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 EEE, The University of Hong Kong [2] United Microelectronics Center (Hong Kong) Limited, HKSPT, Hong Kong *Equal Contribution

IEEE 2022 Winter Conference on Applications of Computer Vision

1. Preliminary

1. 1 HRank

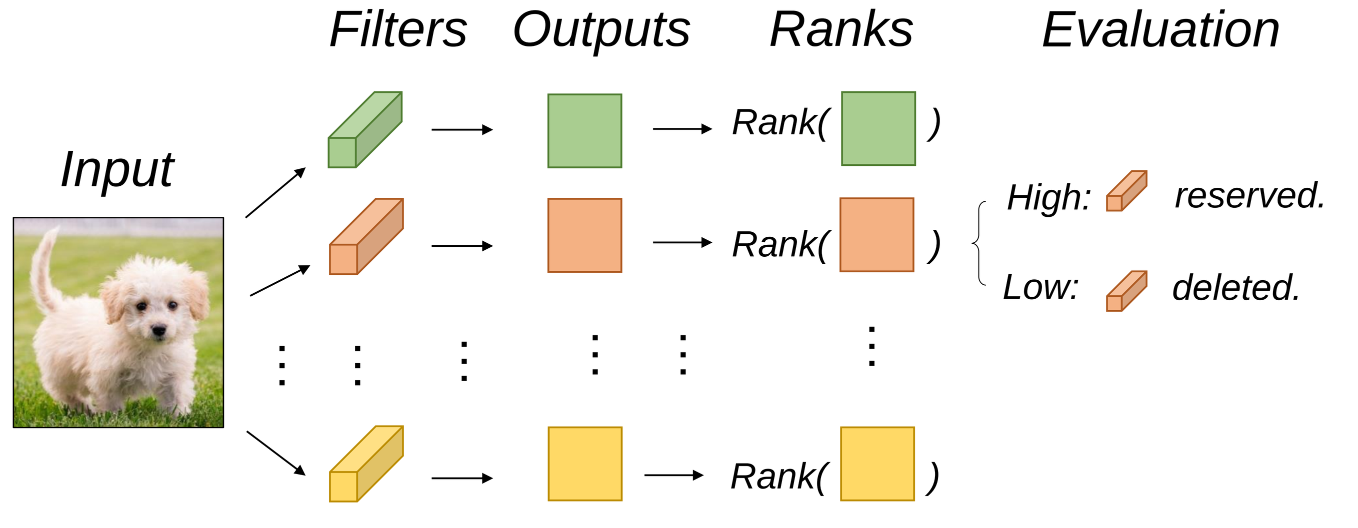


Figure 1. A toy example of HRank [1].

- Filters with **high rank** corresponding output slices will be regarded as **important ones**.
- Filters with **low rank** corresponding output slices will be treated as **trivial ones**.

