LSNET: SEE LARGE, FOCUS SMALL

Ao Wang[1], Hui Chen[2,†], Zijia Lin[1], Jungong Han[3], Guiguang Ding[1]
[1]School of Software, Tsinghua University [2]BNRist, Tsinghua University
[3]Department of Automation, Tsinghua University

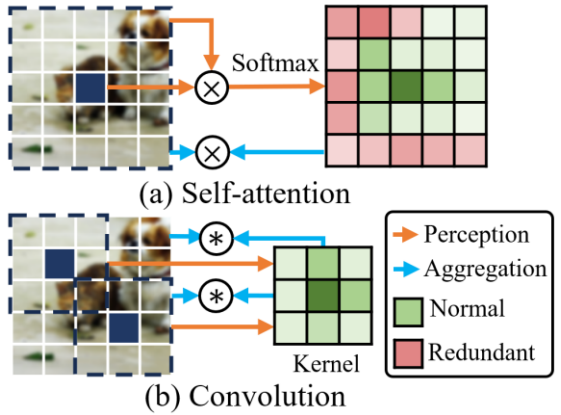
Background

Vision network architecture: CNN and ViT

- They have significantly advanced the field of computer vision.
- However, their notable inference cost makes real-time applications challenging.

Existing lightweight models

- They primarily rely on either self-attention or convolution for token mixing.
- Both of which have inherent limitations in perception and aggregation processes.
- Self-attention introduces unnecessary aggregation on irrelevant regions, with high computational demand.
- Convolution suffers from limited receptive field and static kernel weights, hindering adaptability to diverse contexts.



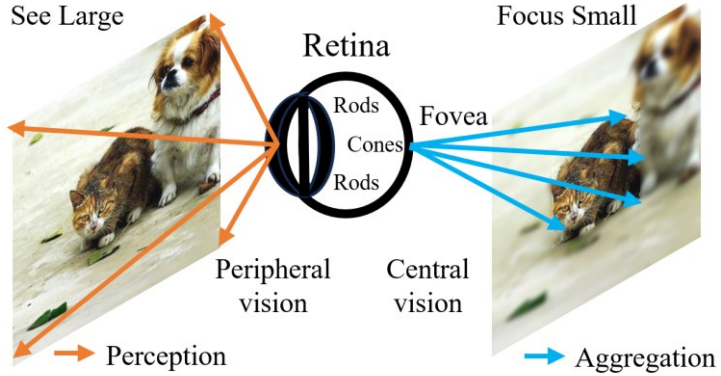
Background

Inspiration from human vision system

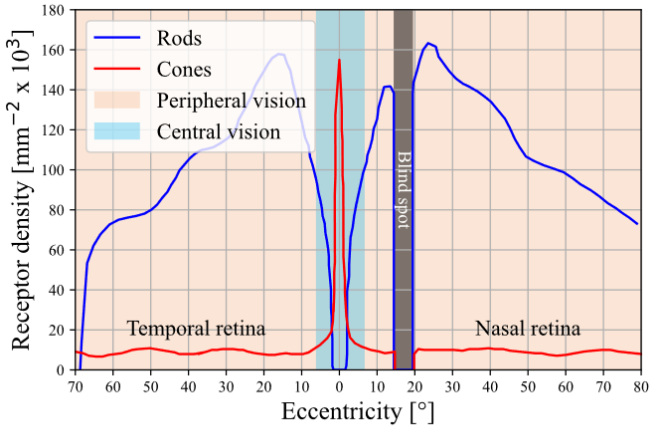
- The retina has two types of photoreceptor cells, rod and cone cells.
- Rod cells: capture a broad view of the scene but with limited detail.
- Cone cells: focus on fine and complex details in a small region.
- It perceives broad contexts with the peripheral vision system and detailed understanding using the central vision system.

Introducing “See Large, Focus Small” and LS convolution

- Large-kernel perception (LKP) well perceive the enlarged neighborhood information, leading to improved relationship modeling, like peripheral vision.
- Small-kernel aggregation (SKA) further adaptively fuse intricate visual features in small surroundings, enabling detailed visual understanding like central vision.



