Learning in the Frequency Domain

Kai Xu[1,2*] Minghai Qin[1] Fei Sun[1]
Yuhao Wang[1] Yen-Kuang Chen[1] Fengbo Ren[2]
[1]DAMO Academy, Alibaba Group [2]Arizona State University

CVPR 2020

Jiajin Zhang
06/29/2021

Motivation

For current neural networks:
To meet the network/GPU requirements → image downsizing in the space domain
→ Inevitably incurs information loss and accuracy degradation

A universal replacement for different networks/tasks:
Reshape/compressing the high-resolution images in the frequency domain

Method

1. Data pre-processing

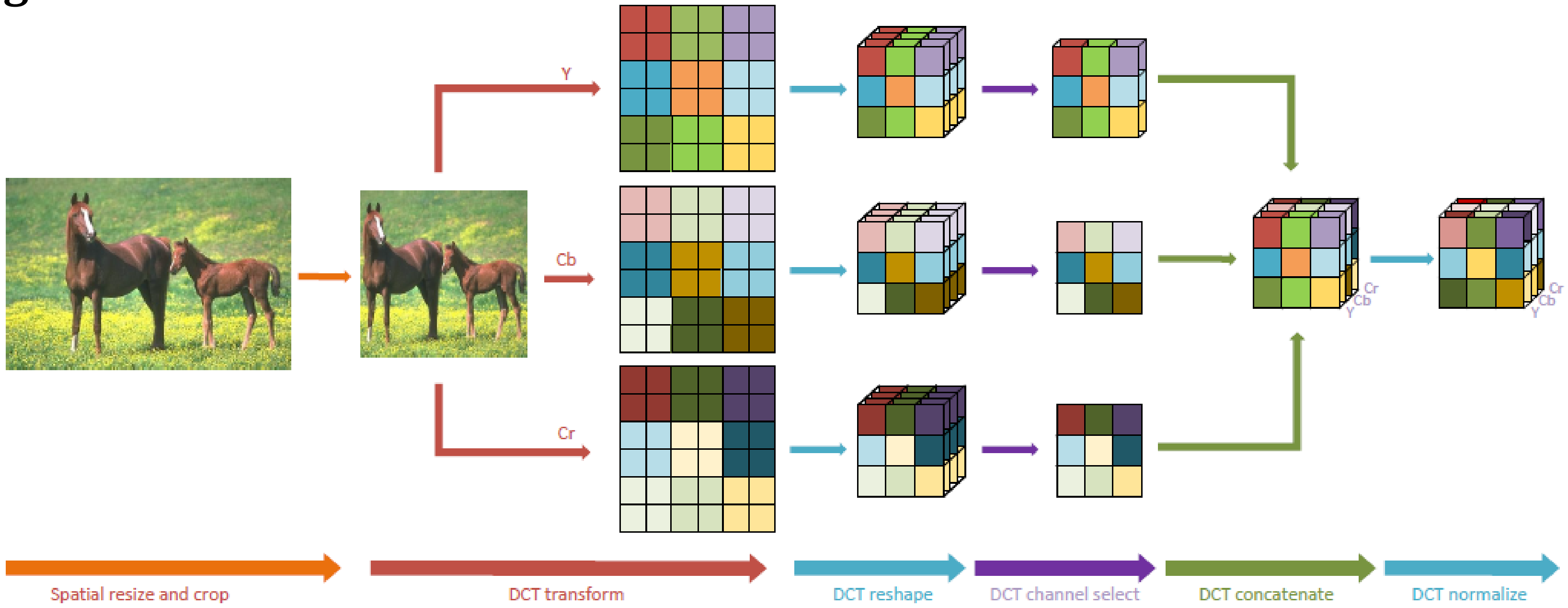


Figure 2: The data pre-processing pipeline for learning in the frequency domain.

1. Pre-processing and augmentation flow in the spatial domain
2. Augmented image → YCbCr color space → DCT transform + reshape
   2D DCT coefficients at the same frequency are grouped into one channel to form 2D DCT cubes
3. Channel selection and re-concatenation
4. Normalization

$$F(u,v) = \left(\frac{2}{N}\right)^{\frac{1}{2}} \left(\frac{2}{M}\right)^{\frac{1}{2}} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \Lambda(i).\Lambda(j).cos\left[\frac{\pi.u}{2.N}(2i+1)\right] cos\left[\frac{\pi.v}{2.M}(2j+1)\right].f(i,j)$$

