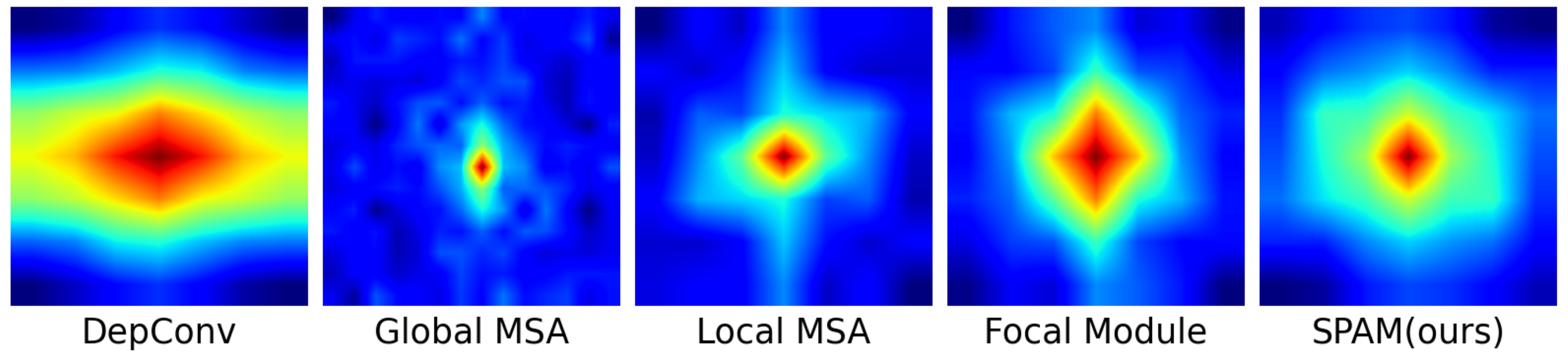
SPANet: Frequency-balancing Token Mixer using Spectral Pooling Aggregation Modulation

ICCV23 PARIS

Guhnoo Yun[1,2] Juhan Yoo[3] Kijung Kim[1,2] Jeongho Lee[1,2] Dong Hwan Kim[1,2]
[1]Korea Institute of Science and Technology [2]Korea University [3]Semyung University