(c) Human vision



(d) Distribution of rods and cones in retina

Methodology

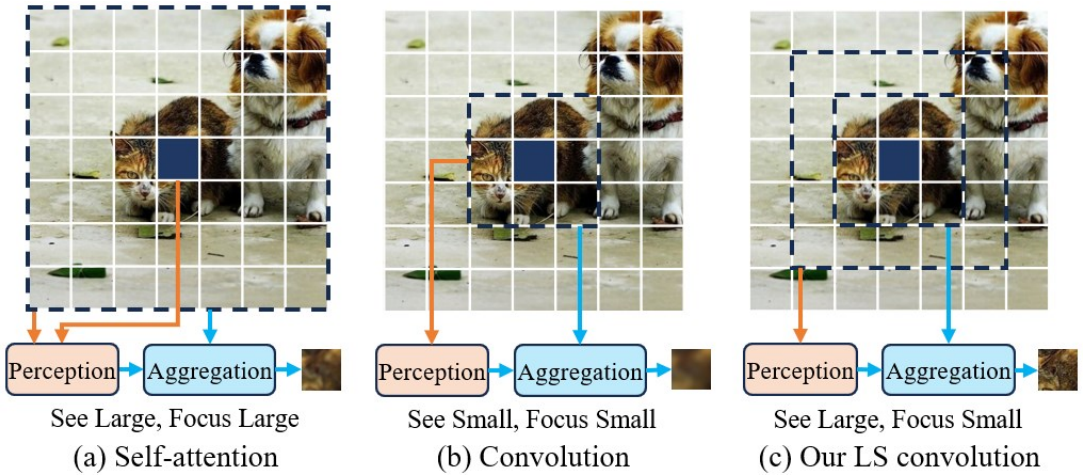
Revisiting Self-Attention and Convolution

- Token mixing ways: Generate feature representation $y_i \in R^C$ for each token $x_i \in R^C$ based on its contextual region $N(x_i)$ by $y_i = A(P(x_i, N(x_i)), N(x_i))$, where perception P extracts the contextual relationships among tokens and aggregation A integrates features based on it for spatial information fusion.
- Self-attention: $y_i = A_{attn}(P_{attn}(x_i, X), X) = P_{attn}(x_i, X)(XW_v)$ and $P_{attn}(x_i, X) = softmax\left((x_i W_q)(XW_k)^T\right)$. X is the global feature map.
- Convolution: $y_i = A_{conv}\left(P_{conv}(x_i, N_K(x_i)), N_K(x_i)\right) = W_{conv} \otimes N_K(x_i)$. W_{conv} is the static kernel weights and $N_K(\cdot)$ is the $K \times K$ neighbor region.

Methodology

See large, focus small: $y_i = A(P(x_i, N_P(x_i)), N_A(x_i))$, where $N_P(x_i) > N_A(x_i)$.

