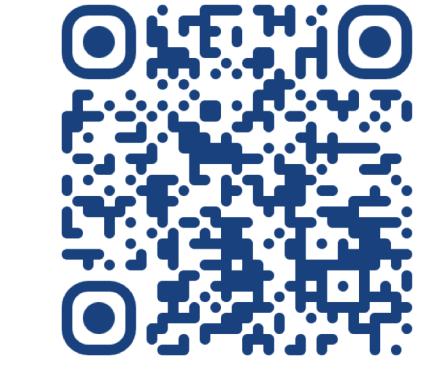


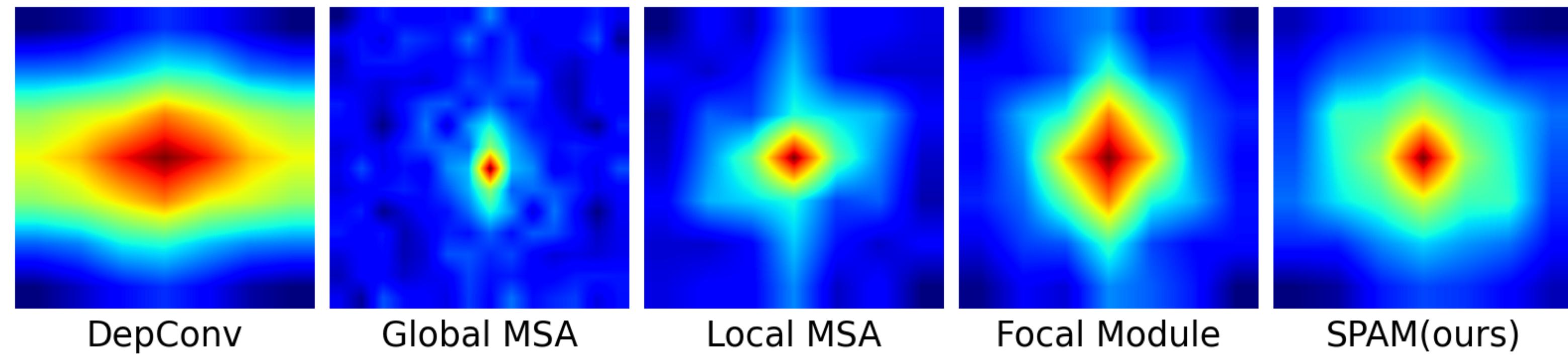
# SPANet: Frequency-balancing Token Mixer using Spectral Pooling Aggregation Modulation



Guhnoo Yun<sup>1,2</sup> Juhan Yoo<sup>3</sup> Kijung Kim<sup>1,2</sup> Jeongho Lee<sup>1,2</sup> Dong Hwan Kim<sup>1,2</sup>  
<sup>1</sup>Korea Institute of Science and Technology <sup>2</sup>Korea University <sup>3</sup>Semyung University