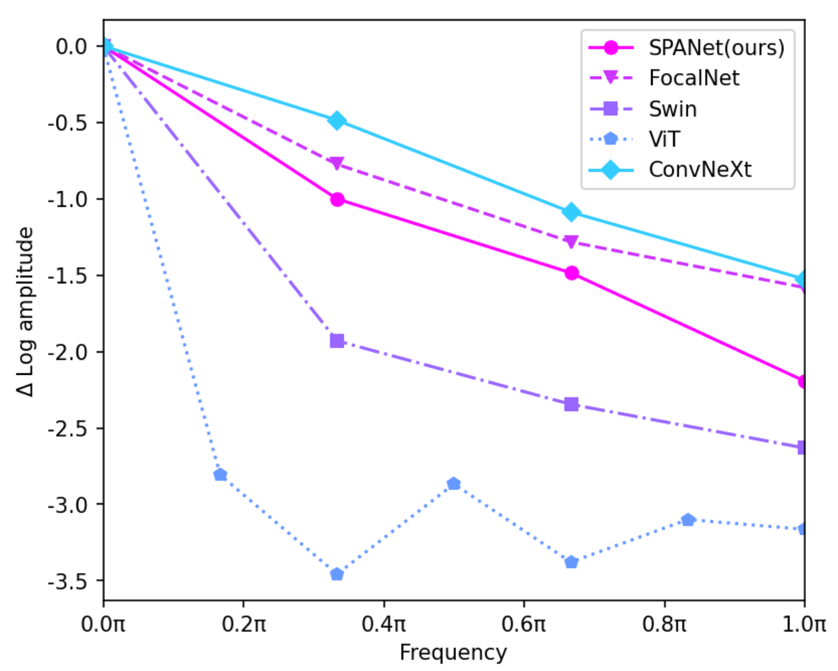
Introduction

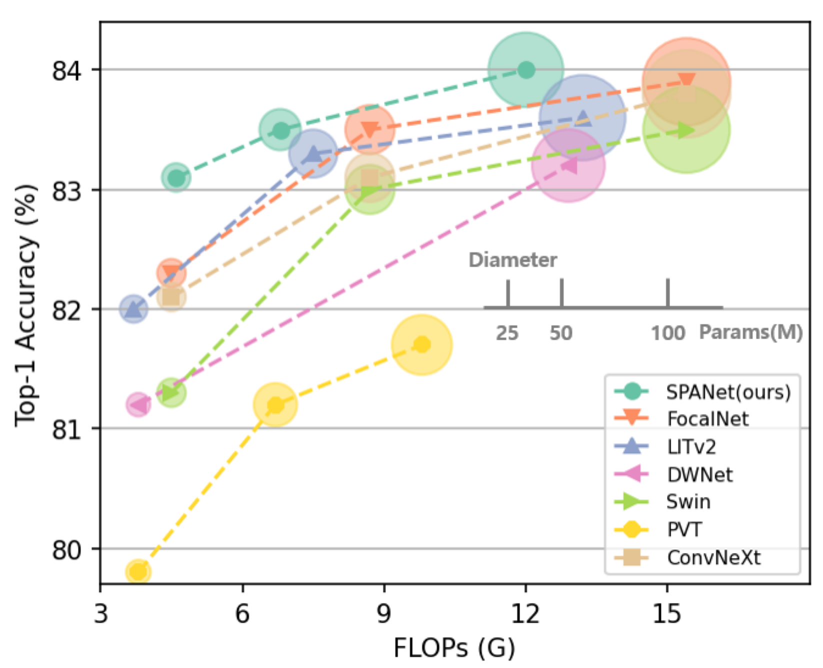
- **Background**: Self-attention acts like a low-pass filter (unlike convolution), and the enhanced high-pass filtering capabilities help improve model performance.
- **Observation**: Better low-pass filtering in convolution operations also can improve performance.
- **Hypothesis**: An ideal token mixer,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

