香港大學
The University of Hong Kong

EZCrop: Energy-Zoned Channels for Robust Output Pruning

Rui Lin[1, *] Jie Ran[1, *] Dongpeng Wang[2] King Hung Chiu[2] Ngai Wong[1]

[1]Dept. of Electrical and Electronic Engineering, The University of Hong Kong

[2]United Microelectronic Center (Hong Kong) Limited

Video Presentation for WACV 2022

* RL, JR contributed equally to this work.

Source Code: https://github.com/RuiLin0212/EZCrop

Content

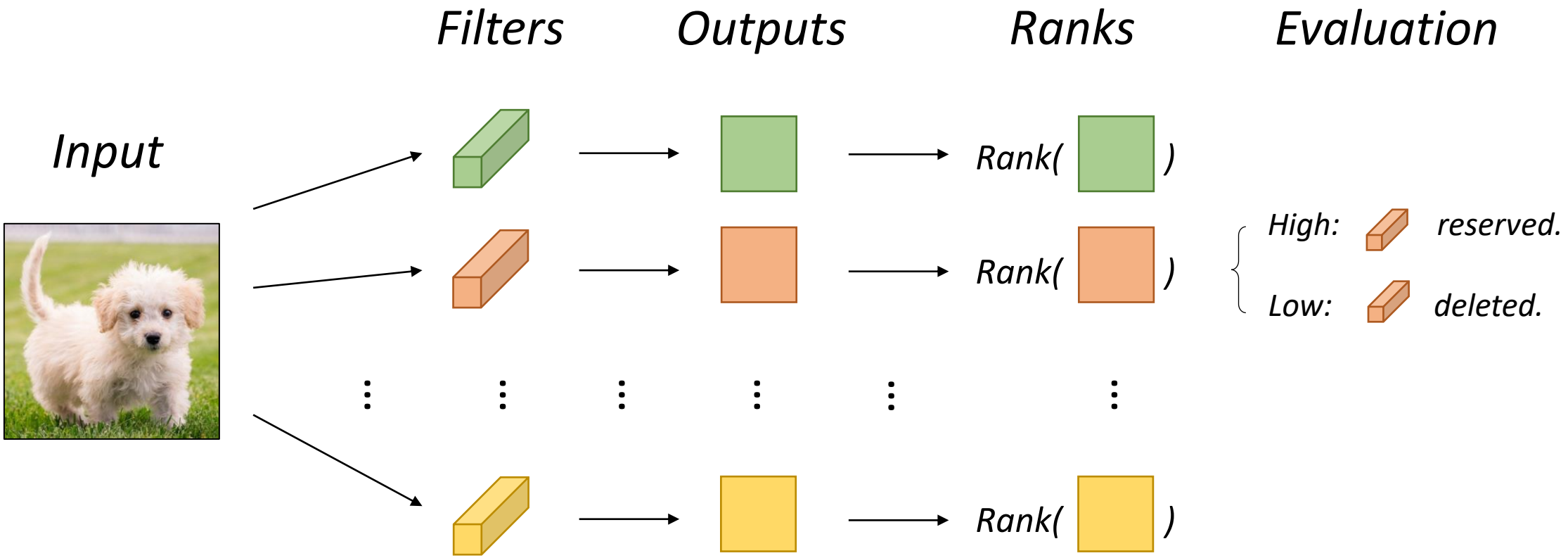
1. ***Preliminary***
   - HRank: a Rank-based Channel Pruning Algorithm
   - Convolution and Matrix Ranks from the Frequency Domain Viewpoint

2. ***EZCrop***

3. ***Experimental Results***

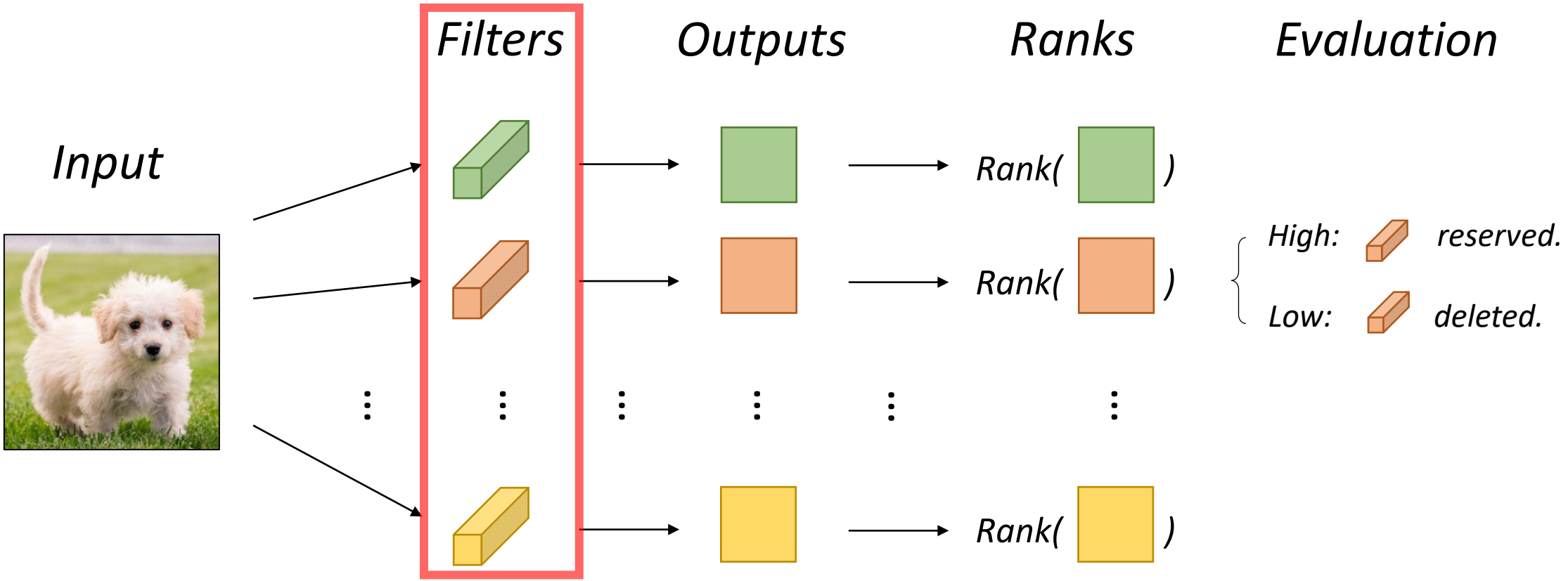
4. ***Conclusion***

HRank[1]: a Rank-based Channel Pruning Algorithm



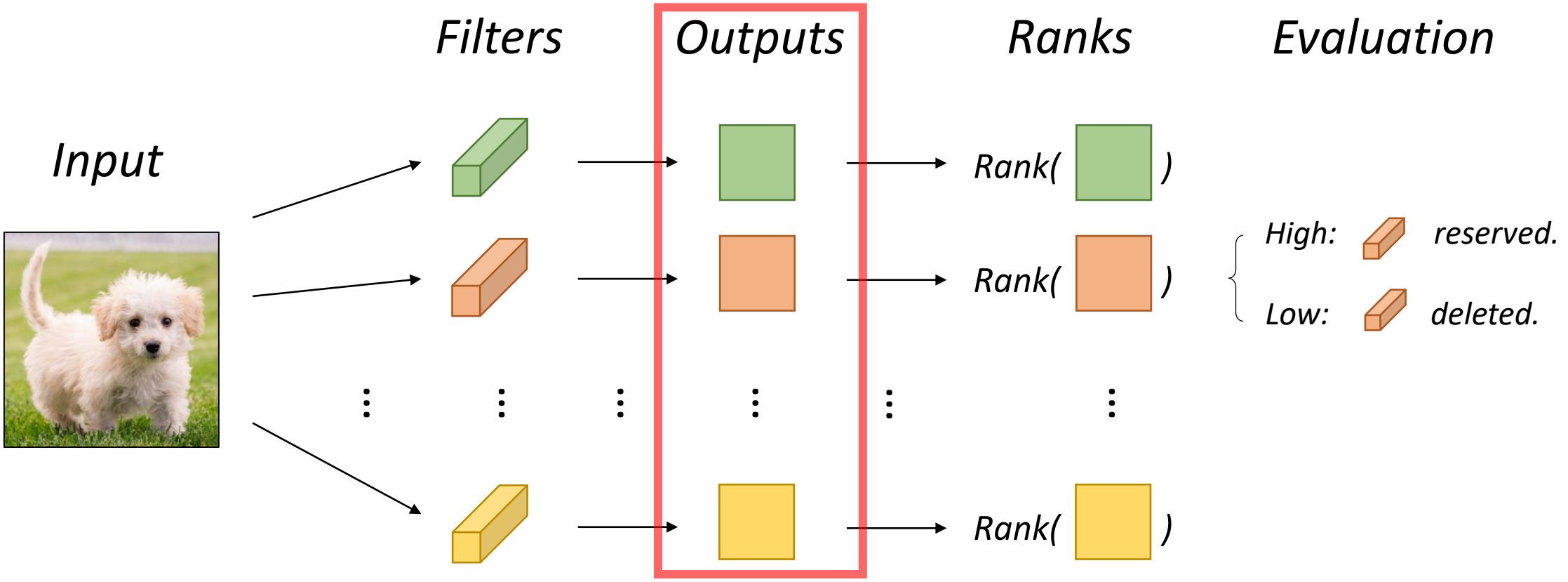
1 Hrank: Filter pruning using high-rank feature map (CVPR 2020)