1. 2 Convolution in the Frequency Domain

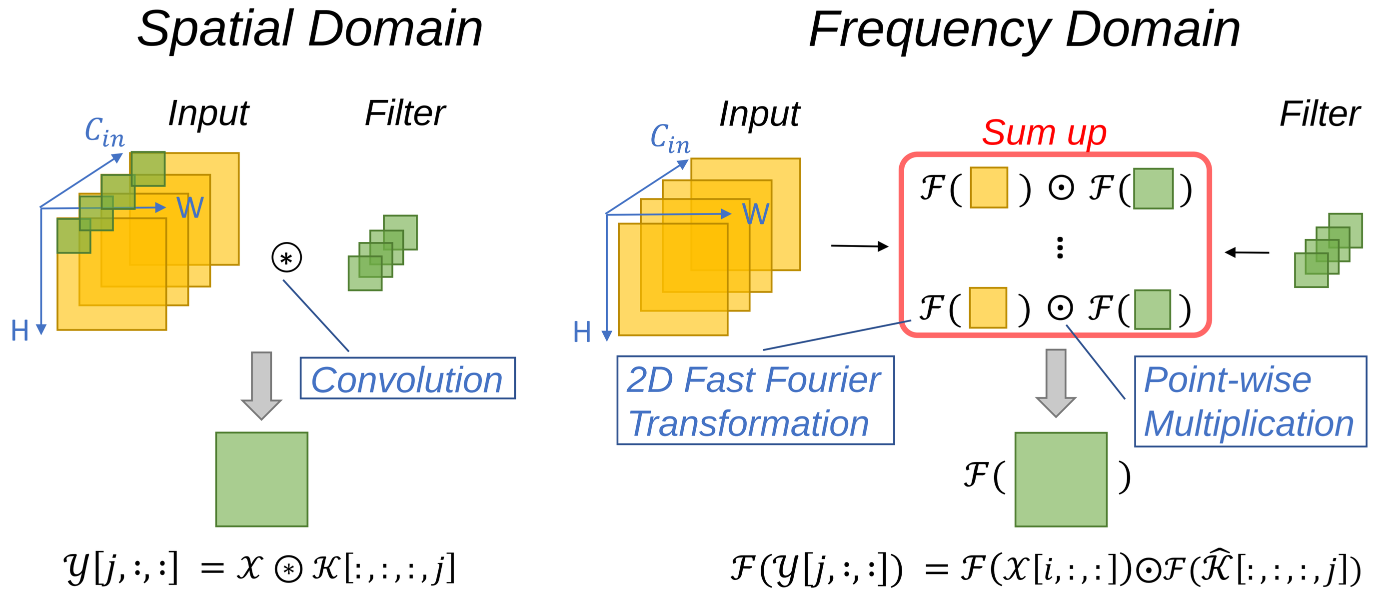


Figure 2. (Left) Convolution in the spatial domain. (Right) Convolution in the frequency domain.

For convolution in the **frequency domain**:

- First, each slice of the input and filter will be **mappe**d into the frequency domain by the **2D fast Fourier transformation.**
- Next, the slices at the same position along the channel axis will do **point-wise multiplication**.
- Finally, all the point-wise multiplication results will be **added up.**

1. 3 Matrix Ranks from the Frequency Domain Viewpoint

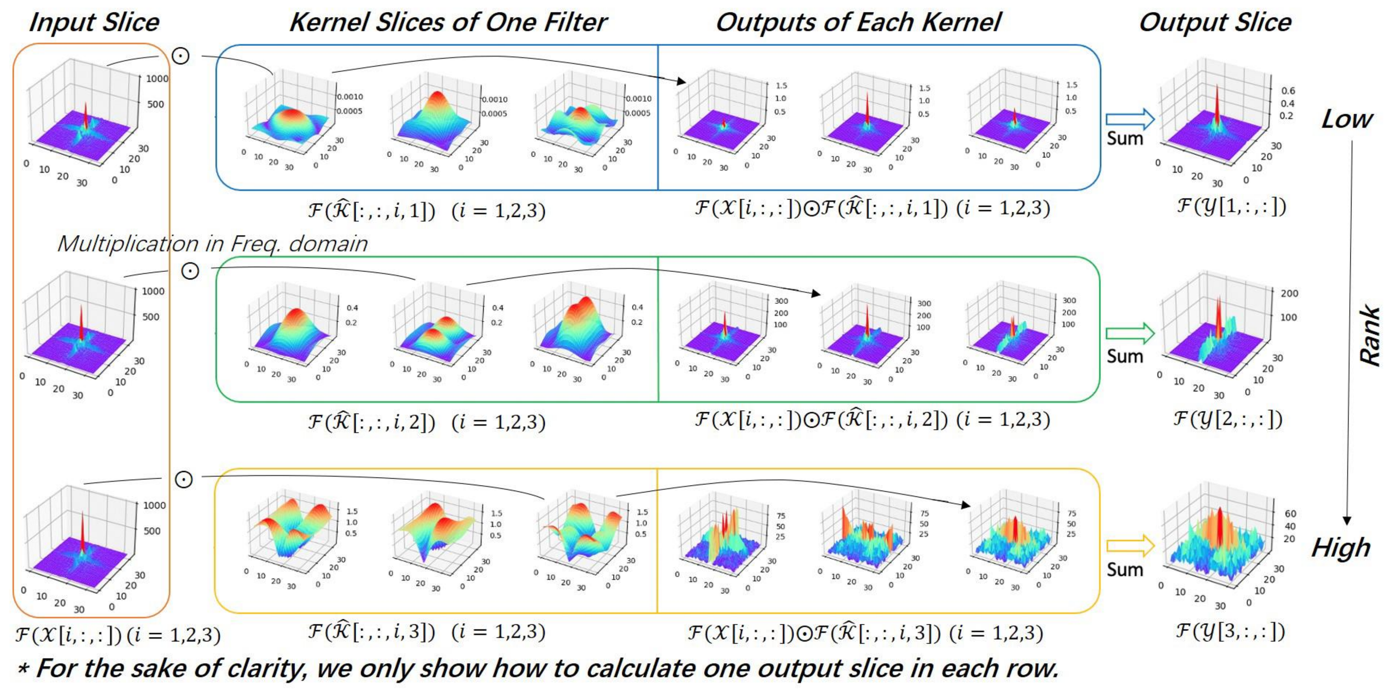


Figure 3. Input slices, kernels and output slices in the frequency domain.