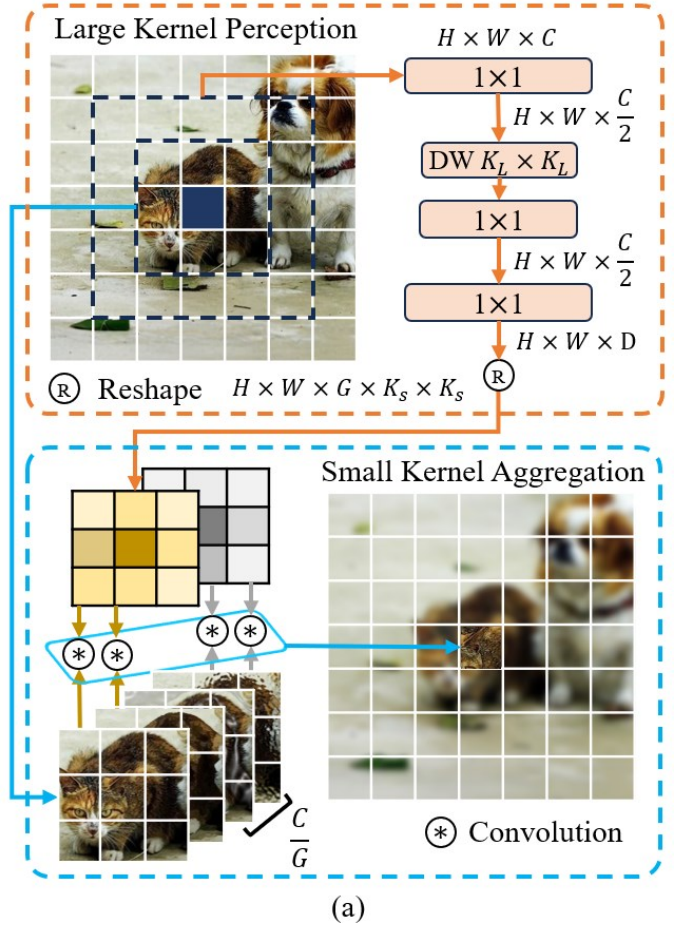
- Perception P and aggregation A involves different contextual scopes, i.e., $N_P(x_i)$ and $N_A(x_i)$, respectively, utilizing heteroscale contextual information and capturing both overall context and fine-grained details.
- For perception with a large spatial extent, cost-effective operations, like large-kernel depth-wise convolution, can be employed. The perception context can thus be enlarged with minimal overhead.
- For the aggregation with a small surrounding region, adaptive weighted feature summation can be adopted. Due to the limited range, the efficiency can be guaranteed and the less important aggregation can be mitigated.



Methodology

LS convolution: See large, focus small

- Large-kernel perception P_{ls} models the neighborhood relationships with enlarged receptive field through large-kernel DW convolutions.
- For P_{ls}, $w_i = PW(DW_{K_L}(PW(N_{K_L}(x_i))))$. K_L is the large kernel size.
- Small-kernel aggregation A_{ls} adaptively integrates highly related surrounding features through small-kernel grouped dynamic conv.
- For A_{ls}, $y_{ic} = w_{ig}^{*} \otimes N_{K_S}(x_{ic})$. K_S is the small kernel size.
- LKP and SKA facilitate the capture of large-field spatial contextual information and enhance fine-grained visual understanding, respectively. They enable more discriminative representations for intricate visual information.