HRank[1]: a Rank-based Channel Pruning Algorithm



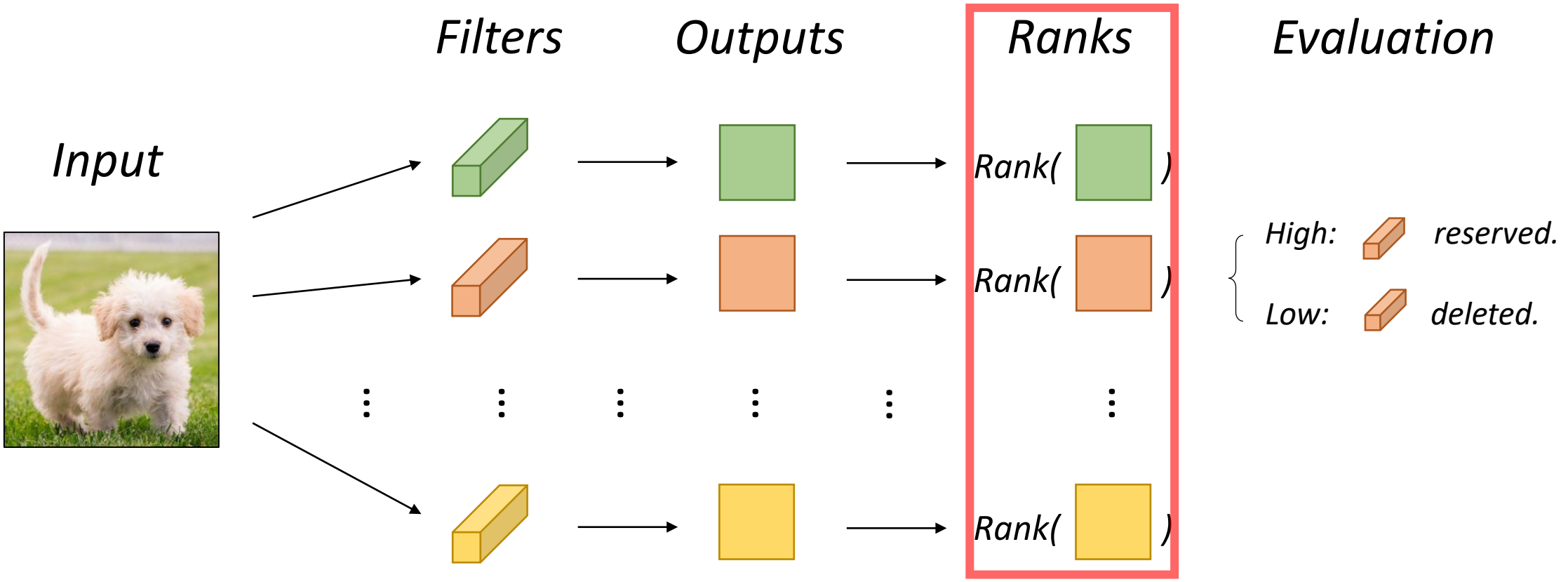
1 Hrank: Filter pruning using high-rank feature map (CVPR 2020)

HRank[1]: a Rank-based Channel Pruning Algorithm



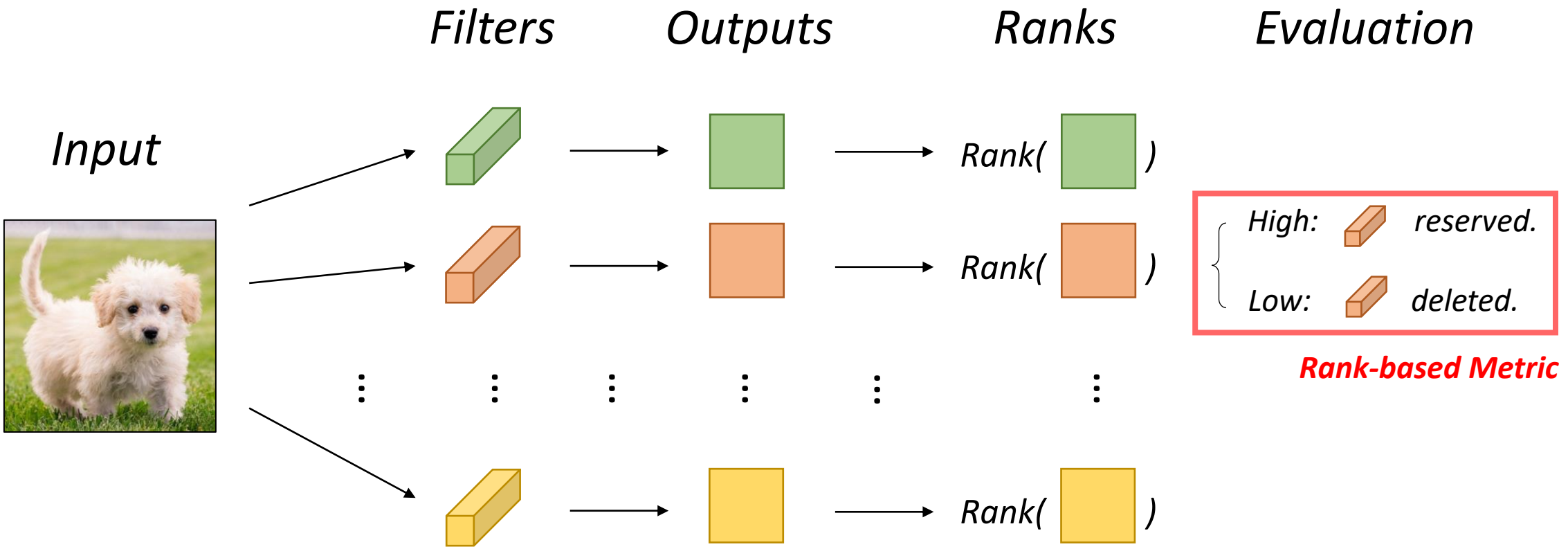
1 Hrank: Filter pruning using high-rank feature map (CVPR 2020)

HRank[1]: a Rank-based Channel Pruning Algorithm



1 Hrank: Filter pruning using high-rank feature map (CVPR 2020)

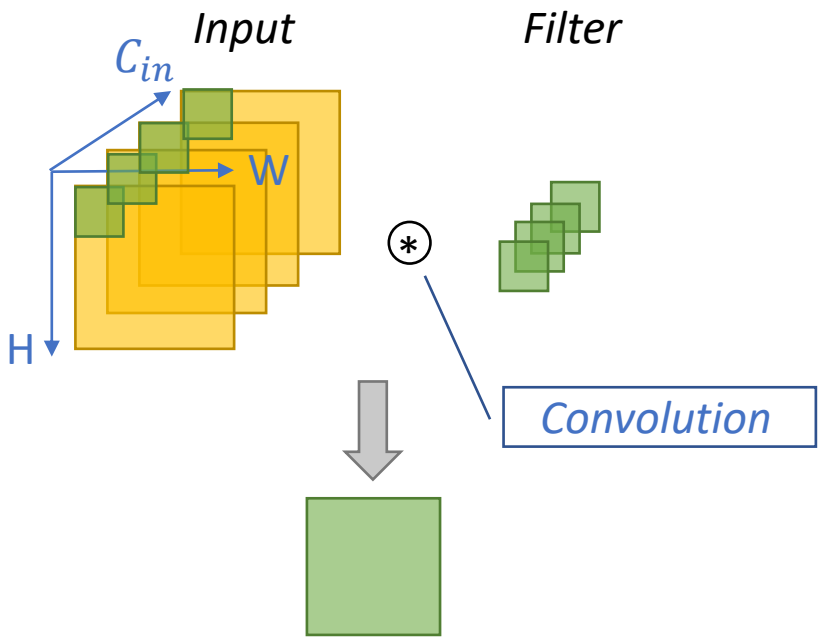
HRank[1]: a Rank-based Channel Pruning Algorithm