- For output slice with **high rank**, the distribution of the **low-frequency** components is **dispersed**.
- For output slice with **low rank**, the distribution of the **low-frequency** components is **concentrated**.

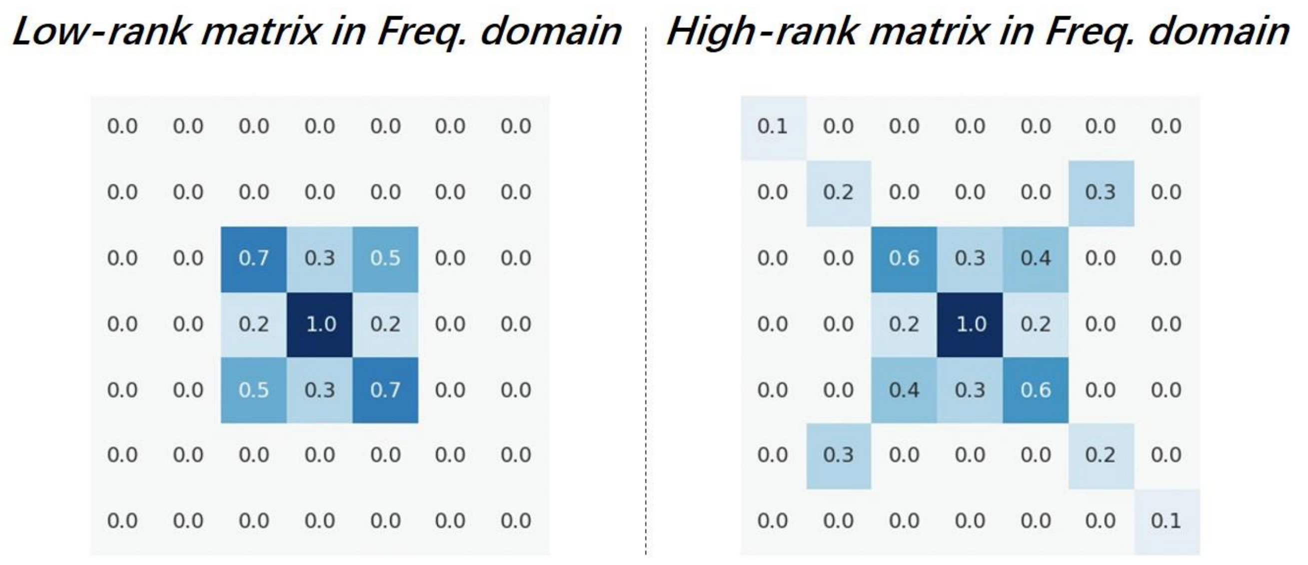
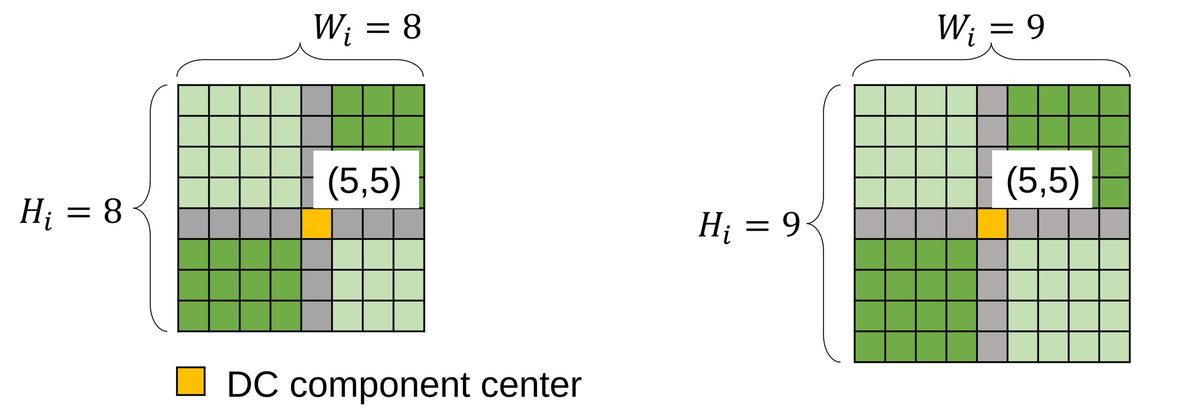


Figure 4. A toy example to conceptually depict a low-rank channel matrix and a high-rank one. We use zero-valued and nonzero-valued elements to represent high-frequency and low-frequency components, respectively.