Method

2. Learning-based Frequency Channel Selection

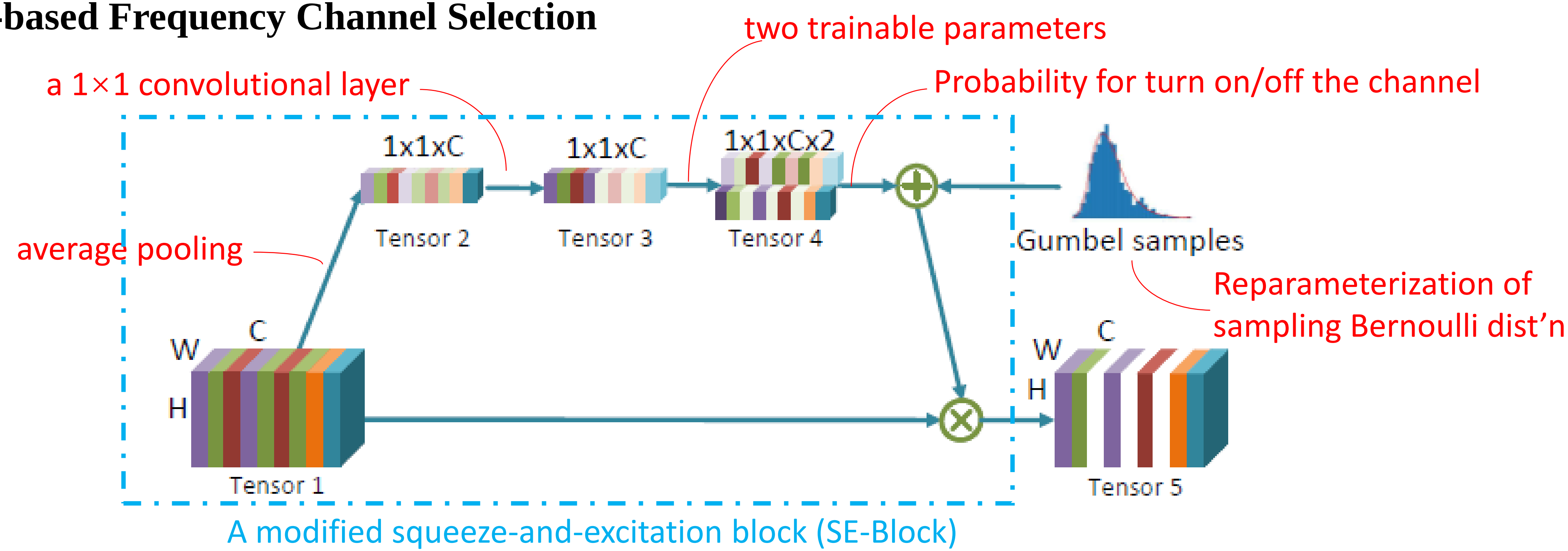
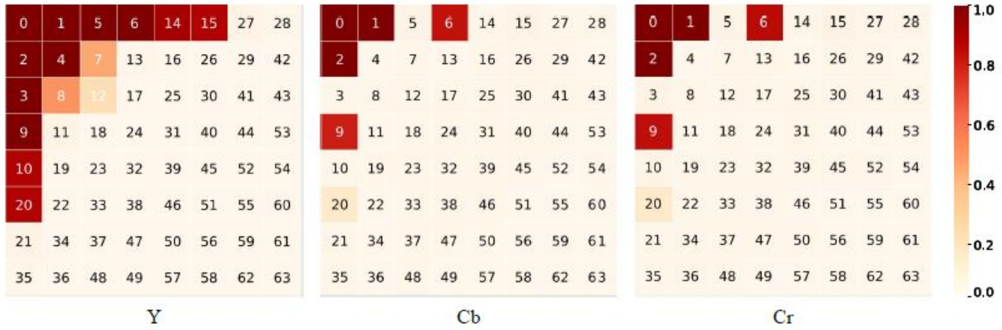


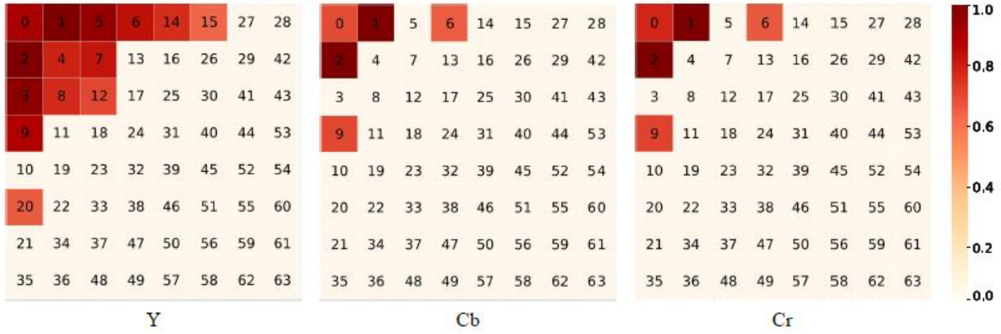
Figure 4: The gate module that generates the binary decisions based on the features extracted by the SE-Block. The white color channels of Tensor 5 indicate the unselected channels.