1 Hrank: Filter pruning using high-rank feature map (CVPR 2020)

Convolution in the Frequency Domain

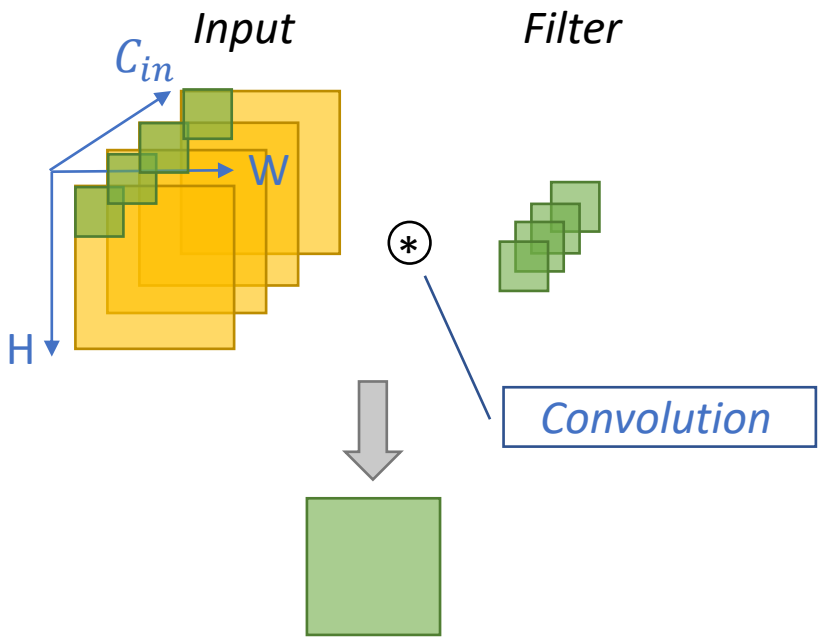
Spatial Domain



$$\mathcal{Y}[j,:,:] = \mathcal{X} \circledast \mathcal{K}[:,:,:,j]$$

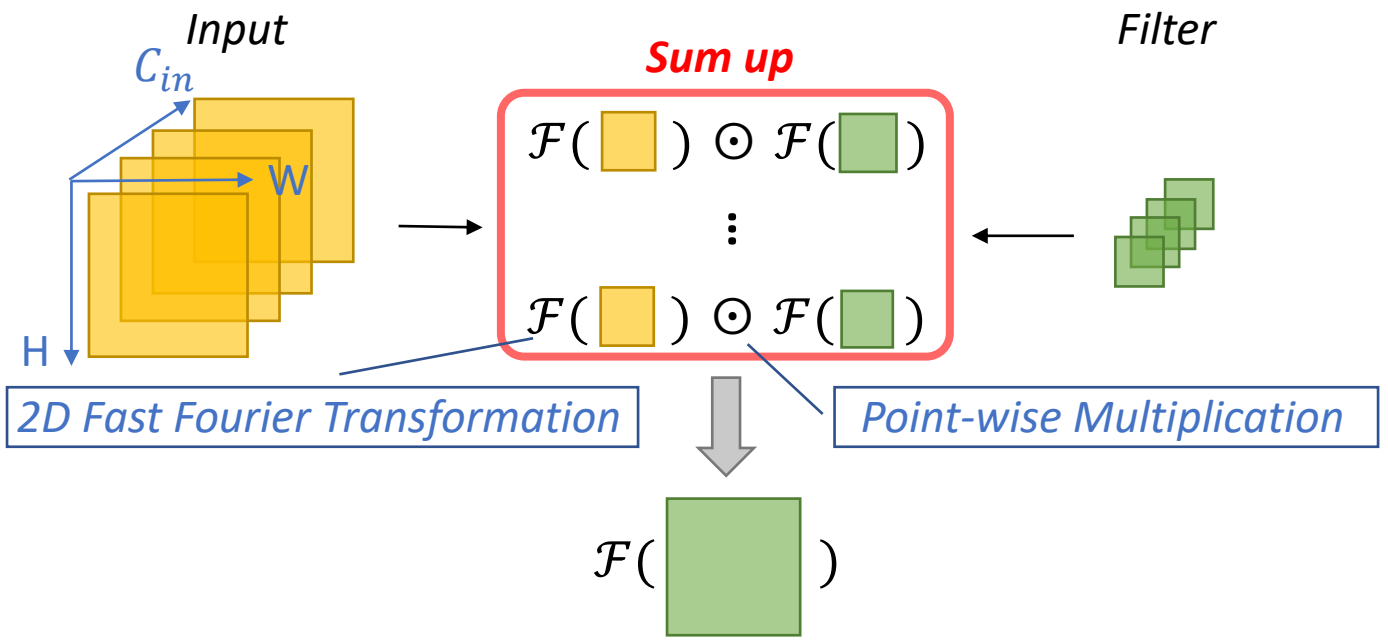
Convolution in the Frequency Domain

Spatial Domain



$$\mathcal{Y}[j,:,:] = \mathcal{X} \circledast \mathcal{K}[:,:,:,j]$$

Frequency Domain



$$\mathcal{F}(\mathcal{Y}[j,:,:]) = \mathcal{F}(\mathcal{X}[i,:,:]) \odot \mathcal{F}(\widehat{\mathcal{K}}[:,:,:,j])$$

Convolution in the Frequency Domain

Spatial Domain

Rank-based *filter importance evaluation*

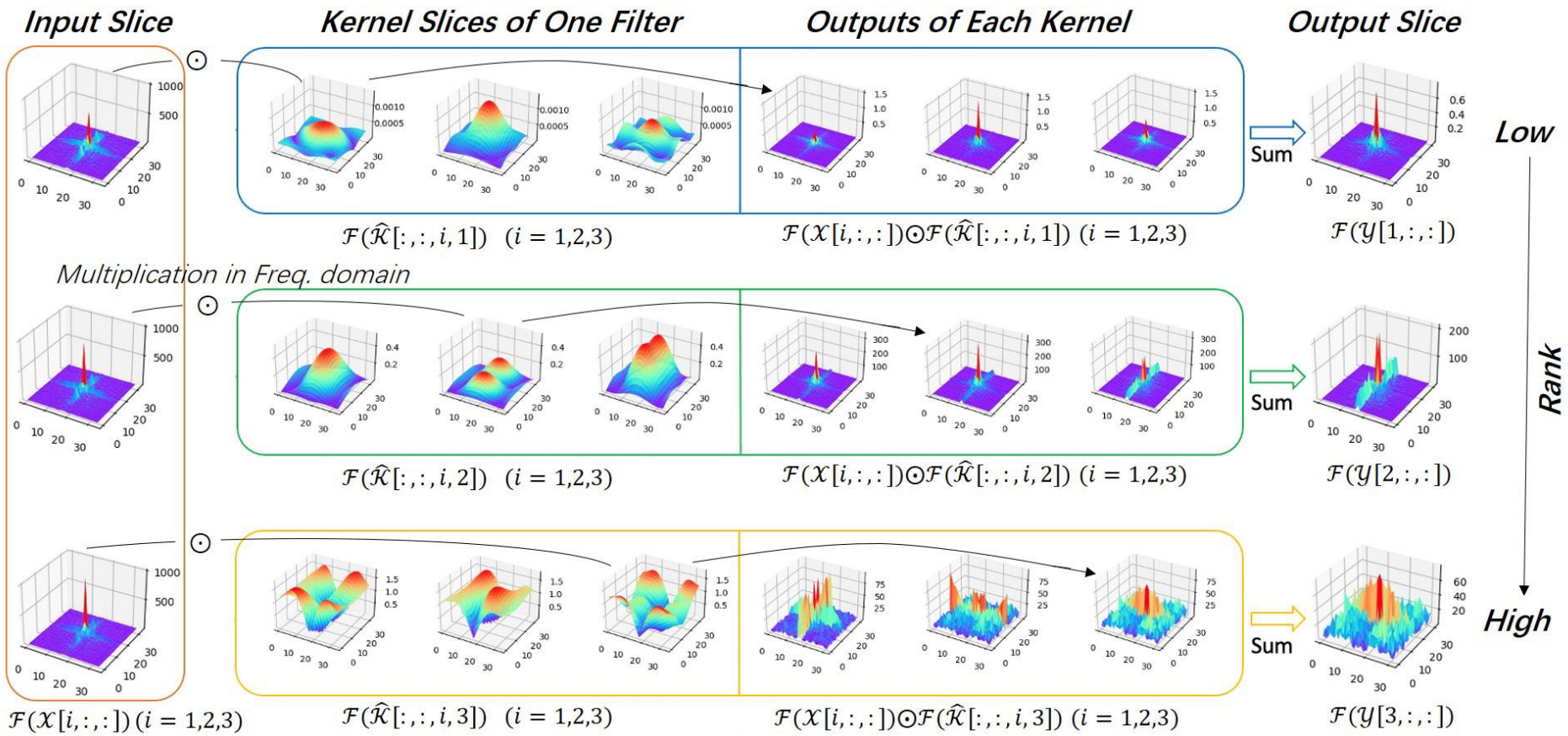
$Rank($  $)$ { *High:*  *reserved.*; *Low:*  *deleted.*

Frequency Domain

What ***metric in the frequency domain*** *can help evaluate the filters' importance?*

