
	SNN	Spikingformer-8-512	224^2	29.68	7.46	4	74.79
	SNN	Spikingformer-8-768	224^2	66.34	13.68	4	75.85
S-Transformer[13]	SNN	S-Transformer-8-384	224^2	16.81	3.90	4	72.28
	SNN	S-Transformer-8-512	224^2	29.68	1.13	1	71.68
	SNN	S-Transformer-8-512	224^2	29.68	4.50	4	74.57
	SNN	S-Transformer-8-768*	288^2	66.34	6.09	4	77.07
ViT[4]	ANN	ViT-B/16	384^2	86.59	254.84	1	77.90
DeiT[32]	ANN	DeiT-B	224^2	86.59	80.50	1	81.80
	ANN	DeiT-B	384^2	86.59	254.84	1	83.10
Swin[8]	ANN	Swin Transformer-B	224^2	87.77	70.84	1	83.50
	ANN	Swin Transformer-B	384^2	87.77	216.20	1	84.50
QKFormer	SNN	HST-10-384	224^2	16.47	15.13	4	78.80
	SNN	HST-10-512	224^2	29.08	21.99	4	82.04
	SNN	HST-10-768	224^2	64.96	8.52	1	81.69
	SNN	HST-10-768	224^2	64.96	38.91	4	84.22
	SNN	HST-10-768*	288^2	64.96	64.27	4	85.25
	SNN	HST-10-768**	384^2	64.96	113.64	4	**85.65**

- **Compared with SNNs:**
QKFormer is the first directly trained SNN model, which has **exceeded 85% accuracy** on ImageNet-1K. The top-5 accuracy of QKFormer (HST-10-768□ □) is 97.74%. Notably, with comparable size to Spikformer (66.34 M, 74.81%), QKFormer (64.96 M) achieves a ground-breaking top-1 accuracy of 85.65% on ImageNet-1k, substantially outperforming Spikformer by **10.84%**.

- **Compared with ANNs:**
QKFormer is a directly trained SNN model that has surpassed many transformer ANNs on ImageNet-1K. Under the same experiment conditions without pre-training or extra training data: **QKFormer (64.96M, 85.65%, SNN) > Swin Transformer(88M, 84.5%, ANN)** > DeiT-B (86M, 83.1%, ANN) > ViT (85.9M, 77.9%, ANN) .

Reuslts: CIFAR10, CIFAR100, DVS128, CIFAR10-DVS

Method	CIFAR10			CIFAR100			DVS128			CIFAR10-DVS		
	Param	T	Acc	Param	T	Acc	Param	T	Acc	Param	T	Acc
Spikformer [11]	9.32	4	95.51	9.32	4	78.21	2.57	16	98.3	2.57	16	80.9
Spikingformer [12]	9.32	4	95.81	9.32	4	78.21	2.57	16	98.3	2.57	16	81.3
CML [14]	9.32	4	96.04	9.32	4	80.02	2.57	16	98.6	2.57	16	80.9
S-Transformer[13]	10.28	4	95.60	10.28	4	78.4	2.57	16	**99.3**	2.57	16	80.0
STSA[15]	–	–	–	–	–	–	1.99	16	98.7	1.99	16	79.93
ResNet-19 (ANN)	12.63	1	94.97	12.63	1	75.35	–	–	–	–	–	–
Trasnformer (ANN)	9.32	1	96.73	9.32	1	81.02	–	–	–	–	–	–
QKFormer	6.74	4	**96.18**	6.74	4	**81.15**	1.50	16	98.6	1.50	16	**84.0**

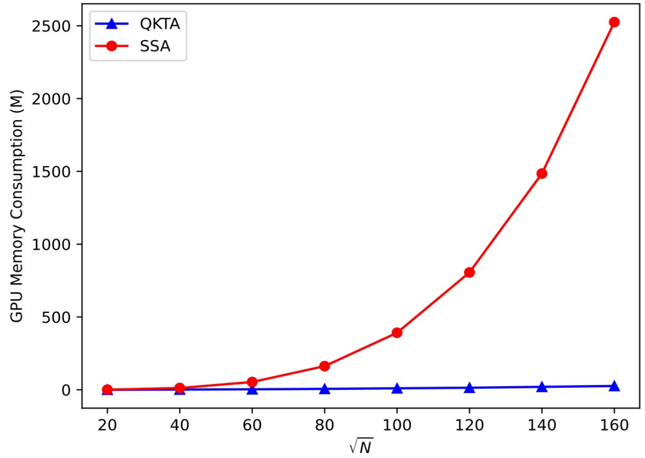
Model	CIFAR100 (Acc)	CIFAR10-DVS (Acc)
QKFormer (QKTA + SSA, baseline)	81.15%	84.00%
QKFormer (QKTA + SSA, w/o SPEDS)	80.08%	83.40%
Spikformer (SSA, w/o scaling)	76.95%	79.30%
Spikformer (SSA)	78.21%	80.90%
Spikformer (SSA) + SPEDS	80.26%	82.20%

