NEURAL INFORMATION PROCESSING SYSTEMS

Revisiting the Integration of Convolution and Attention for Vision Backbone

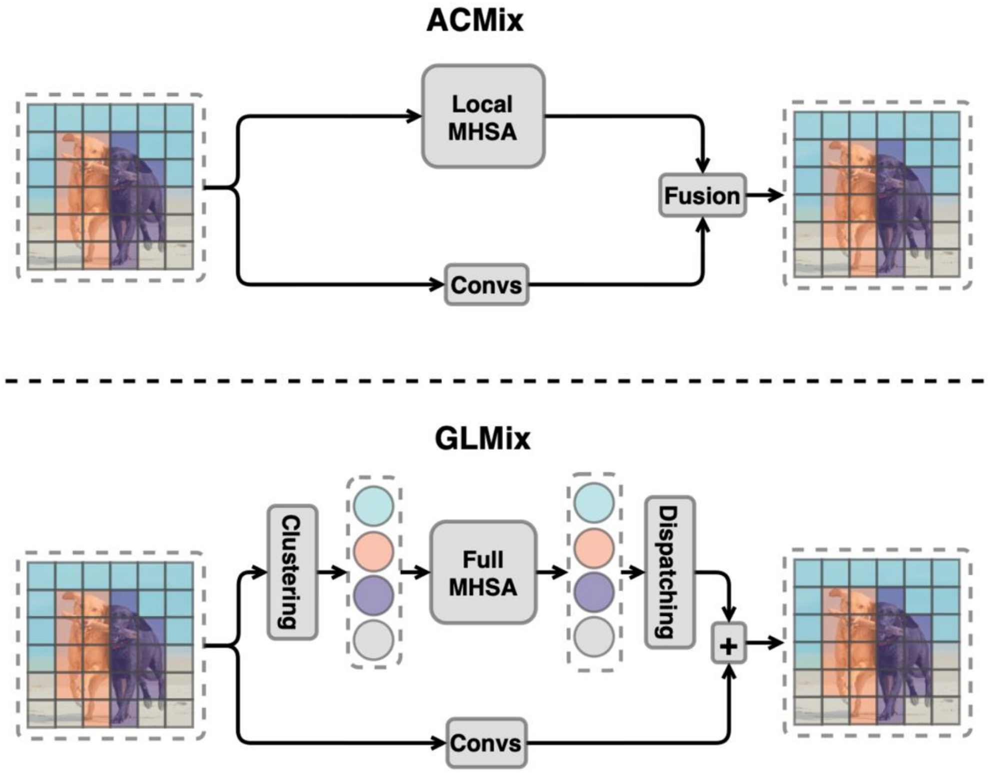
Lei Zhu[1], Xinjiang Wang[2], Wayne Zhang[2] and Rynson Lau[1]