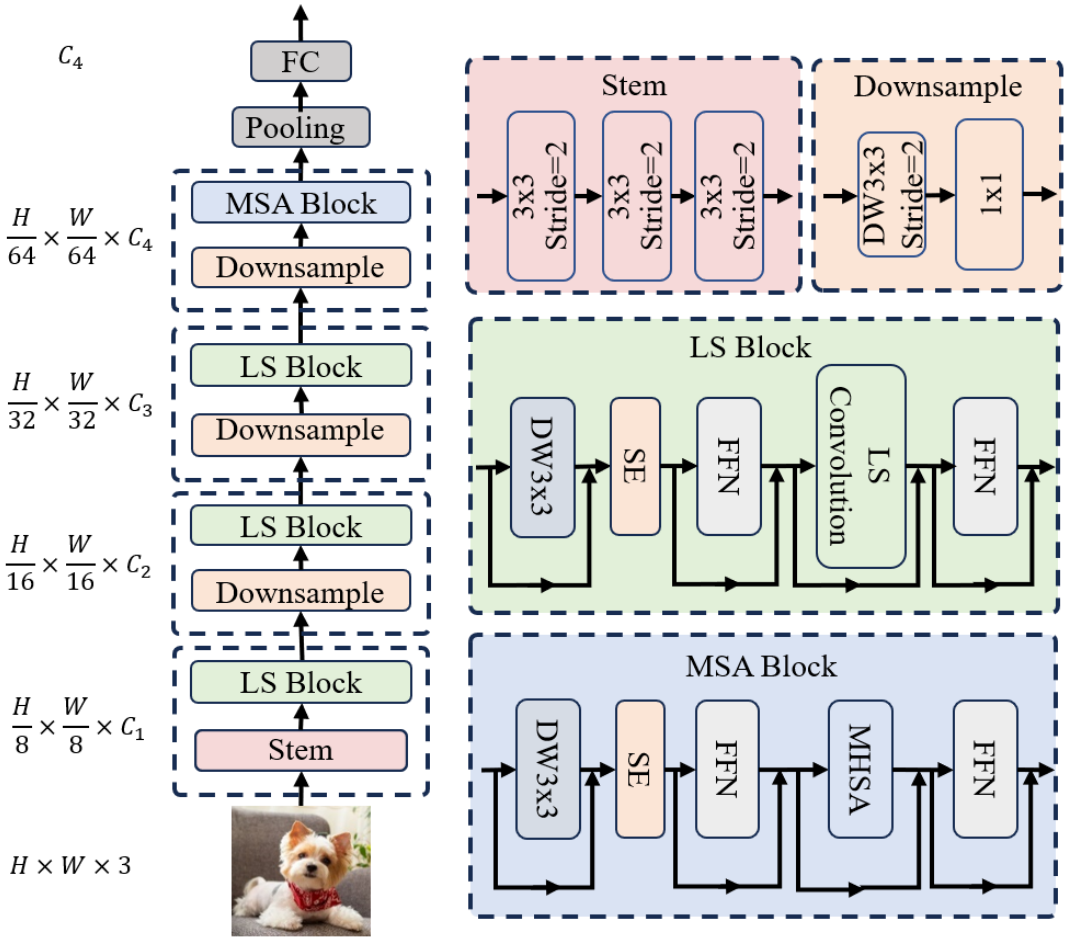


(a)

Methodology

LSNet architecture

- Stem: overlapping patch embedding to project the input image into the visual feature map.
- Downsample: DW and PW convolution to reduce spatial resolution and modulate the channel dimension, respectively
- Top three stages: LS convolution for effective token mixing, with extra DW and SE for enhancement.
- Last stage: MSA block to capture long-range dependencies due to the small resolution



(b)

Experiments

For classification on ImageNet-1K, LSNet consistently outperforms existing lightweight models.

Model	Params (M)	FLOPs (G)	Throughput (img/s)	Top-1 (%)
EdgeNeXt-XXS [48]	1.3	0.3	5089	71.2
FasterNet-T0 [2]	3.9	0.3	14467	71.9
ShuffleNetV2 [46]	3.5	0.3	9593	72.6
AFFNet-ET [28]	1.4	0.4	2877	73.0
EfficientViT-M3 [43]	6.9	0.3	14613	73.4
StarNet-S1 [47]	2.9	0.4	5034	73.5
LSNet-T	11.4	0.3	14708	**74.9**
LSNet-T*	11.4	0.3	14708	**76.1**
EdgeNeXt-XS [48]	2.3	0.5	3118	75.0
PVT-Tiny [73]	13.2	1.9	2125	75.1
MobileNetV3-L [24]	5.4	0.2	7921	75.2
FastViT-T8 [68]	3.6	0.7	3909	75.6
EFormerV2-S0* [39]	3.5	0.4	1329	75.7
FasterNet-T1 [2]	7.6	0.9	8660	76.2
UniRepLKNet-A [10]	4.4	0.6	3931	77.0
EfficientNet-B0 [63]	5.3	0.4	4481	77.1
PoolFormer-S12 [82]	12.0	1.8	2769	77.2
SHViT-S3 [83]	14.2	0.6	8993	77.4
RepViT-M0.9 [71]	5.1	0.8	4817	77.4
LSNet-S	16.1	0.5	9023	**77.8**
LSNet-S*	16.1	0.5	9023	**79.0**
EdgeViT-XS [53]	6.7	1.1	2751	77.5
SwiftFormer-S* [60]	6.1	1.0	3376	78.5
UniRepLKNet-F [10]	6.2	0.9	3209	78.6
FastViT-T12 [68]	6.8	1.4	2586	79.1
EFormer-L1* [38]	12.3	1.3	3280	79.2
EdgeNeXt-S [48]	5.6	1.3	2128	79.4
RepViT-M1.1 [71]	8.2	1.3	3604	79.4
PVT-Small [73]	24.5	3.8	1160	79.8
AFFNet [28]	5.5	1.5	1355	79.8
LSNet-B	23.2	1.3	3996	**80.3**
LSNet-B*	23.2	1.3	3996	**81.6**