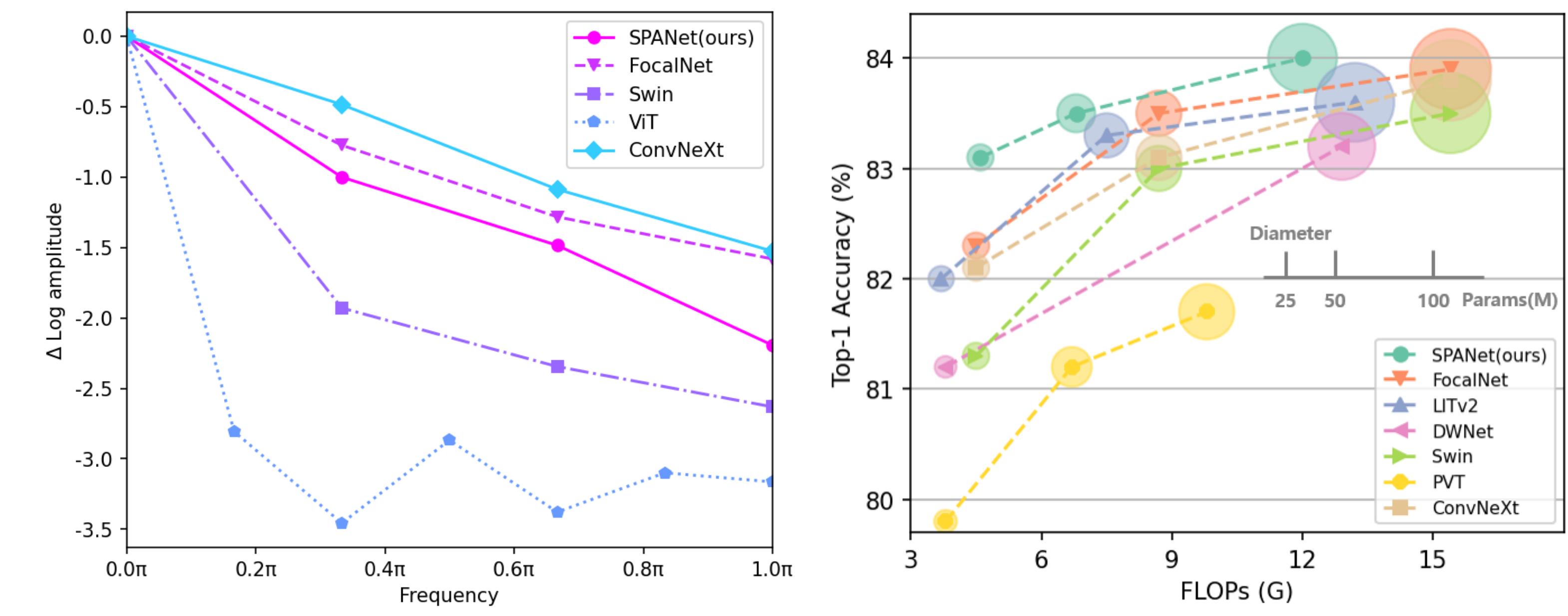
## Introduction

- Background:** Self-attention acts like a low-pass filter (unlike convolution), and the enhanced high-pass filtering capabilities help improve model performance.
- Observation:** Better low-pass filtering in convolution operations also can improve performance.
- Hypothesis:** An ideal token mixer, optimizing the balance of high- and low-frequency features, can enhance model performance.