- The left matrix is only of rank 3 while the right is full-rank. The spectral ranks also translate to the spatial ranks due to **rank-invariant domain transforms**.
- The right matrix with a **high rank** has more **dispersed nonzero elements.**

2. EZCrop

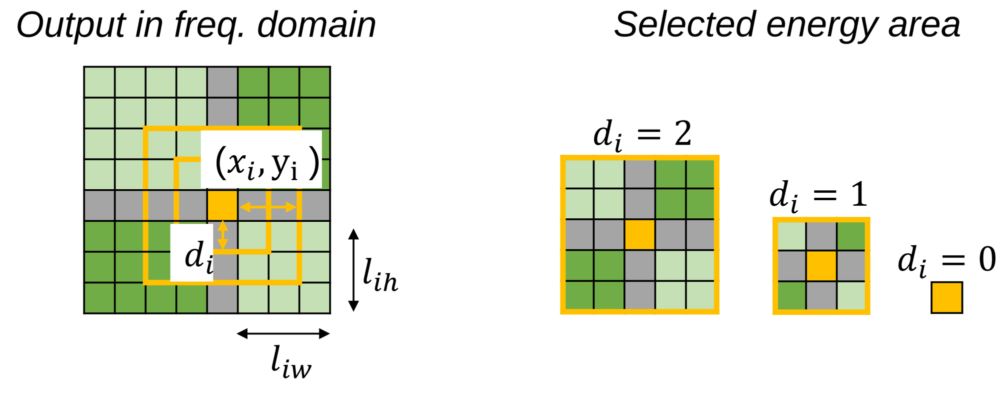
Step 1: Find the Square Center



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

Figure 5. The first step of EZCrop is to find the DC component center. (Left) when the height and width are even numbers. (Right) when the height and width are odd numbers.

Step 2: Decide the Expanding Distance



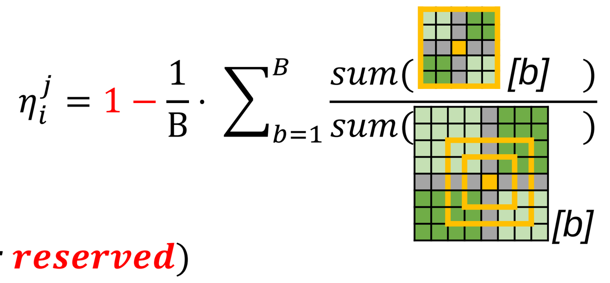
$$l_{ih} = H_i - x_i \qquad l_{iw} = W_i - y_i$$

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

Figure 5. The second step of EZCrop is to decide the size of the energy area. (Left) find the distance between the DC component center and the last row and column. (Right) if l_{ih} or l_{iw} is too small, the distance is set to be zero, else we use a hyperparameters to decide it.

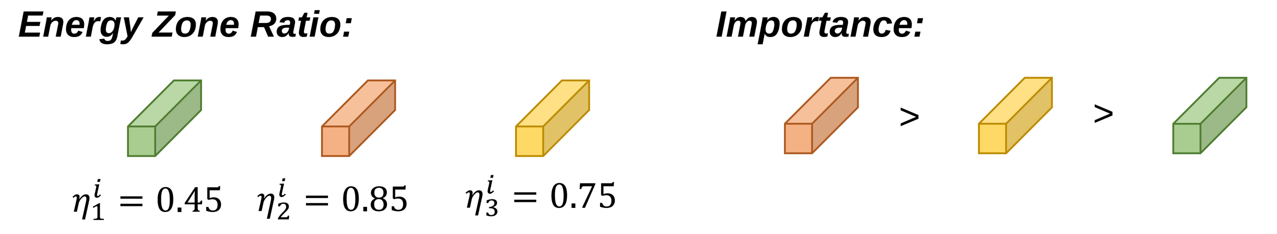
Step 3: Calculate the Energy Zone Ration

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



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

Step 4: Sort the Filters



3. Selected Experimental Results

3.1 Time Comparison

Dataset	Model	HRank [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%)

3.1 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

Acknowledgement

This work is supported in part by the General Research Fund (GRF) projects 17209721 & 17206020, and in part by the in-kind support of United Microelectronics Centre (Hong Kong) Limited.

Main References

Lin, M., T et al. (2020). "Hrank: Filter pruning using high-rank feature map". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1529-1538).

Victor Podlozhnyuk (2007). "Fft-based 2d convolution". NVIDIA white paper, 32.