[1]City University of Hong Kong, [2]SenseTime Research

CityU 香港城市大學 City University of Hong Kong

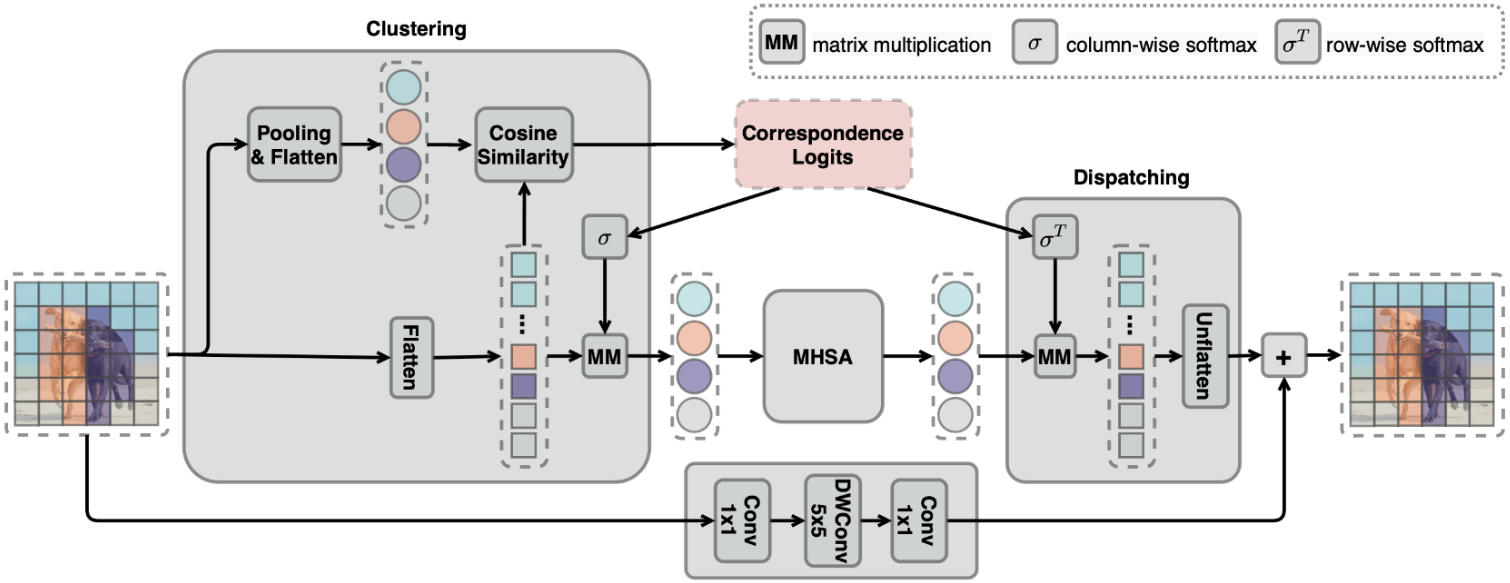
sensetime

Motivation



- Integrating Convs & MHSAs in vision backbones has shown better accuracy than using a single one of them (e.g. ACMix, CVPR 2022)
- However, **do we need both Convs and MHSAs at the finest pixel/token level ?**
- GLMix: apply Convs and MHSAs **at different granularity levels**
  - (light-weight) Convs for finegrained feature grids
  - (heavy) MHSAs on a set of coarse-grained semantic slots

Methodology



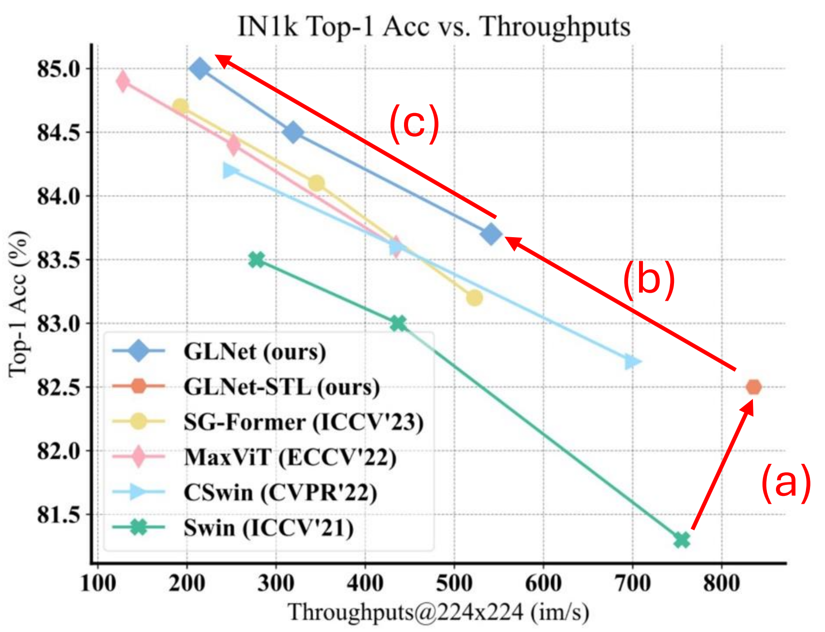
- Parallel design with a **G**lobal branch using attention and a **L**ocal branch using Convs