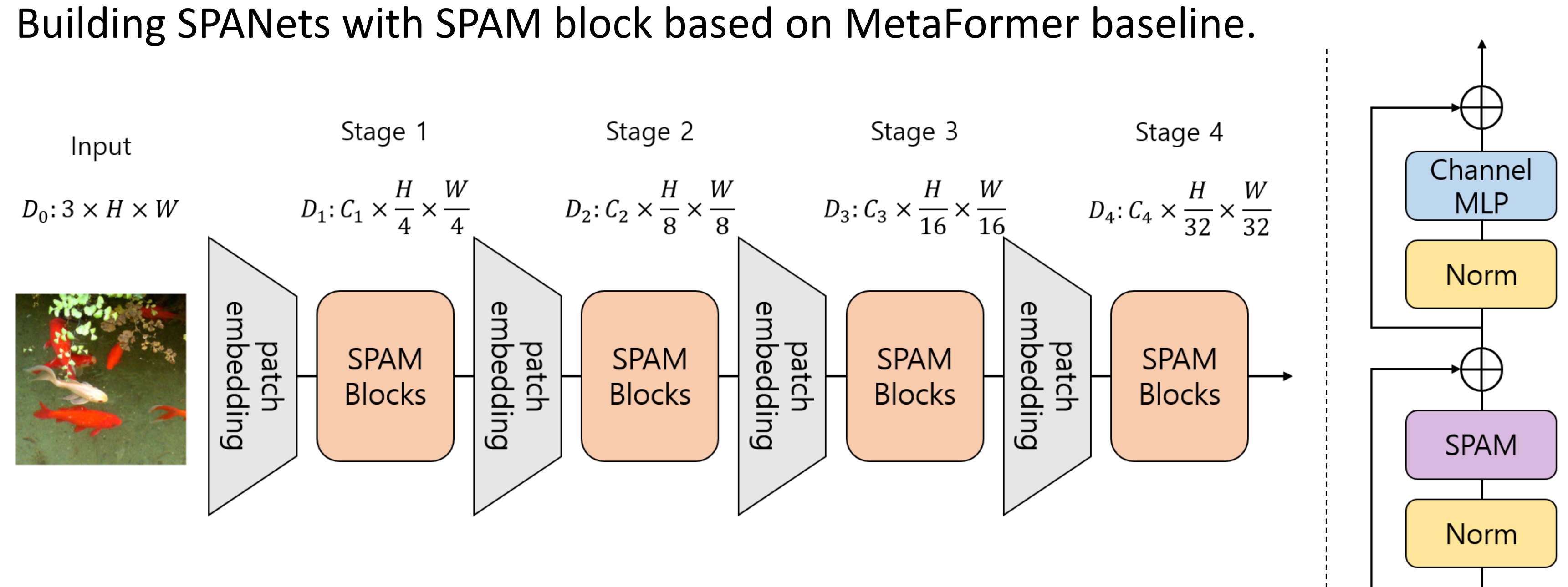




Methodology

1. SPANet

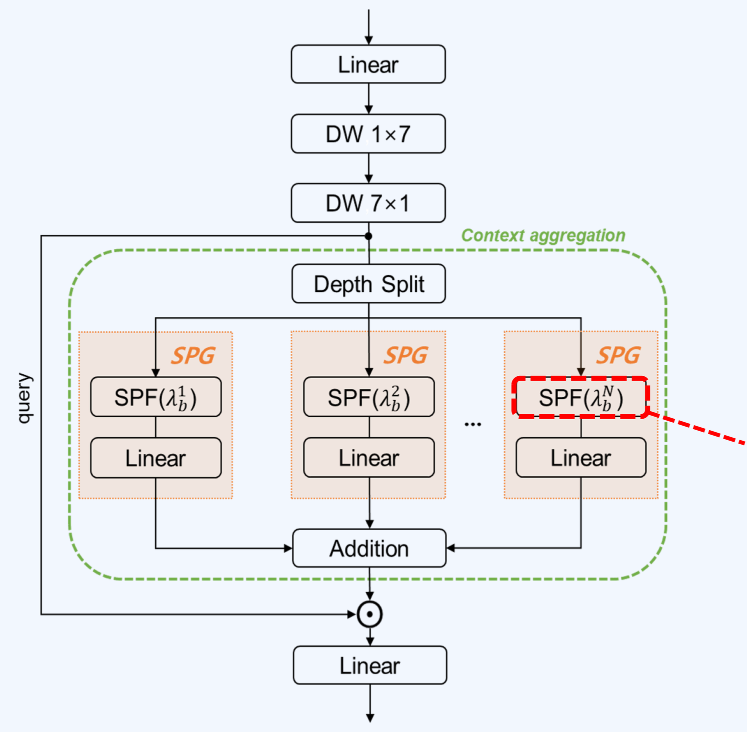
- A new token mixer called **SPAM** (spectral pooling aggregation modulation) can balance high/low-frequency components of visual features.
- Building SPANets with SPAM block based on MetaFormer baseline.



SPANet architecture and details of SPAM block

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.



Frequency-balancing formulation

$$\tilde{\boldsymbol{x}}_c = \lambda_b f_{lp}(\boldsymbol{x}_c) + (1-\lambda_b) f_{hp}(\boldsymbol{x}_c) \in \mathbb{R}^{H \times W}$$

Mask filtering in the frequency domain

$$\tilde{\boldsymbol{x}}_c = \mathcal{F}^{-1}(\mathcal{G}^{-1}(\boldsymbol{M}^f \odot \mathcal{G}(\boldsymbol{X}_c)))$$

$$\boldsymbol{M}^f = \begin{cases} \lambda_b & (u,v) \in \mathbf{A}^{lf} \\ 1-\lambda_b & \text{otherwise} \end{cases}$$