Method

3. Static Frequency Channel Selection



(a) Heat maps of Y, Cb, and Cr components on the ImageNet validation dataset.



(b) Heat maps of Y, Cb, and Cr components on the COCO validation dataset

Figure 5: A heat map visualization of input frequency channels on the ImageNet validation dataset for image c and COCO validation dataset for instance segmentation. The numbers in each square represent the correspond indices. The color from bright to dark indicates the possibility of a channel being selected from low to high.

- The low-frequency channels (boxes with small indices) are selected much more often than the high-frequency channels (boxes with with large indices). This demonstrates that low-frequency channels are more informative than high-frequency channels in general for vision inference tasks.
- The frequency channels in luma component Y are selected more often than the channels in chroma components Cb and Cr. This indicates that the luma component is more informative for vision inference tasks.
- The heat maps share a common pattern between the classification and segmentation tasks. This indicates that the above-mentioned two observations are not specific to one task and is very likely to be general to more high-level vision tasks.
- Interestingly, some lower frequency channels have lower probability of being selected than the slightly higher frequency channels. For example, in Cb and Cr components, both tasks favor Channel 6 and 9 over Channel 5 and 3.

Those observations imply that the CNN models may indeed exhibit similar characteristics to the HVS, and the image compression standards (*e.g.*, JPEG) targeting human eyes may be suitable for the CNN models as well.

Method

Overview of the proposed pipeline

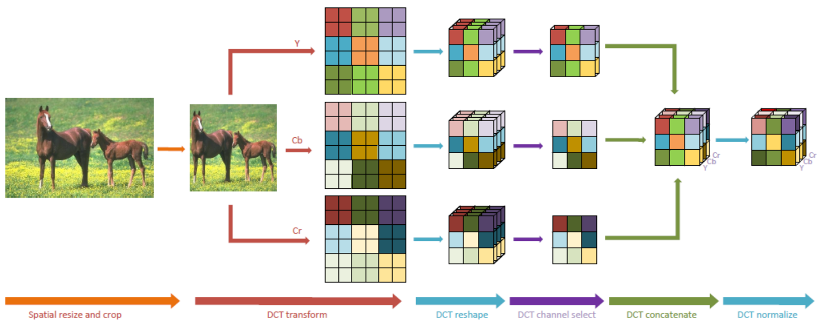


Figure 2: The data pre-processing pipeline for learning in the frequency domain.

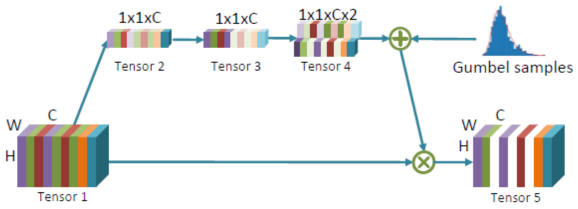


Figure 4: The gate module that generates the binary decisions based on the features extracted by the SE-Block. The white color channels of Tensor 5 indicate the unselected channels.

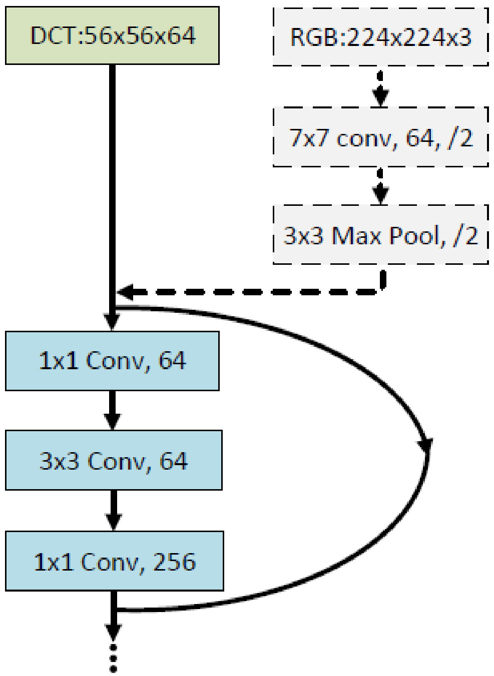


Figure 3: Connecting the pre-processed input features in the frequency domain to ResNet-50. The three input layers (the dashed gray blocks) in a vanilla ResNet-50 are removed to admit the 56×56×64 DCT inputs. We take 64 channels as an example. This value can vary based on the channel selection. In learning-based channel selection, all 192 channels are analyzed for their importance to accuracy, based on which only a subset ($\ll$ 192 channels) is used in the static selection approach.