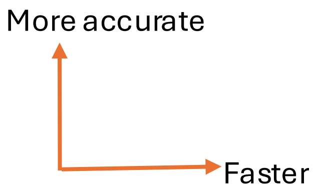
- The heavy attention operator only process a coarse **set** of semantic slots (e.g. 64 slots)

- The finegrained feature **grid** is processed by lightweight convolutions

- A pair of soft clustering (grid -> set) and dispatching (set -> grid) modules are introduced to bridge the set and grid representations

Methodology





- We start by creating a **S**win-**T**iny-**L**ayout architecture GLNet-**STL**
  - (a) Replacing the window attention in Swin-Tiny with GLMix. the GLNet-STL is both efficient and effective
- To compare with recent SOTA models
  - (b) We then adopt the several advanced architectural designs from existing works to derive GLNet-4G; and
  - (c) scale up the model by the width (channels) to derive GLNet-GLNet-9G and GLNet-16G
- The GLNet family push the Pareto frontier of accuracy-throughput further to the upper-right corner
- Detailed comparisons with more models and on more tasks (e.g., object detection, instance segmentation, and semantic segmentation) can be found in the paper.

Ablation Study

Model	Slot init.	Slot number	Conv k.s.	FLOPs (G)	Params (M)	Throu. (im/s)	IN1k Top-1 (%)
GLNet-STL	pooling	64	5	4.4	30.3	835.9	82.5
local branch only	pooling	-	5	3.8	26.4	999.7	81.8
global branch only	pooling	64	-	3.8	28.3	982.4	78.0
sequential (global $\rightarrow$ local)	pooling	64	5	4.4	30.3	860.1	80.6
sequential (local $\rightarrow$ global)	pooling	64	5	4.4	30.3	825.9	79.6
local branch w/ W-MHSA†	pooling	64	w7	5.0	32.2	660.9	81.1
k-means clustering‡	hashing	64	5	5.2	30.3	440.6	N/A
static slot initialization	param.	64	5	4.4	30.5	852.0	82.1
local w/ 7×7 DWConv	pooling	64	7	4.4	30.3	855.2	82.4
local w/ 3×3 DWConv	pooling	64	3	4.3	30.4	823.9	82.4
global w/ 9 slots	pooling	9	5	3.9	30.3	893.6	81.9
global w/ 25 slots	pooling	25	5	4.0	30.3	880.8	82.1
global w/ 36 slots	pooling	36	5	4.1	30.3	880.0	82.3
global w/ 49 slots	pooling	49	5	4.2	30.3	866.6	82.3
global w/ 81 slots	pooling	81	5	4.5	30.3	790.0	82.4