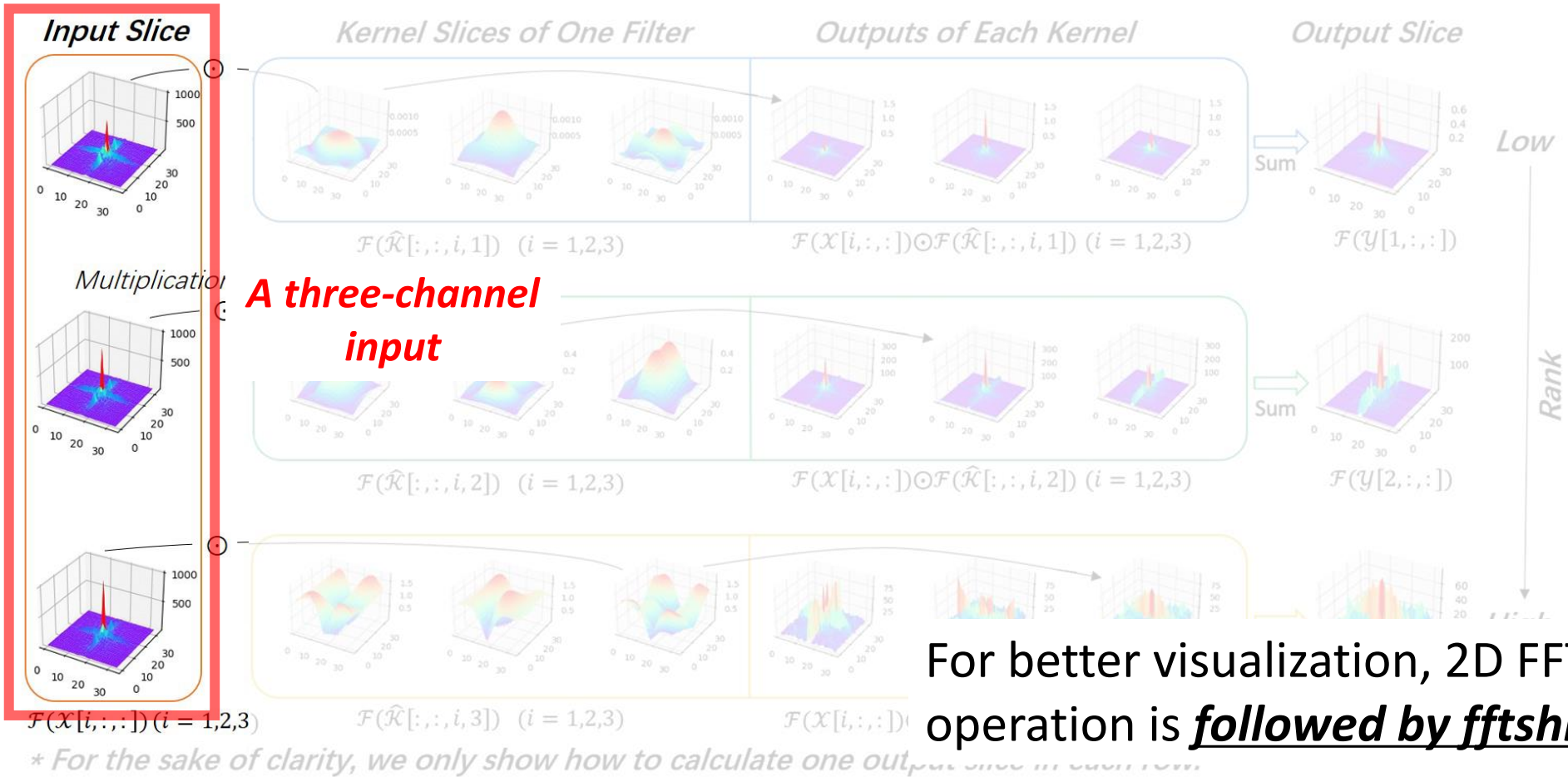
$\mathcal{F}($  $)$

Matrix Ranks from the Frequency Domain Viewpoint



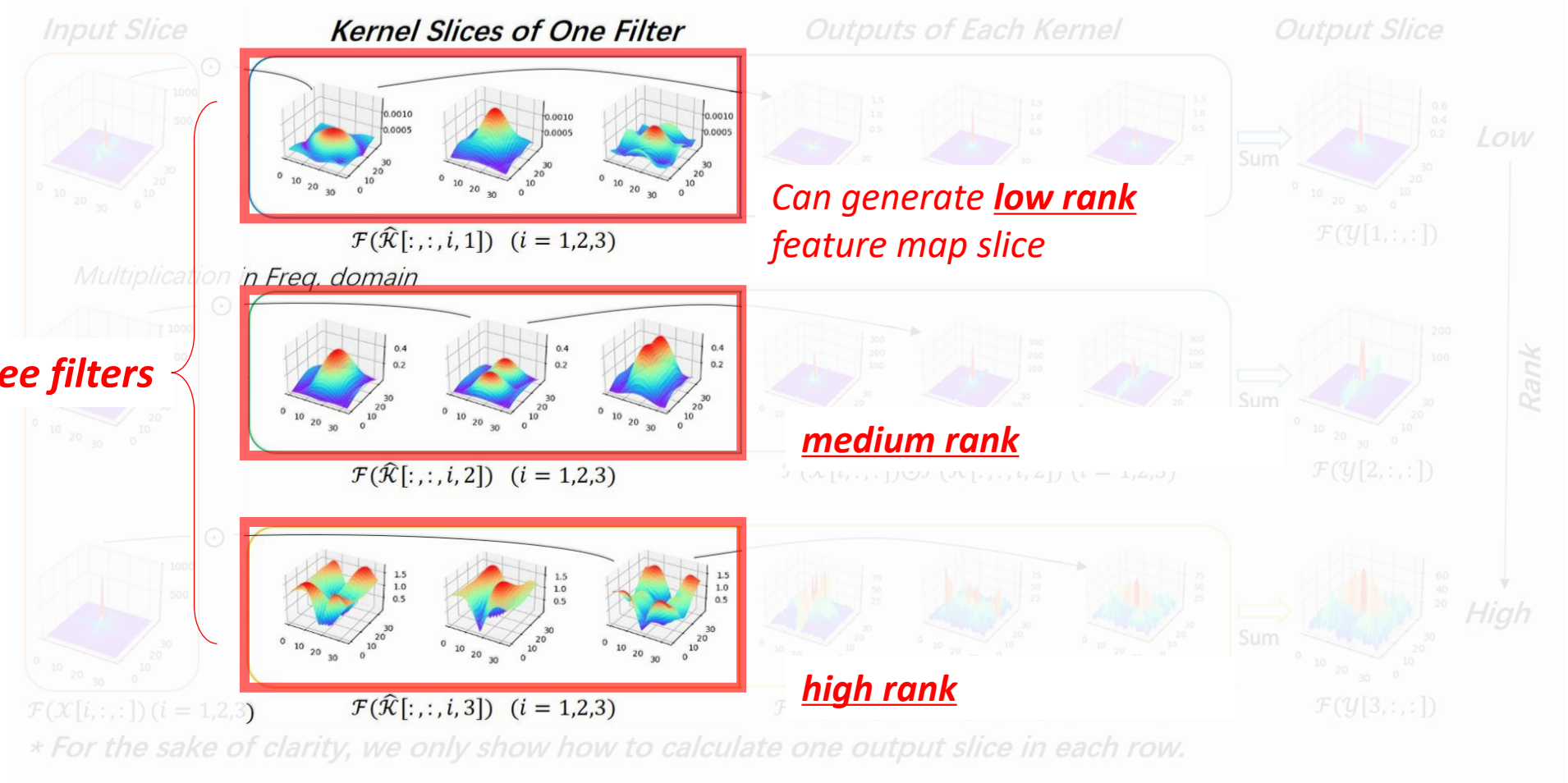
* For the sake of clarity, we only show how to calculate one output slice in each row.