## Contribution

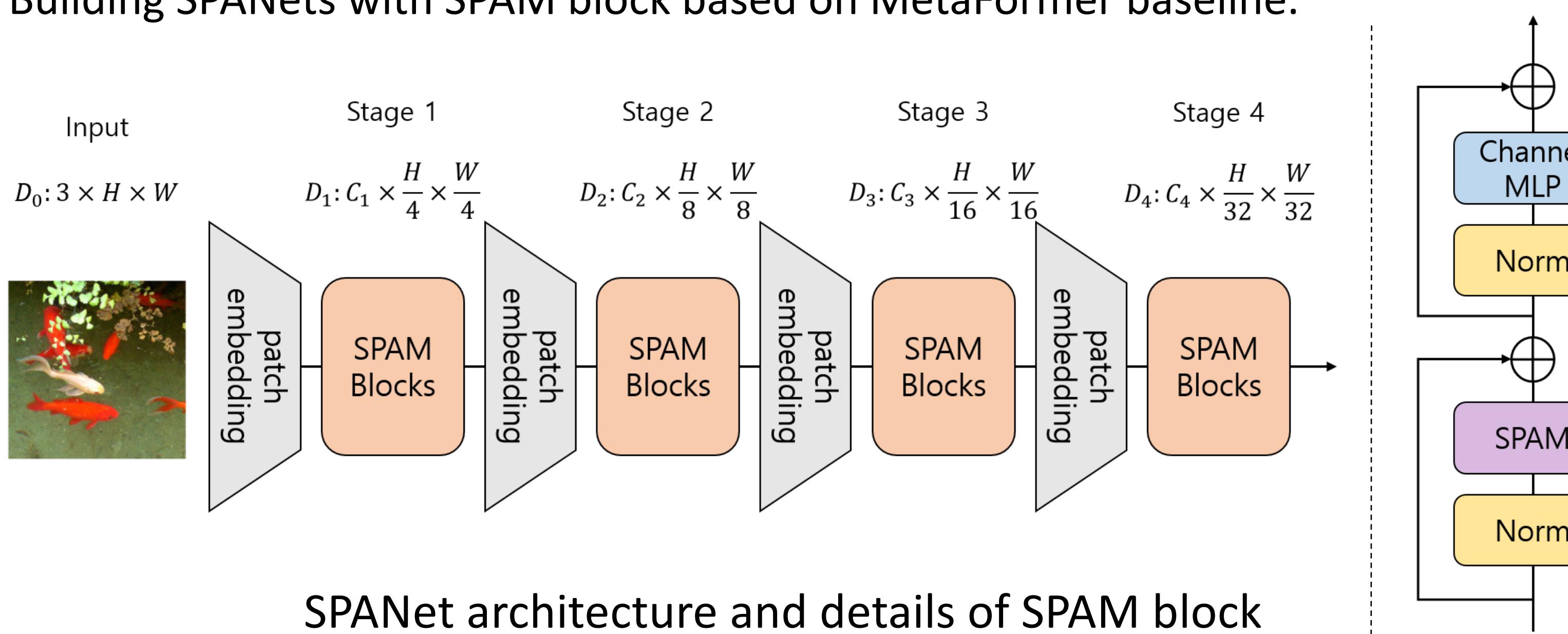
- Performance improvements by **adjusting spectral filtering capabilities** of token mixers.
- SPAM optimizes **high/low-frequency component balance**.
- SoTA on image classification and semantic segmentation, and competitive results for object detection and instance segmentation.