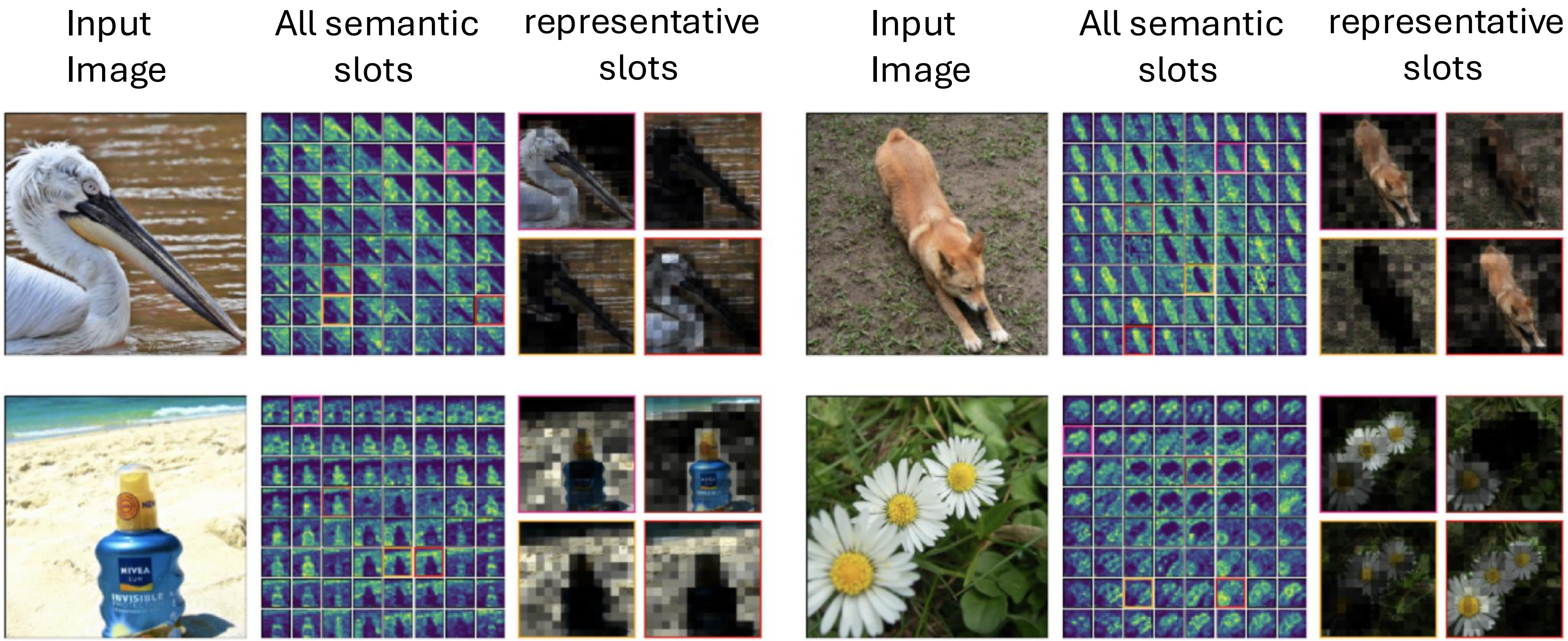
- Local-global collaboration
  - Local + global > only local or only global
  - Parallel > sequential
  - Using convs in local branch is better than window attention

- Clustering strategy
  - Soft clustering ,instead of the hard one with k-means, is crucial for both stable training and efficiency (throughput)
  - Initialization with **per-image** adaptive pooling is better than using shared static parameters

