Hard-Attention for Scalable Image Classification

Athanasios Papadopoulos, Paweł Korus, Nasir Memon

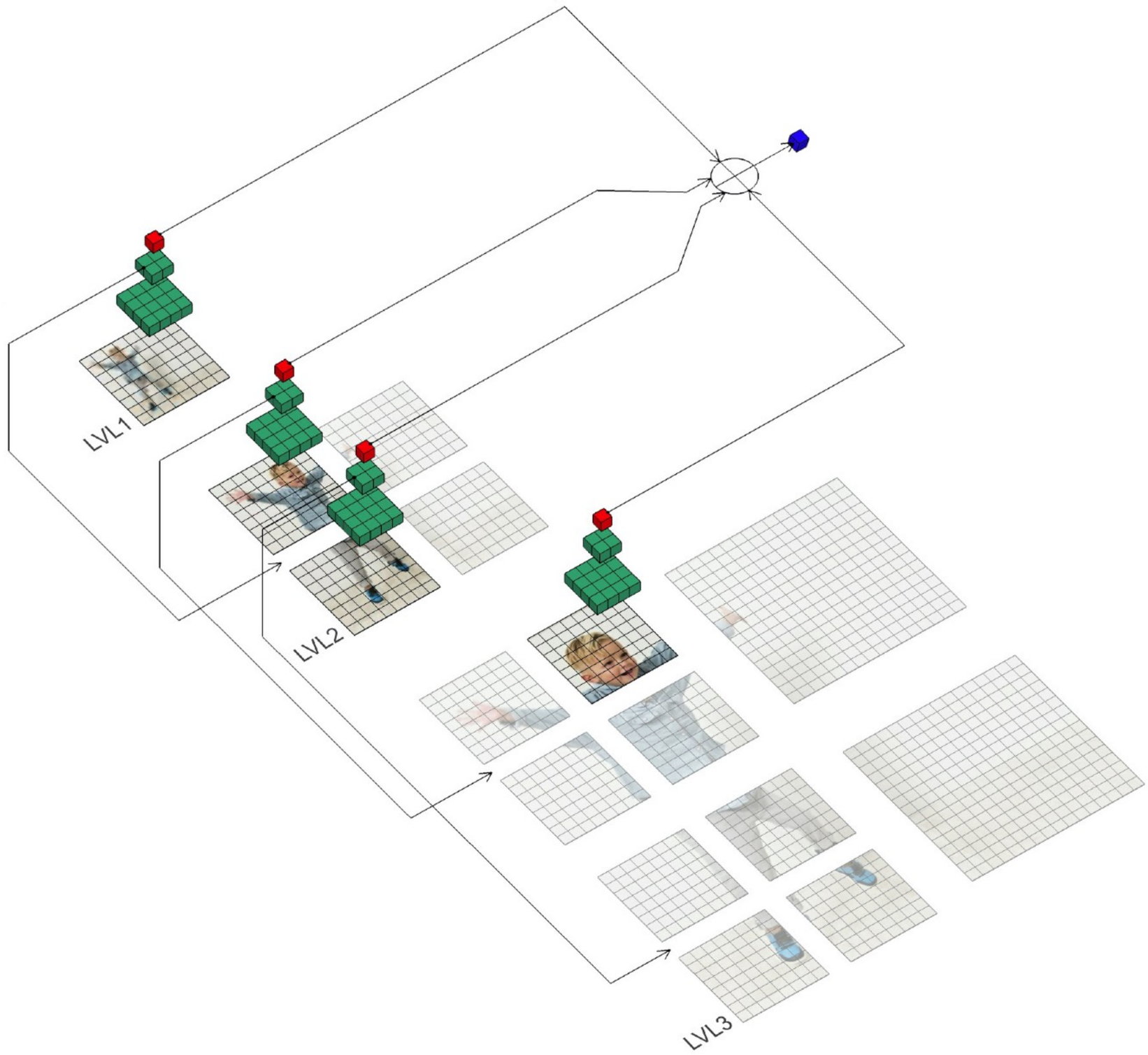
Motivation

- Linear increase to input spatial dimensions, leads to quadratic increase in computation and memory.