## Methodology

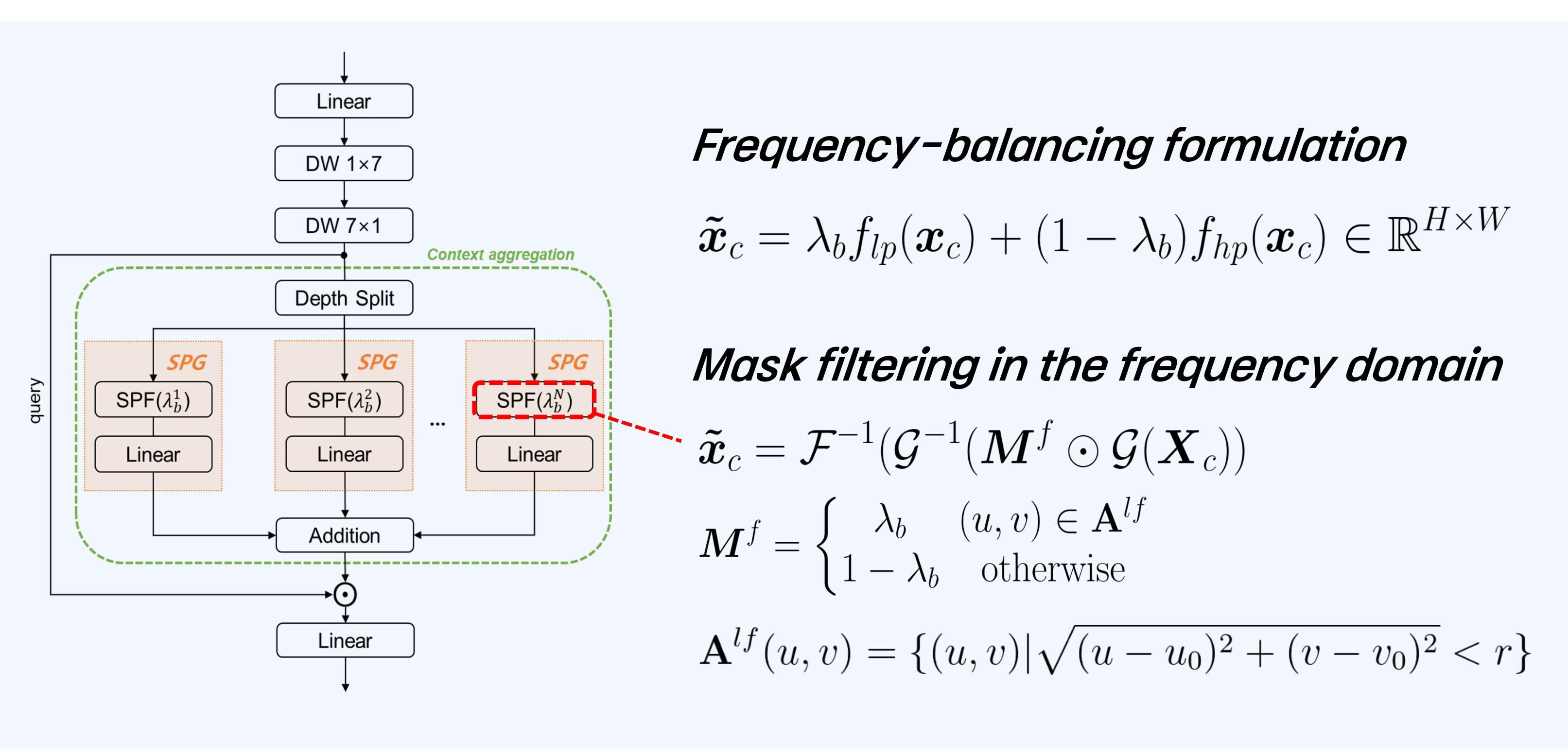
### 1. SPANet

- A new token mixer called **SPAM** (spectral pooling aggregation modulation) can balance high/low-frequency components of visual features.
- Building SPANets with SPAM block based on MetaFormer baseline.



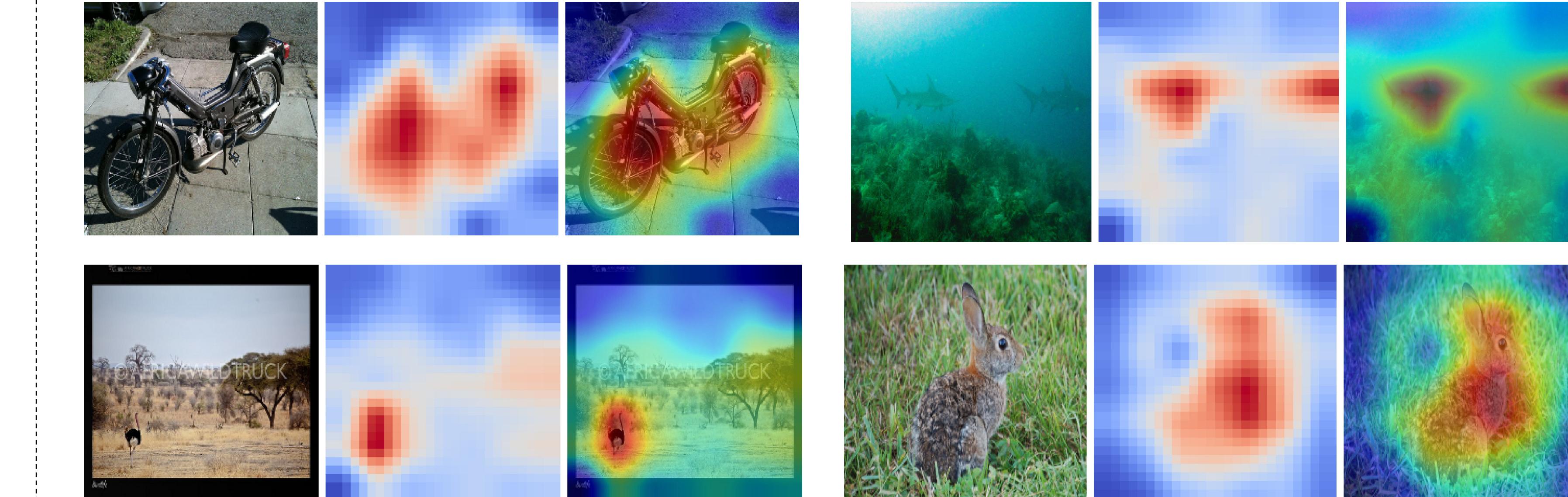
### 2. SPAM, The Token Mixer

- The frequency-balancing formula is changed into a mask-filtering problem in the frequency domain. This mask filter is defined as **spectral pooling filter (SPF)**.
- Context is aggregated with multiple spectral pooling gates (SPGs).
- Query and context are modulated by the Hadamard product.