for each frequency channel x_i. Then x_i is selected if

$$\mathbf{F}(x_i) \neq 0, \text{ i.e., } \mathbf{F}(x_i) \odot x_i \neq \mathbf{0}, \tag{1}$$

where $\odot$ is the element-wise product.

$$\mathcal{L} = \mathcal{L}_{Acc} + \lambda \cdot \sum_{i=1}^{C} \mathbf{F}(x_i), \tag{2}$$

Experiments & Results
Image Classification

Table 1: ResNet-50 classification results on ImageNet (validation). The input size of each method is normalized over the baseline ResNet-50. The input frequency channels are selected with the square and triangle channel selection pattern if the postfix S and T is specified, respectively.

ResNet-50	#Channels	Size Per Channel	Top-1	Top-5	Normalized Input Size
RGB	3	224×224	75.780	92.650	1.0
YCbCr	3	224×224	75.234	92.544	1.0
DCT-192 [17]	192	28×28	76.060	93.020	1.0
DCT-192 (ours)	192	56×56	77.194	93.454	4.0
DCT-24D (ours)	24	56×56	77.166	93.560	0.5
DCT-24S (ours)	24	56×56	77.196	93.504	0.5
DCT-24T (ours)	24	56×56	77.148	93.326	0.5
DCT-48S (ours)	48	56×56	77.384	93.554	1.0
DCT-48T (ours)	48	56×56	77.338	93.614	1.0
DCT-64S (ours)	64	56×56	77.232	93.624	1.3
DCT-64T (ours)	64	56×56	77.280	93.456	1.3

Table 2: MobileNetV2 classification results on ImageNet (validation).

MobileNetV2	#Channels	Size Per Channel	Top-1	Top-5	Normalized Input Size
RGB	3	224×224	71.702	90.415	1.0
DCT-6S (ours)	6	112×112	71.776	90.258	0.5
DCT-12S (ours)	12	112×112	72.156	90.634	1.0
DCT-24S (ours)	24	112×112	72.364	90.606	2.0
DCT-32S (ours)	32	112×112	72.282	90.592	2.7

Note that DCT-12S and DCT-6S select 12 and 6 frequency channels

Experiments & Results

Instance Segmentation

Table 3: Bbox AP results of Mask R-CNN using different backbones on COCO 2017 validation set. The baseline Mask R-CNN uses a ResNet-50-FPN as the backbone. The DCT method uses the frequency-domain ResNet-50-FPN as the backbone.

Backbone	#Channels	Size Per Channel	bbox					
			AP	AP@0.5	AP@0.75	AP_S	AP_M	AP_L
ResNet-50-FPN (RGB)	3	800×1333	37.3	59.0	40.2	21.9	40.9	48.1
DCT-24S (ours)	24	200×334	37.7	59.2	40.9	21.7	41.4	49.1
DCT-48S (ours)	48	200×334	38.1	59.5	41.2	22.0	41.3	49.8
DCT-64S (ours)	64	200×334	38.1	59.6	41.1	22.5	41.6	49.7

Table 4: Mask AP results of Mask R-CNN using different backbones on COCO 2017 validation set.

Backbone	#Channels	Size Per Channel	mask					
			AP	AP@0.5	AP@0.75	AP_S	AP_M	AP_L
ResNet-50-FPN (RGB)	3	800×1333	34.2	55.9	36.2	15.8	36.9	50.1
DCT-24S (ours)	24	200×334	34.6	56.1	36.9	16.1	37.4	50.7
DCT-48S (ours)	48	200×334	35.0	56.6	37.2	16.3	37.5	52.3
DCT-64S (ours)	64	200×334	35.0	56.5	37.4	16.9	37.6	51.6

evaluated. For the mask AP, we also report AP@0.5 and AP@0.75 at the IoU threshold of 0.5 and 0.75 respectively, as well as AP_S, AP_M, and AP_L at different scales.

Thanks！