LSNet transfers well for object detection & instance segmentation on COCO

Backbone	FLOPs	RetinaNet						Mask R-CNN					
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
MobileNetV2 [59]	1.6	28.3	46.7	29.3	14.8	30.7	38.1	29.6	48.3	31.5	27.2	45.2	28.6
MobileNetV3 [24]	1.1	29.9	49.3	30.8	14.9	33.3	41.1	29.2	48.6	30.3	27.1	45.5	28.2
FairNAS-C [6]	1.7	31.2	50.8	32.7	16.3	34.4	42.3	31.8	51.2	33.8	29.4	48.3	31.0
EfficientViT-M4 [43]	1.6	32.7	52.2	34.1	17.6	35.3	46.0	32.8	54.4	34.5	31.0	51.2	32.2
StarNet-S1 [47]	2.2	33.6	53.3	35.1	**18.3**	36.0	47.0	33.8	56.1	35.5	31.9	52.9	33.4
LSNet-T	1.5	**34.2**	**54.6**	**35.2**	17.8	**37.1**	**48.5**	**35.0**	**57.0**	**37.3**	**32.7**	**53.8**	**34.3**
ResNet18 [18]	9.5	31.8	49.6	33.6	16.3	34.3	43.2	34.0	54.0	36.7	31.2	51.0	32.7
DFvT-T [14]	6.9	-	-	-	-	-	-	34.8	56.9	37.0	32.6	53.7	34.5
EfficientViT-M5 [43]	2.8	34.3	54.2	36.1	18.0	36.9	48.2	34.9	57.0	37.0	32.8	53.7	34.6
SHViT-S3 [83]	3.0	36.1	56.6	38.0	19.9	39.1	50.8	36.9	59.4	39.6	34.4	56.3	36.1
LSNet-S	2.6	**36.7**	**57.2**	**38.6**	**20.0**	**39.7**	**51.8**	**37.4**	**59.9**	**39.8**	**34.8**	**56.8**	**36.6**
ResNet50 [18]	21.4	36.3	55.3	38.6	19.3	40.0	48.8	38.0	58.6	41.4	34.4	55.1	36.7
PVT-Tiny [73]	11.8	36.7	56.9	38.9	**22.6**	38.8	50.0	36.7	59.2	39.3	35.1	56.7	37.3
PoolFormer-S12 [82]	9.5	36.2	56.2	38.2	20.8	39.1	48.0	37.3	59.0	40.1	34.6	55.8	36.9
FasterNet-S [2]	23.8	-	-	-	-	-	-	39.9	61.2	43.6	36.9	58.1	39.7
FastViT-SA12 [68]	7.7	-	-	-	-	-	-	38.9	60.5	42.2	35.9	57.6	38.1
RepViT-M1.1 [71]	7.0	-	-	-	-	-	-	39.8	61.9	43.5	37.2	58.8	**40.1**
LSNet-B	6.2	**39.2**	**60.0**	**41.5**	22.1	**43.0**	**52.9**	**40.8**	**63.4**	**44.0**	**37.8**	**60.5**	**40.1**

Experiments

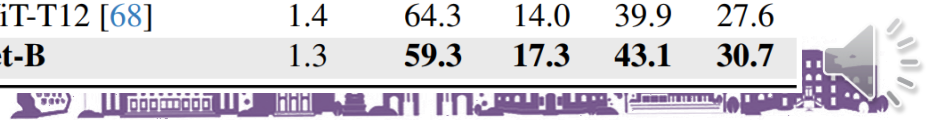
LSNet transfers well for semantic segmentation on ADE20K

Backbone	FLOPs	mIoU
StarNet-S1	2.2	36.0
MobileNetV3	1.1	37.0
PVTv2-B0	3.8	37.2
VAN-B0	4.5	38.5
LSNet-T	1.5	**40.1**
EdgeViT-XXS	3.2	39.7
SHViT-S3	3.0	40.0
FastViT-SA12	7.7	38.0
RepViT-M1.1	7.0	40.6
LSNet-S	2.6	**41.0**