Matrix Ranks from the Frequency Domain Viewpoint



For better visualization, 2D FFT operation is ***followed by fftshift(·).***

Matrix Ranks from the Frequency Domain Viewpoint

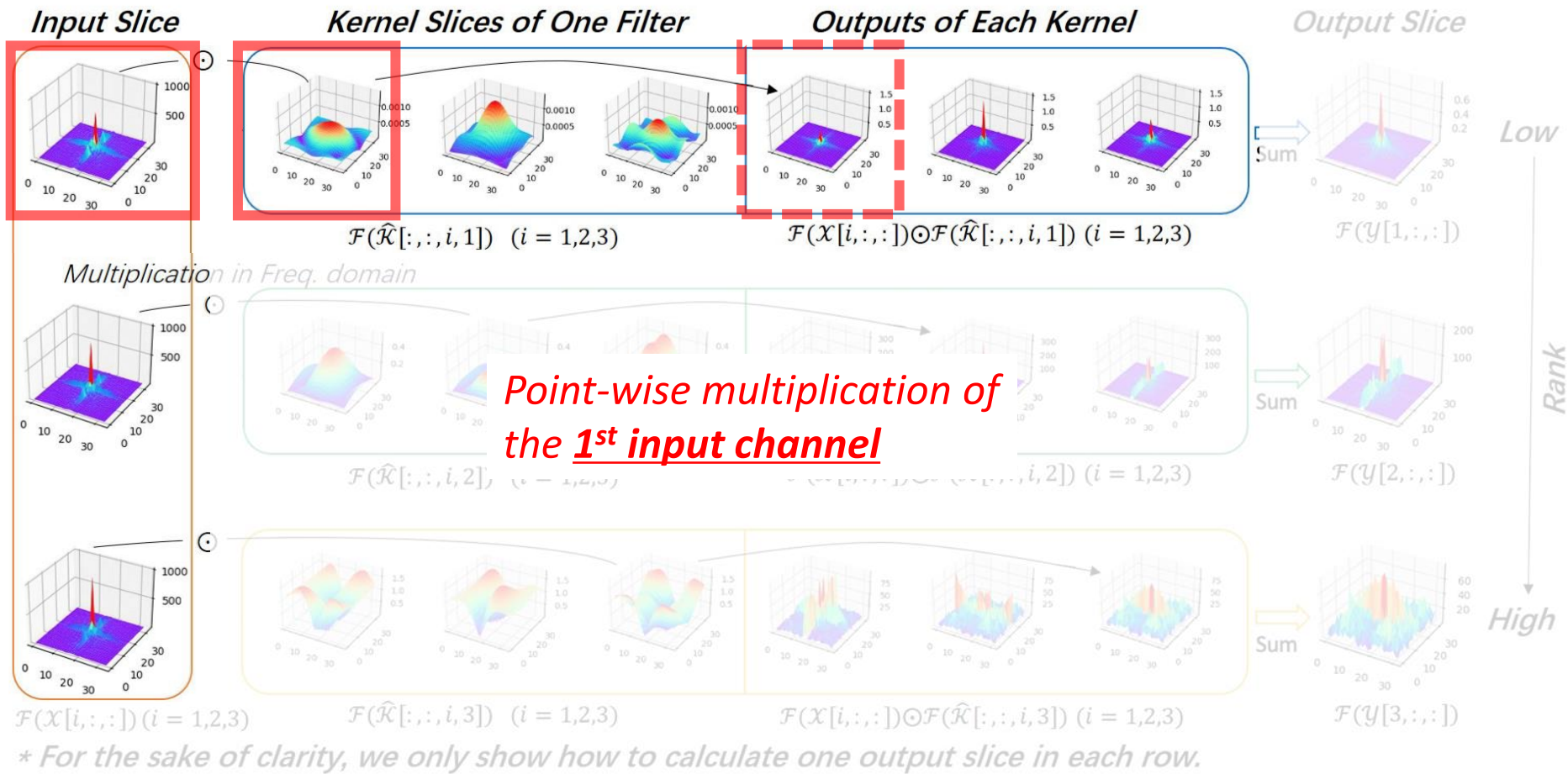


*Can generate **low rank** feature map slice*

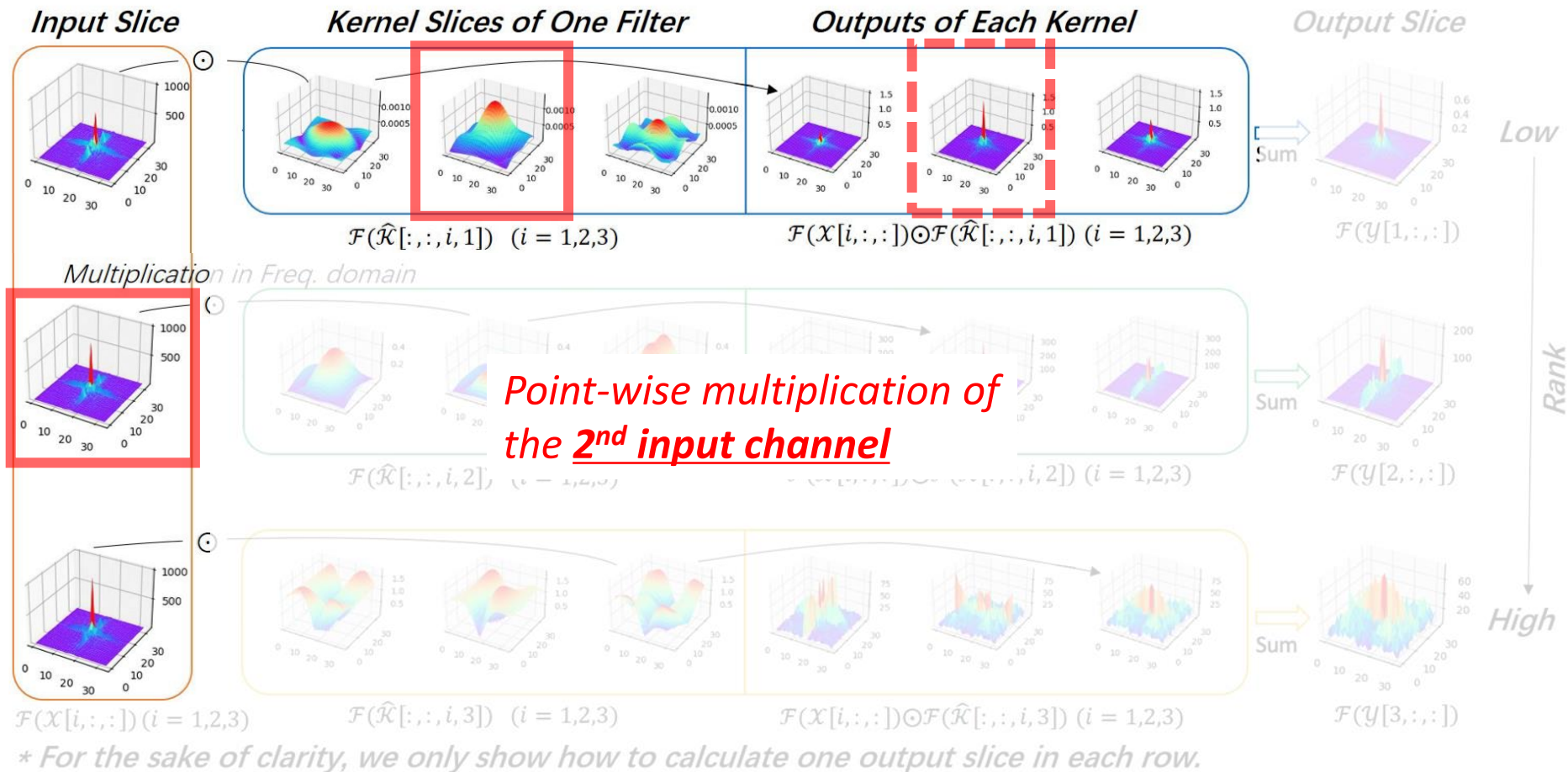
medium rank

high rank

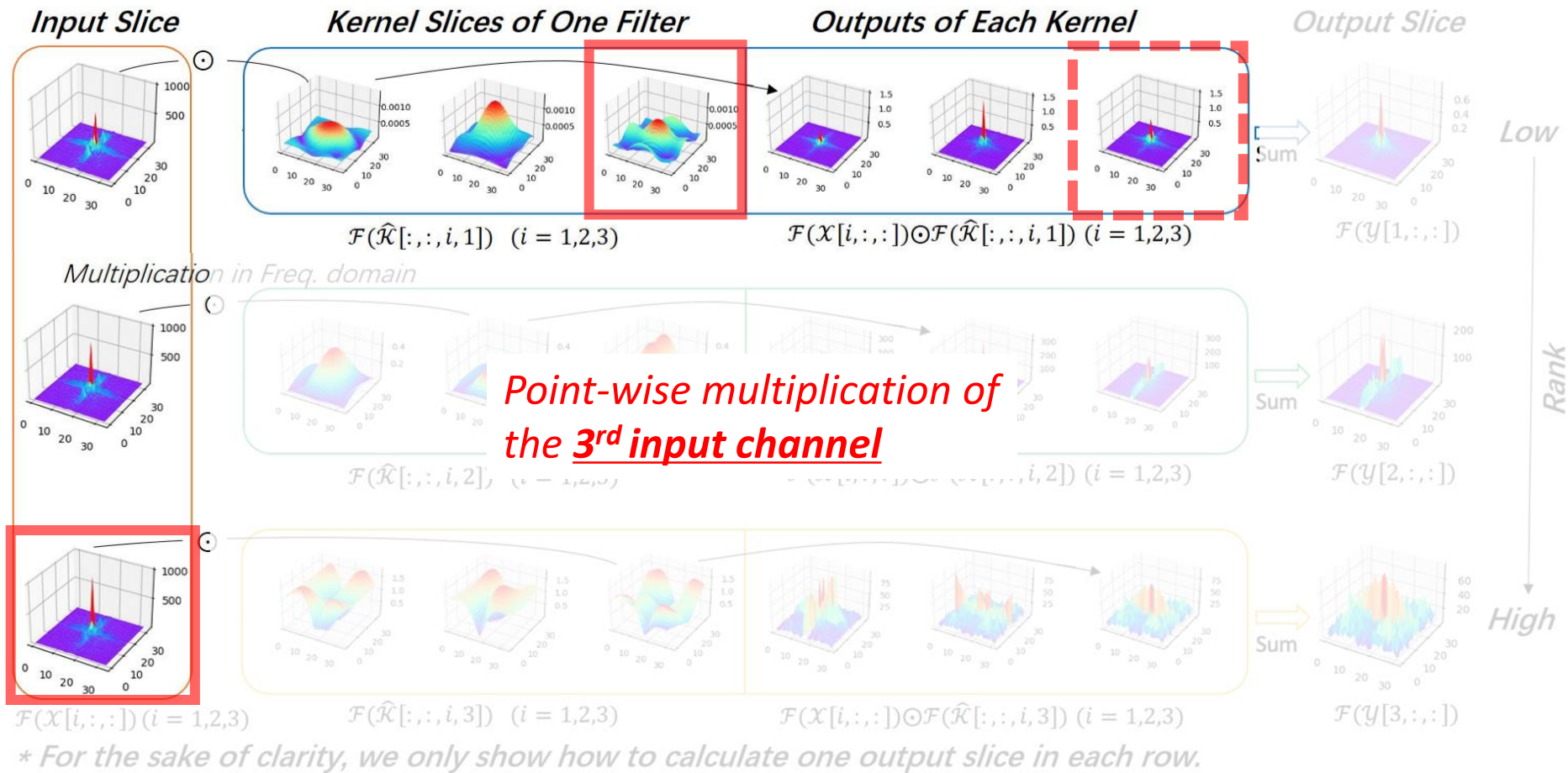
Matrix Ranks from the Frequency Domain Viewpoint