## Experiments

### Aggregated Context



### Image Classification on ImageNet-1K

Model	General Arch.	Token Mixer	Params (M)	FLOPs (G)	Top-1 (%)
RSB-ResNet-50 [23, 66]	CNN	-	26	4.1	79.8
ConvNeXt-T [35]			29	4.5	82.1
PoolFormer-S24 [72]	Pooling		21	3.4	80.3
PVT-Small [63]	Attention		25	3.8	79.8
Swin-T [34]			29	4.5	81.3
LITv2-S [42]			28	3.7	82.0
GFNet-H-S [46]			32	4.6	81.5
DWNet-Hiny [21]			24	3.8	81.2
FocalNet-T [69]	Convolution		29	4.5	82.3
SPANet-S (ours)		SPAM	29	4.6	<b>83.1</b>
	Norm				
RSB-ResNet-101 [23, 66]	CNN	-	45	7.9	81.3
ConvNeXt-S [35]			50	8.7	83.1
PoolFormer-M36 [72]	Pooling		56	8.8	82.1
PVT-Medium [63]	Attention		44	6.7	81.2
Swin-S [34]			50	8.7	83.0
LITv2-M [42]			49	7.5	83.3
GFNet-H-B [46]	Convolution		54	8.6	82.9
FocalNet-S [69]			50	8.7	83.5
SPANet-M (ours)		SPAM	42	8.8	<b>83.5</b>
	Norm				
RSB-ResNet-152 [23, 66]	CNN	-	60	11.6	81.8
ConvNeXt-B [35]			89	15.4	83.8
PoolFormer-M48 [72]	Pooling		73	11.6	82.5
PVT-B16 [15]	Attention		86	17.6	79.7
Swin-B [34]			61	9.8	81.7
LITv2-B [42]			88	15.4	83.5
GFNet-B [46]	Convolution		87	13.2	83.6
FocalNet-B [69]			74	12.9	83.2
SPANet-B (ours)		SPAM	89	15.4	83.9
	Norm				
			76	12.0	<b>84.0</b>

### Semantic Segmentation on ADE20K

Backbone	Params (M)	FLOPs (G)	mIoU(%)
ResNet50 [23]	29	46	36.7
PVT-Small [63]	28	45	39.8
Swin-T [34]	32	46	41.5
LITv2-S [42]	31	41	44.3
SPANet-S (ours)	32	46	<b>45.4</b>
ResNet101 [23]	48	65	38.8
PVT-Medium [63]	48	61	41.6
Swin-S [34]	53	70	45.2
LITv2-M [42]	52	63	45.7
SPANet-M (ours)	45	57	<b>46.2</b>

### Object Detection / Instance Segmentation on COCO

Backbone	RetinaNet 1×						Mask R-CNN 1×							
	Param (M)	AP	AP <sub>50</sub>	AP <sub>75</sub>	AP <sub>S</sub>	AP <sub>M</sub>	AP <sub>L</sub>	Param (M)	AP <sup>b</sup>	AP <sub>50</sub> <sup>b</sup>	AP <sub>75</sub> <sup>b</sup>	AP <sup>m</sup>	AP <sub>50</sub> <sup>m</sup>	AP <sub>75</sub> <sup>m</sup>
ResNet50 [23]	38	36.3	55.3	38.6	19.3	40.0	48.8	44	38.0	58.6	41.4	34.4	55.1	36.7
PVT-Small [63]	34	40.4	61.3	43.0	25.0	42.9	55.7	44	40.4	62.9	43.8	37.8	60.1	40.3
Swin-T [34]	39	41.5	62.1	44.2	25.1	44.9	55.5	48	42.2	64.6	46.2	39.1	61.6	42.0
LITv2-S [42]	38	43.7	-	-	-	-	-	47	44.7	-	-	40.7	-	-
SPANet-S (ours)	38	43.3	<b>63.7</b>	<b>46.5</b>	<b>25.8</b>	<b>47.7</b>	<b>57.0</b>	48	<b>44.7</b>	<b>65.7</b>	<b>48.8</b>	40.6	<b>62.9</b>	<b>43.8</b>
ResNet101 [23]	57	38.5	57.8	41.2	21.4	42.6	51.1	63	40.4	61.1	44.2	36.4	57.7	38.8
PVT-Medium [63]	54	41.9	63.1	44.3	25.0	44.9	57.6	64	42.0	64.4	45.6	39.0	61.6	42.1
Swin-S [34]	60	44.5	<b>65.7</b>	<b>47.5</b>	<b>27.4</b>	<b>48.0</b>	<b>59.9</b>	69	44.8	<b>66.6</b>	48.9	40.9	63.4	<b>44.2</b>
LITv2-M [42]	59	<b>45.8</b>	-	-	-	-	-	68	<b>46.5</b>	-	-	<b>42.0</b>	-	-
SPANet-M (ours)	51	44.0	64.3	47.0	25.9	<b>48.0</b>	58.7	61	45.2	66.3	<b>49.6</b>	41.0	<b>63.5</b>	44.0