- The receptive field of the local branch does not matter

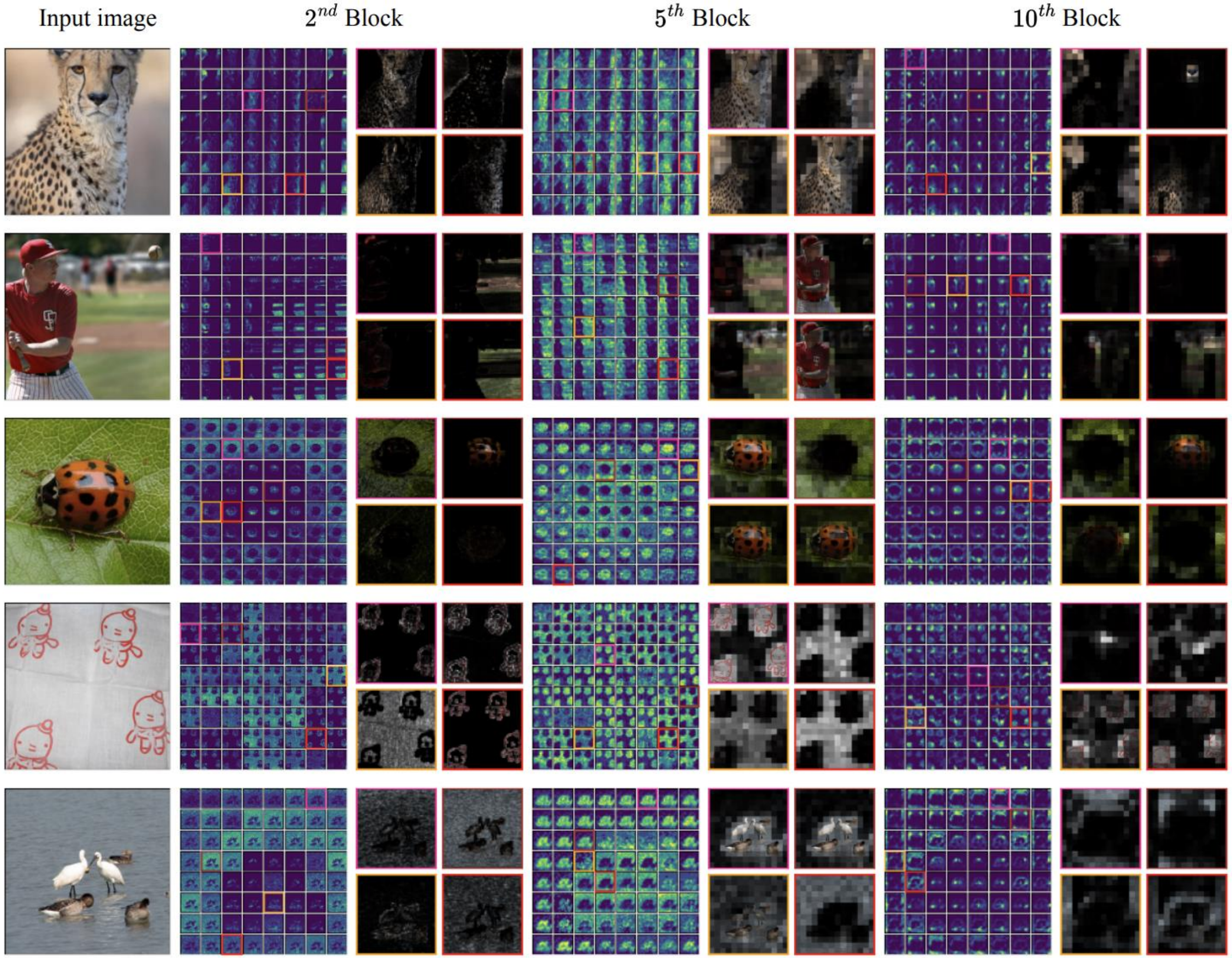
- It is sufficient to use 64 semantic slots in the global branch