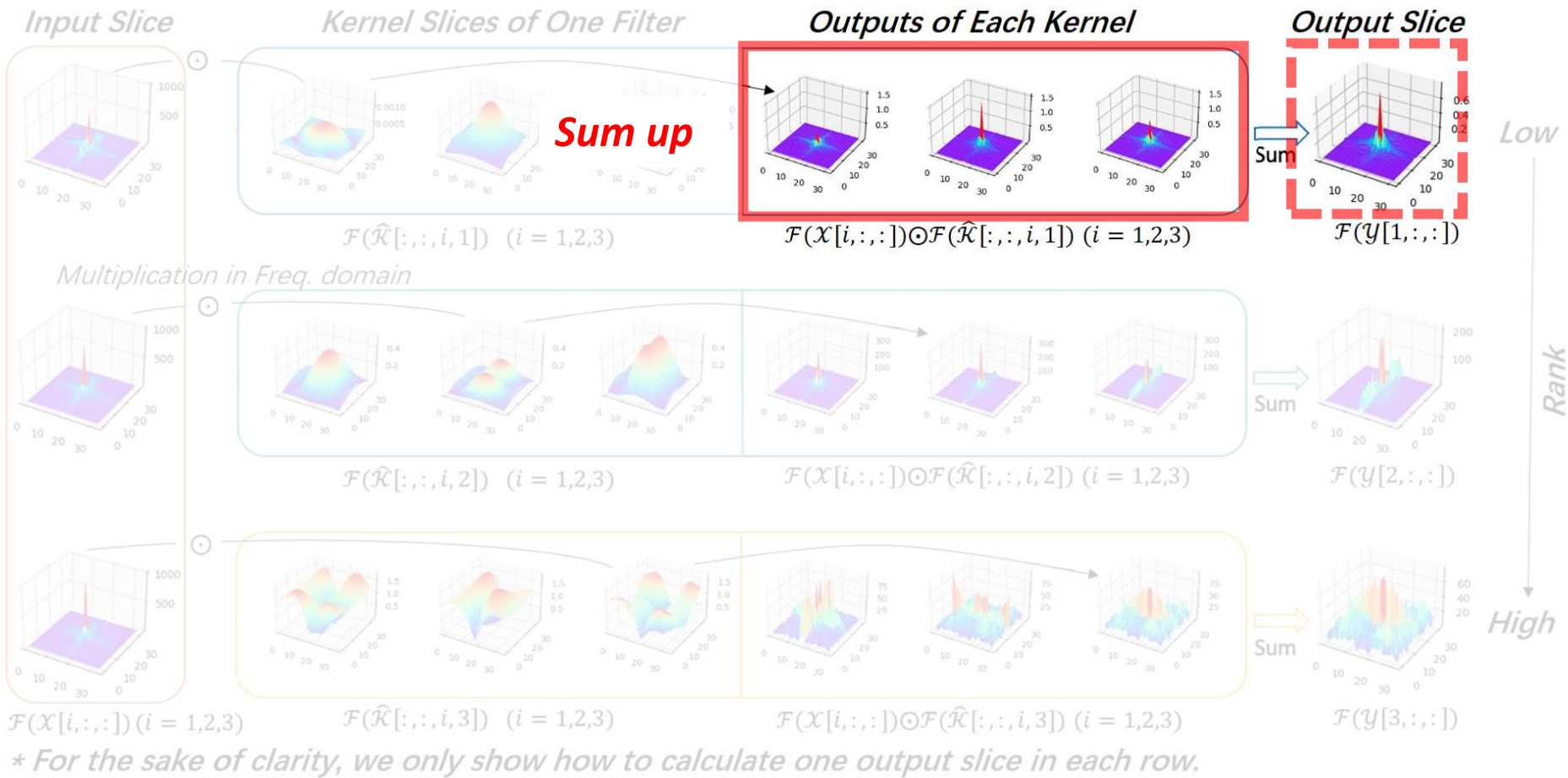
Matrix Ranks from the Frequency Domain Viewpoint



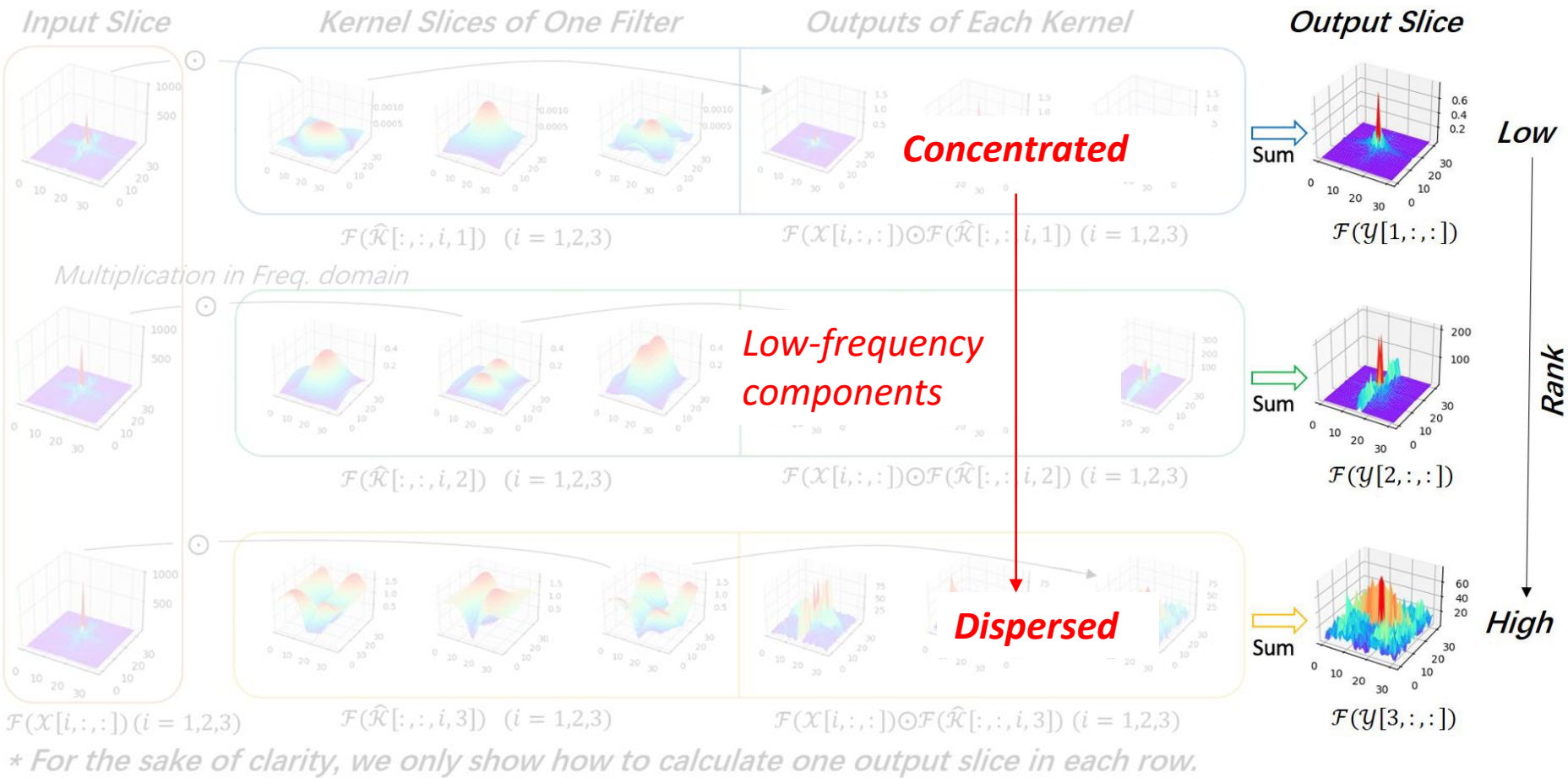
Matrix Ranks from the Frequency Domain Viewpoint



Matrix Ranks from the Frequency Domain Viewpoint



Matrix Ranks from the Frequency Domain Viewpoint



Matrix Ranks from the Frequency Domain Viewpoint

Spatial Domain

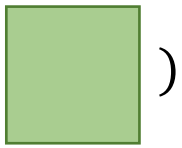
Rank-based *filter importance evaluation*

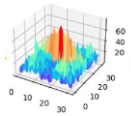
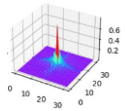
$Rank($  $)$