Model	CIFAR100 (Acc, Param)	CIFAR10-DVS (Acc, Param)
QKFormer (QKTA + SSA, baseline)	81.15%, 6.74M	84.00%, 1.50M
QKFormer (QKCA + SSA)	81.07%, 6.74M	84.30%, 1.50M
QKFormer (QKTA + QKCA)	81.04%, 6.44M	83.10%, 1.44M
QKFormer (SSA)	81.23%, 6.79M	84.10%, 1.52M
QKFormer (QKCA)	81.00%, 6.44M	80.70%, 1.44M
QKFormer (QKTA)	79.09%, 6.44M	80.70%, 1.44M

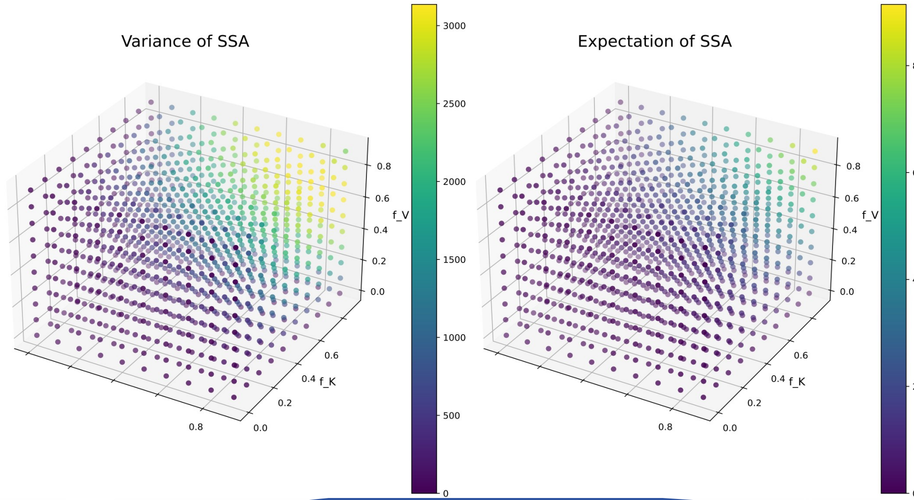
- QKFormer achieved SOTA performance on both CIFAR and Neuronoiphic datasets: fewer parameters, higher performance.
- SPEDS module is essential to QKFormer on both static and neuromorphic datasets. In addition, the addition of SPEDS to Spikformer leads to great gains.
- Mixed spiking attention solutions, such as QKFormer(QKTA + SSA),can achieve comparable performance to QKFormer(SSA) while requiring fewer parameters and much fewer memory resources.

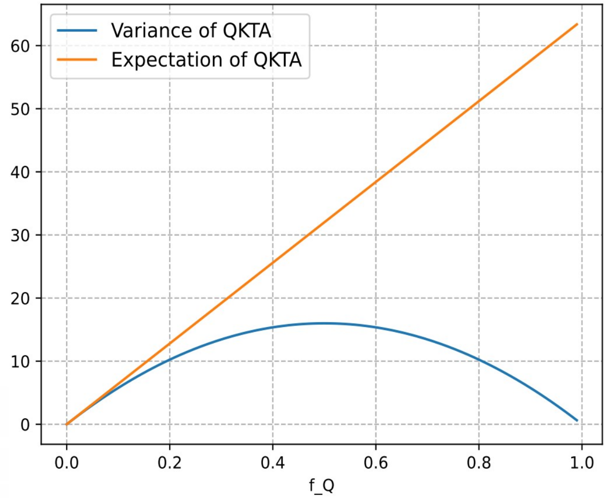
Reuslts: More Analyses

QKFormer Block		Stage1 (fr)	Stage2 (fr)
QKTA	$\mathbf{Q}$	0.0432	0.0231
	$\mathbf{K}$	0.1784	0.0847
	$\mathbf{A}_t$	0.3477	0.2655
	$\mathbf{X}'$	0.0832	0.0350
	$\mathbf{X}''$	0.1478	0.0577
SMLP	Layer1	0.0518	0.0246
	Layer2	0.2733	0.1869



- **Low firing rate.**
- **Low Computational Complexity.**





- **More stable Variance and Expectation.**

Reuslts: Conclusion & Discussion

This work achieves a large improvement (+10.84%) above the state of the art in spiking neural networks.
With its powerful performance, we aim for our investigations to instill optimism in the application of SNNs.

Thanks for your attention!

If you have any question or suggestion, please feel free to contact:
zhouchl@pcl.ac.cn or *zhouchenlin19@mails.ucas.ac.cn.*