Backbone	FLOPs	mIoU
EFormer-L1	6.8	38.9
PVT-Small	23.1	39.8
PoolFormer-S24	17.8	40.3
FastViT-SA24	15.0	41.0
EdgeViT-XS	6.3	41.4
SwiftFormer-L1	8.3	41.4
Swin-T	25.6	41.5
EFormerV2-S2	7.3	42.4
PVTv2-B1	12.8	42.5
LSNet-B	6.2	**43.0**

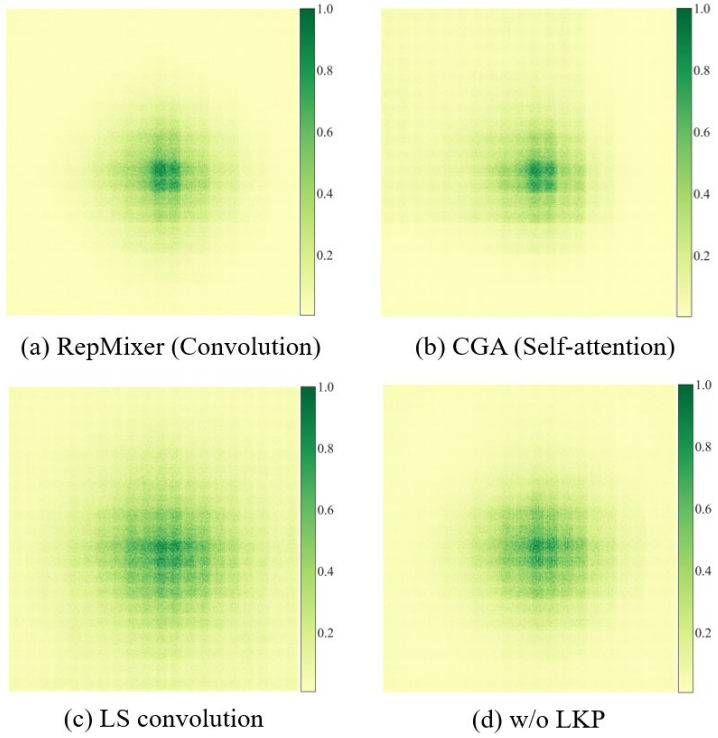
LSNet shows strong domain generalization capabilities and promising robustness on ImageNet-C/A/R/Sketch

Model	FLOPs	C (↓)	A	R	SK
FasterNet-T0 [2]	0.3	89.8	2.3	28.6	16.3
EdgeNeXt-XXS [48]	0.3	94.6	3.6	29.5	18.5
EfficientViT-M3 [43]	0.3	71.1	5.2	36.1	23.4
StarNet-S1 [47]	0.4	77.5	4.5	34.1	21.8
LSNet-T	0.3	**68.2**	**6.7**	**38.5**	**25.5**
FastViT-T8 [68]	0.7	72.1	6.9	36.8	25.5
PVTv2-B0 [74]	0.6	75.4	4.2	34.2	21.5
EdgeNeXt-XS [48]	0.5	88.4	6.3	32.5	22.0
UniRepLKNet-A [10]	0.6	67.0	8.4	37.9	26.0
LSNet-S	0.5	**65.7**	**9.6**	**39.4**	**27.5**
PVT-Tiny [73]	1.9	79.6	7.9	33.9	21.5
PoolFormer-S12 [82]	1.8	67.7	6.9	37.7	25.2
FasterNet-T2 [2]	1.9	70.8	8.7	40.5	27.2
EdgeNeXt-S [48]	1.3	72.1	11.9	40.1	28.8
PVTv2-B1 [74]	2.1	62.2	14.6	41.8	28.9
FastViT-T12 [68]	1.4	64.3	14.0	39.9	27.6
LSNet-B	1.3	**59.3**	**17.3**	**43.1**	**30.7**



Experiments

LS convolution enables broad peripheral perception and central view focusing simultaneously, showing large and natural effective receptive field.



(a) RepMixer (Convolution) (b) CGA (Self-attention)

(c) LS convolution (d) w/o LKP

LS convolution can help the model to grasp the critical visual information under limited computational costs, enhancing both efficiency and effectiveness.

清華大學
Tsinghua University

THANKS!