- *High:*  *reserved.*
- *Low:*  *deleted.*

Frequency Domain

Low frequency *distribution-based filter importance evaluation*

$\mathcal{F}($  $)$

- *Dispersed*  *:*  *reserved.*
- *Concentrated*  *:*  *deleted.*

Matrix Ranks from the Frequency Domain Viewpoint

Low-rank matrix in Freq. domain

0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.7	0.3	0.5	0.0	0.0
0.0	0.0	0.2	1.0	0.2	0.0	0.0
0.0	0.0	0.5	0.3	0.7	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0

High-rank matrix in Freq. domain

0.1	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.2	0.0	0.0	0.0	0.3	0.0
0.0	0.0	0.6	0.3	0.4	0.0	0.0
0.0	0.0	0.2	1.0	0.2	0.0	0.0
0.0	0.0	0.4	0.3	0.6	0.0	0.0
0.0	0.3	0.0	0.0	0.0	0.2	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.1

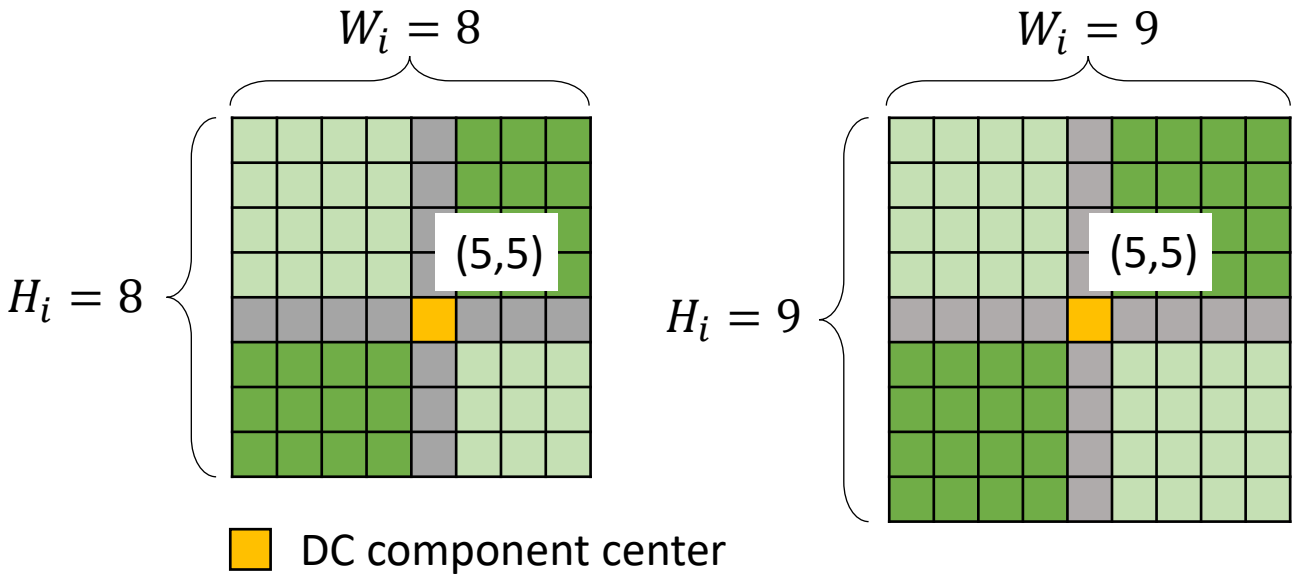
EZCrop

Step 1: Find the square Center

Step 2: Decide the Expanding Distance

Step 3: Calculate the Energy Zone Ratios

Step 4: Sort the filter



$$x_i = \begin{cases} \frac{H_i}{2} + 1, & H_i \text{ is even.} \\ \frac{H_i + 1}{2}, & H_i \text{ is odd.} \end{cases} \qquad y_i = \begin{cases} \frac{W_i}{2} + 1, & W_i \text{ is even.} \\ \frac{W_i + 1}{2}, & W_i \text{ is odd.} \end{cases}$$

EZCrop

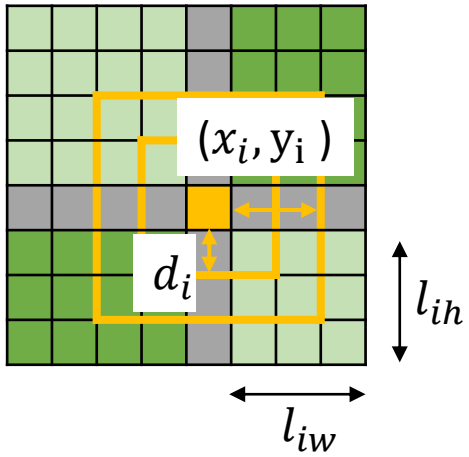
Step 1: Find the square Center

Step 2: Decide the Expanding Distance

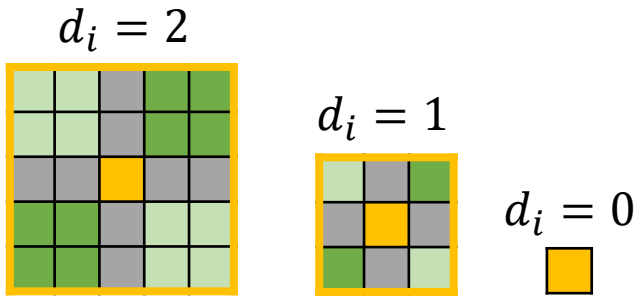
Step 3: Calculate the Energy Zone Ratios

Step 4: Sort the filter

Output in freq. domain



Selected energy area



$$l_{ih} = H_i - x_i$$
$$l_{iw} = W_i - y_i$$

$$d_i = \begin{cases} 0, l_{ih} \leq 1 \; or \; l_{iw} \leq 1 \\ ceil(\boldsymbol{\beta} \cdot \min(l_{ih}, l_{iw})), else. \end{cases}$$

EZCrop

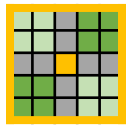
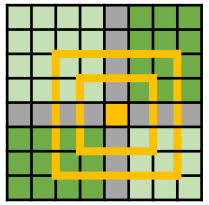
Step 1: Find the square Center

Step 2: Decide the Expanding Distance

Step 3: Calculate the Energy Zone Ratios

Step 4: Sort the filter

$$\eta_i^j = 1 - \frac{1}{B} \cdot \sum_{b=1}^{B} \frac{S(d_i[b])}{S(E_i^j[b,:,:])}$$

$$\eta_i^j = 1 - \frac{1}{\mathrm{B}} \cdot \sum_{b=1}^{B} \frac{sum(\ \text{  }\ [b]\ )}{sum(\ \text{  }\ [b])}$$

$$\eta_i^j = \begin{cases} large, & dispersed\ (filter\ \textbf{reserved}) \\ small, & concentrated\ (filter\ \textbf{deleted}) \end{cases}$$