Visualization



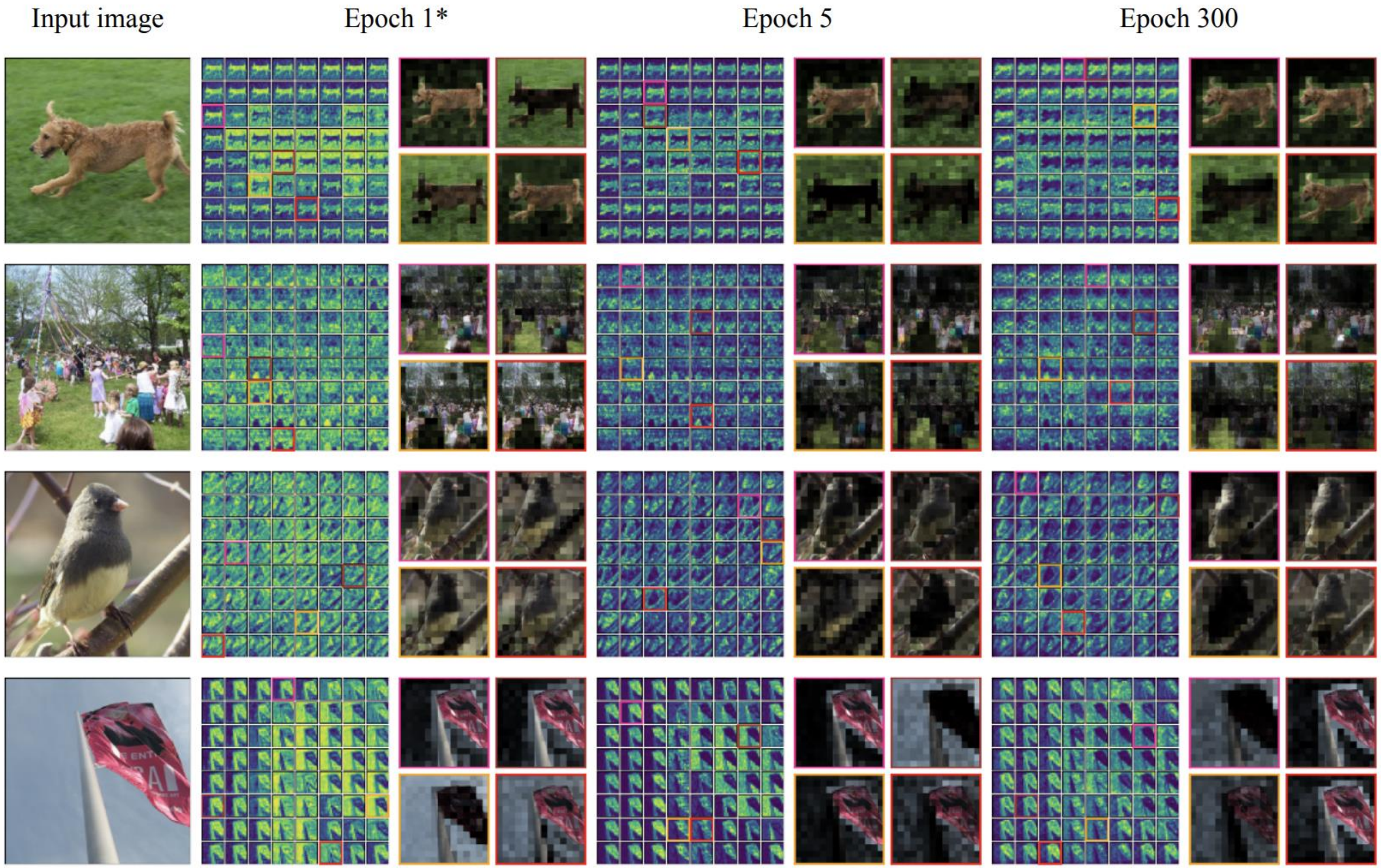
- The 64 semantic slots are visualized by pseudo-colorizing the assignment weights in clustering
- The 4 representative slots are selected automatically by the k-medoids algorithm
- Meaningful semantic grouping effect emerges in the soft clustering module **with only image-level supervision**
- You can find more visualizations for layers at different depths and over the training epochs in our paper

Visualization



- lower block (2^{nd} block) tends to group pixels according to color cues.
- At the middle block (5^{th} block), an object-level grouping effect has emerged.
- The upper block (10^{th} block) pays attention to discriminative local regions.

Visualization



During the training, we found that :

- At the end of the 1st epoch, we can already distinguish the foreground objects and the backgrounds, although the grouping has not very concentrated patterns
- At the end of the 5th epoch, the semantic grouping becomes more concentrated and similar to that of the final stage.

Conclusion

- We propose a novel integration scheme of Convs and MHSAs by applying the two operators at **different granularity levels**
- Through extensive experiments, it is discovered that by offloading the burden of fine-grained features into lightweight Convs, MHSAs can be aggressively applied to a few (e.g. 64) semantic slots
- It‘s observed that meaningful semantic grouping effects emerge in the soft clustering module, which is introduced to bridge the feature grid and semantic slots