$$\mathbf{A}^{lf}(u,v) = \{(u,v) \mid \sqrt{(u-u_0)^2 + (v-v_0)^2} < r\}$$

Experiments

Aggregated Context

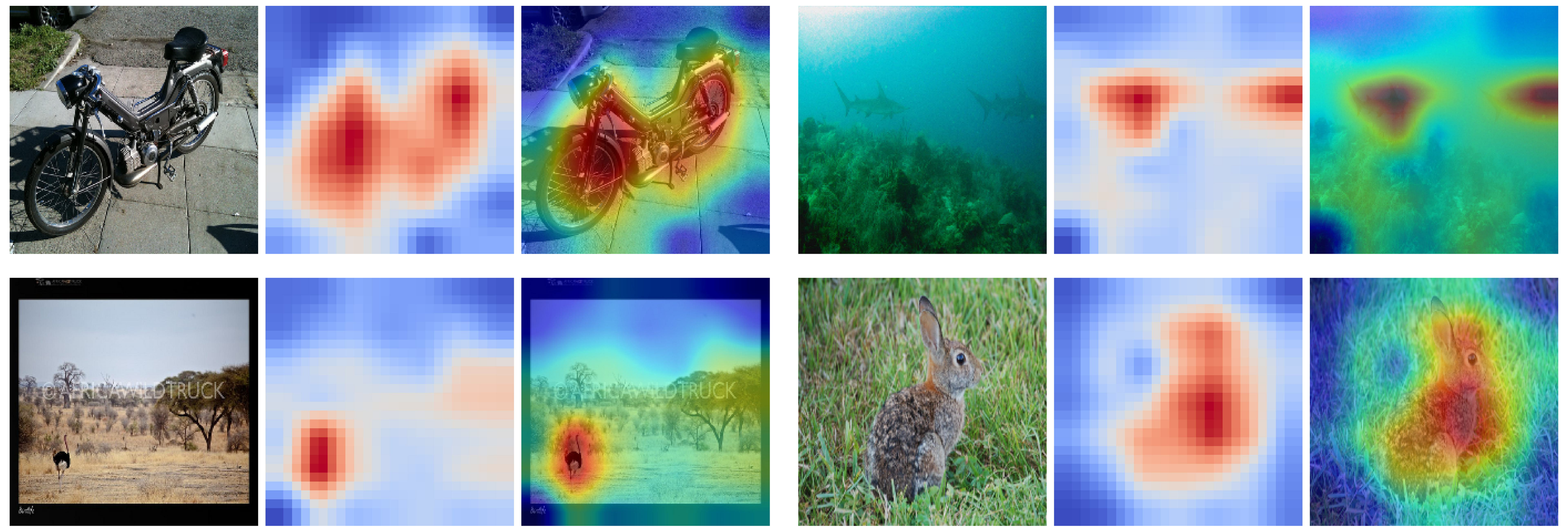


Image Classification on ImageNet-1K

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]	MetaFormer	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]		Convolution	32	4.6	81.5
DWNet-tiny [21]			24	3.8	81.2
FocalNet-T [69]			29	4.5	82.3
SPANet-S (ours)			29	4.6	**83.1**
RSB-ResNet-101 [23, 66]	CNN	-	45	7.9	81.3
ConvNeXt-S [35]			50	8.7	83.1
PoolFormer-M36 [72]	MetaFormer	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)			42	6.8	**83.5**
RSB-ResNet-152 [23, 66]	CNN	-	60	11.6	81.8
ConvNeXt-B [35]			89	15.4	83.8
PoolFormer-M48 [72]	MetaFormer	Pooling	73	11.6	82.5
ViT-B/16 [15]		Attention	86	17.6	79.7
PVT-Large [63]			61	9.8	81.7
Swin-B [34]			88	15.4	83.5
LITv2-B [42]			87	13.2	83.6
DWNet-base [21]		Convolution	74	12.9	83.2
FocalNet-B [69]			89	15.4	83.9
SPANet-B (ours)			76	12.0	**84.0**

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	**45.4**
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	**46.2**

Object Detection / Instance Segmentation on COCO

Backbone	RetinaNet 1×							Mask R-CNN 1×						
	Param (M)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Param (M)	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
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	**63.7**	**46.5**	**25.8**	**47.7**	**57.0**	48	**44.7**	**65.7**	**48.8**	40.6	**62.9**	**43.8**
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	**65.7**	**47.5**	**27.4**	**48.0**	**59.9**	69	44.8	**66.6**	48.9	40.9	63.4	**44.2**
LITv2-M [42]	59	**45.8**	-	-	-	-	-	68	**46.5**	-	-	**42.0**	-	-
SPANet-M (ours)	51	44.0	64.3	47.0	25.9	**48.0**	58.7	61	45.2	66.3	**49.6**	41.0	**63.5**	44.0