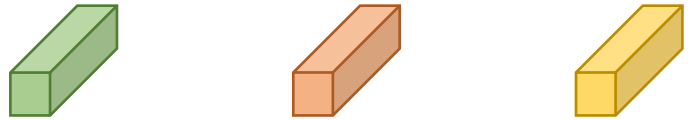
EZCrop

Step 1: Find the square Center

Step 2: Decide the Expanding Distance

Step 3: Calculate the Energy Zone Ratios

Step 4: Sort the filter

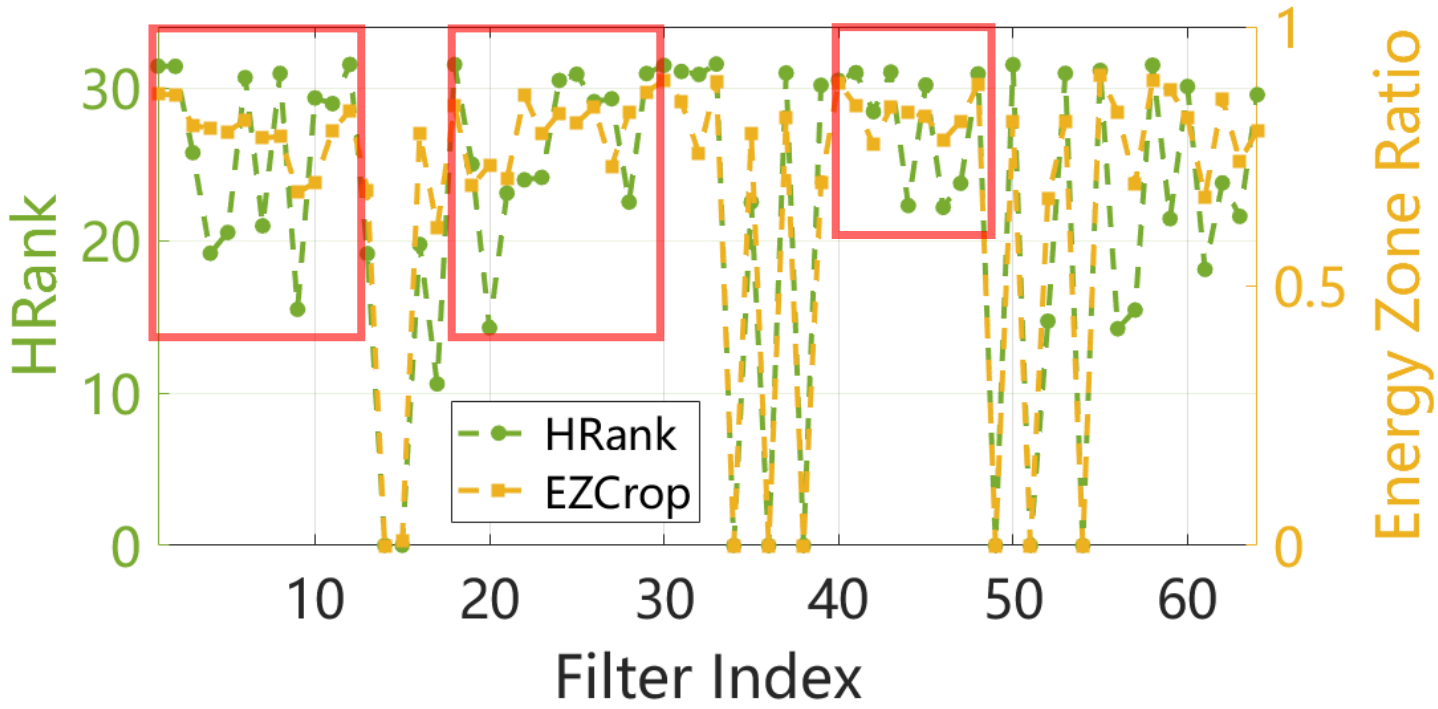


$\eta_1^i = 0.45$ $\eta_2^i = 0.85$ $\eta_3^i = 0.75$

Importance:

EZCrop

*The two lines **generally track each other**, but **some** evaluation results are **slightly different**.*



Time Comparison

Dataset	Model	HRrank [23]	**EZCrop** (↓)
CIFAR-10	VGGNet	1505.54s	**356.94**s (76.29%)
	ResNet-56	1247.51s	**381.97**s (69.38%)
	DenseNet-40	473.17s	**171.50**s (63.76%)
ImageNet	ResNet-50	7.96h	**3.45**h (56.66%)

Much more efficient!

VGG-16 on CIFAR-10

Model	Top-1%	FLOPs (↓)	Params (↓)
VGGNet	93.96	313.73M(0.0%)	14.98M(0.0%)
L1 [20]	93.40	206.00M(34.3%)	5.40M(64.0%)
SSS [15]	93.02	183.13M(41.6%)	3.93M(73.8%)
Zhao *et al.* [46]	93.18	190.00M(39.1%)	3.92M(73.3%)
GAL-0.05 [27]	92.03	189.49M(39.6%)	3.36M(77.6%)
GAL-0.1 [27]	90.78	171.89M(45.2%)	2.67M(82.2%)
FPGM [10]	94.00	201.10M(35.9%)	–
PScratch [44]	93.63	156.86M(50.0%)	–
HRank [23]	93.73	131.17M(58.1%)	2.76M(81.6%)
EZCrop	**94.01**	**131.17M**(58.1%)	**2.76M**(81.6%)
HRank [23]	93.56	104.78M(66.6%)	2.50M(83.3%)
EZCrop	**93.70**	**104.78M**(66.6%)	**2.50M**(83.3%)

Better *performance when compression level is* ***similar****.*

ResNet-50 on ImageNet

Model	Top-1%	Top-5%	FLOPs	Params
ResNet-50 [32]	76.15	92.87	4.09B	25.50M
He *et al.* [11]	72.30	90.80	2.73B	–
ThiNet-50 [32]	68.42	88.30	1.10B	8.66M
SSS-26 [15]	71.82	90.79	2.33B	15.60M
SSS-32 [15]	74.18	91.91	2.82B	18.60M
GDP-0.5 [26]	69.58	90.14	1.57B	–
GDP-0.6 [26]	71.19	90.71	1.88B	–
GAL-0.5 [27]	71.95	90.94	2.33B	21.20M
GAL-1 [27]	69.88	89.75	1.58B	14.67M
GAL-0.5-joint [27]	71.80	90.82	1.84B	19.31M
GAL-1-joint [27]	69.31	89.12	1.11B	10.21M
FPGM [10]	75.91	92.63	2.36B	–
MetaPruning [30]	75.40	–	2.29B	–
DMCP [6]	76.20	–	2.20B	–
EagleEye [19]	76.40	92.89	2.00B	–
ABCPrunner-80% [21]	73.86	91.69	1.89B	11.75M
HRank [23]	75.56	92.63	2.26B	15.09M
EZCrop	**75.68**	**92.70**	2.26B	15.09M
HRank [23]	74.19	91.94	1.52B	11.05M
EZCrop	**74.33**	**92.00**	1.52B	11.05M

Better *performance when compression level is* ***similar****.*

Repetitive Pruning of ResNet-56 on CIFAR-10

Repetitive Pruning:

1. Do filter importance evaluation
2. Prune some trivial filters
3. Fine tune the model (300 epochs)
4. ***Repeat step 1-3*** for 2 more times

#Passes (#epochs)	FLOPs	Params	HRank [23]	**EZCrop**	*Acc. gap*
1 (300)	$90.86M$	$0.63M$	93.76%	**93.95%**	0.19%
2 (300)	$66.25M$	$0.46M$	93.15%	**93.42%**	0.27%
3 (300)	$36.03M$	$0.22M$	91.58%	**92.18%**	0.60%

*EZCrop enjoys **higher** robustness*

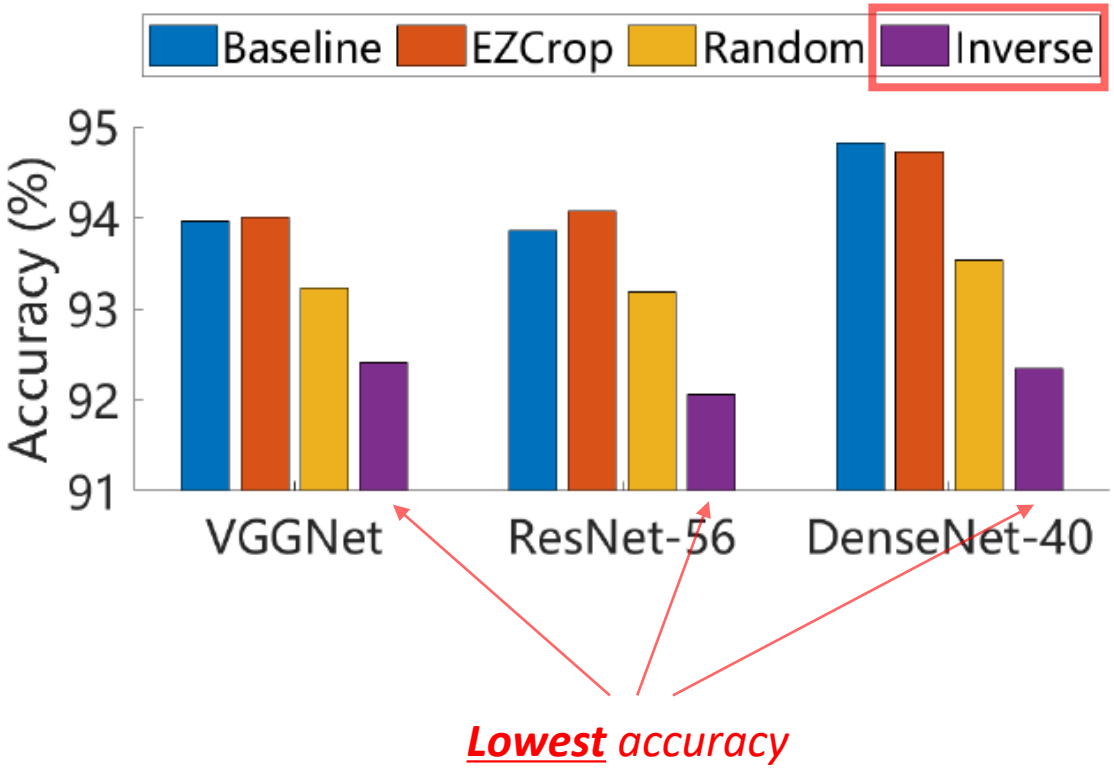
Standard Deviation (STD) Analysis

Each experiment is repeated for 20 times.

Mean Acc. (%) / STD	VGGNet	ResNet-56	DenseNet-40
EZCrop	**93.98** / **0.073**	**93.99** / **0.097**	**94.63** / **0.066**
HRank [23]	93.75 / 0.140	93.76 / 0.168	94.26 / 0.165

*EZCrop makes **more accurate and robust** evaluation of filter importance.*

Effectiveness of Energy-zone Rate



Conclusion

1. This work connected the ***matrix rank in the spatial domain*** to the ***low frequency component distribution in the frequency domain***.

2. The proposed FFT-based metric for filter importance evaluation is ***efficient***.

3. EZCrop brings ***higher resolution*** in channels' importance evaluation.

4. EZCrop constitutes a ***robust*** way for repetitive channel pruning.

Thanks for Your Attention!

If you have any question, please contact us.