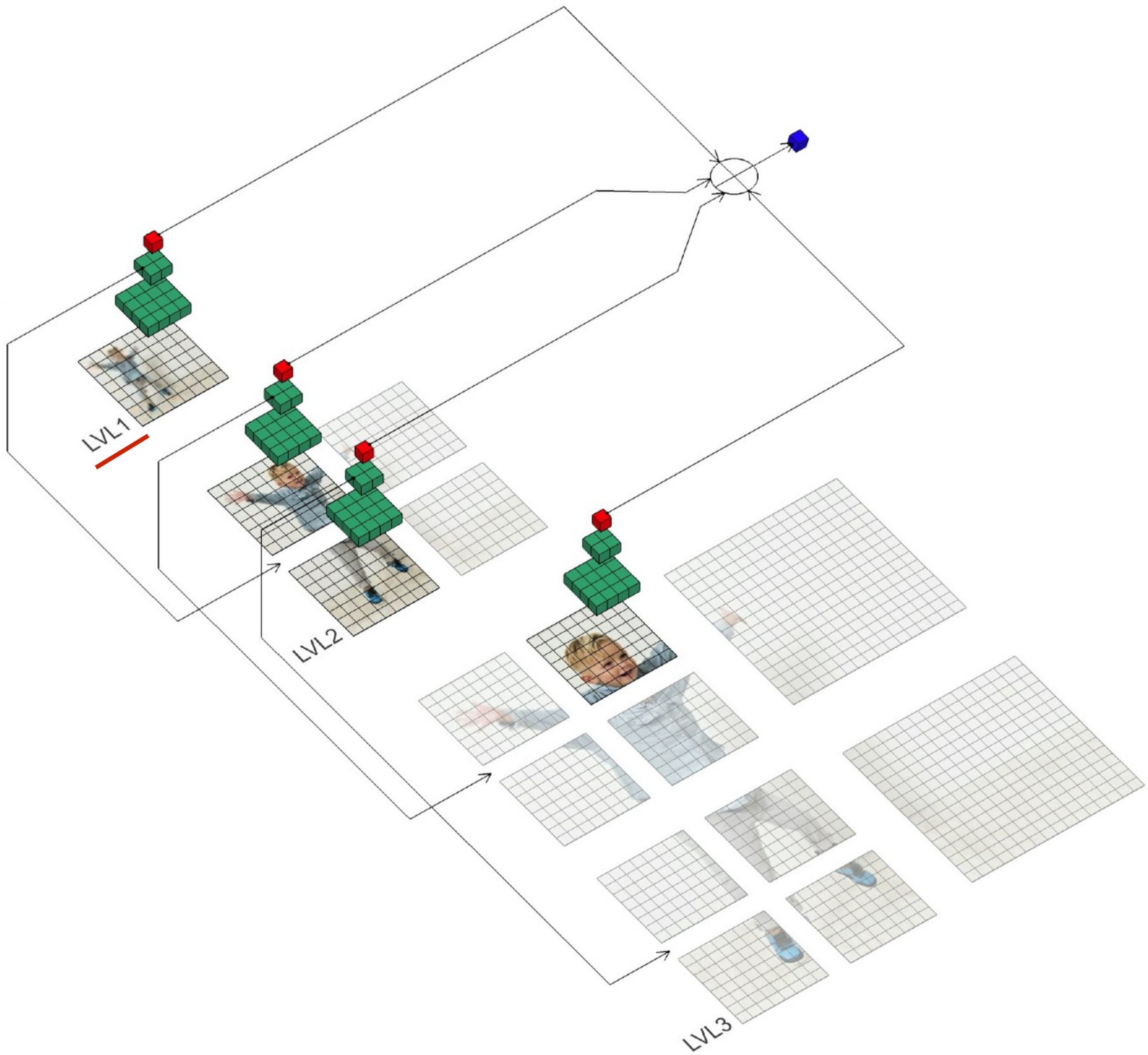
- Can we leverage high-resolution information, without the unsustainable quadratic complexity to input scale?

- Selective processing:
  - Process part of the input
  - Top-down understanding – context drives focus to finer details

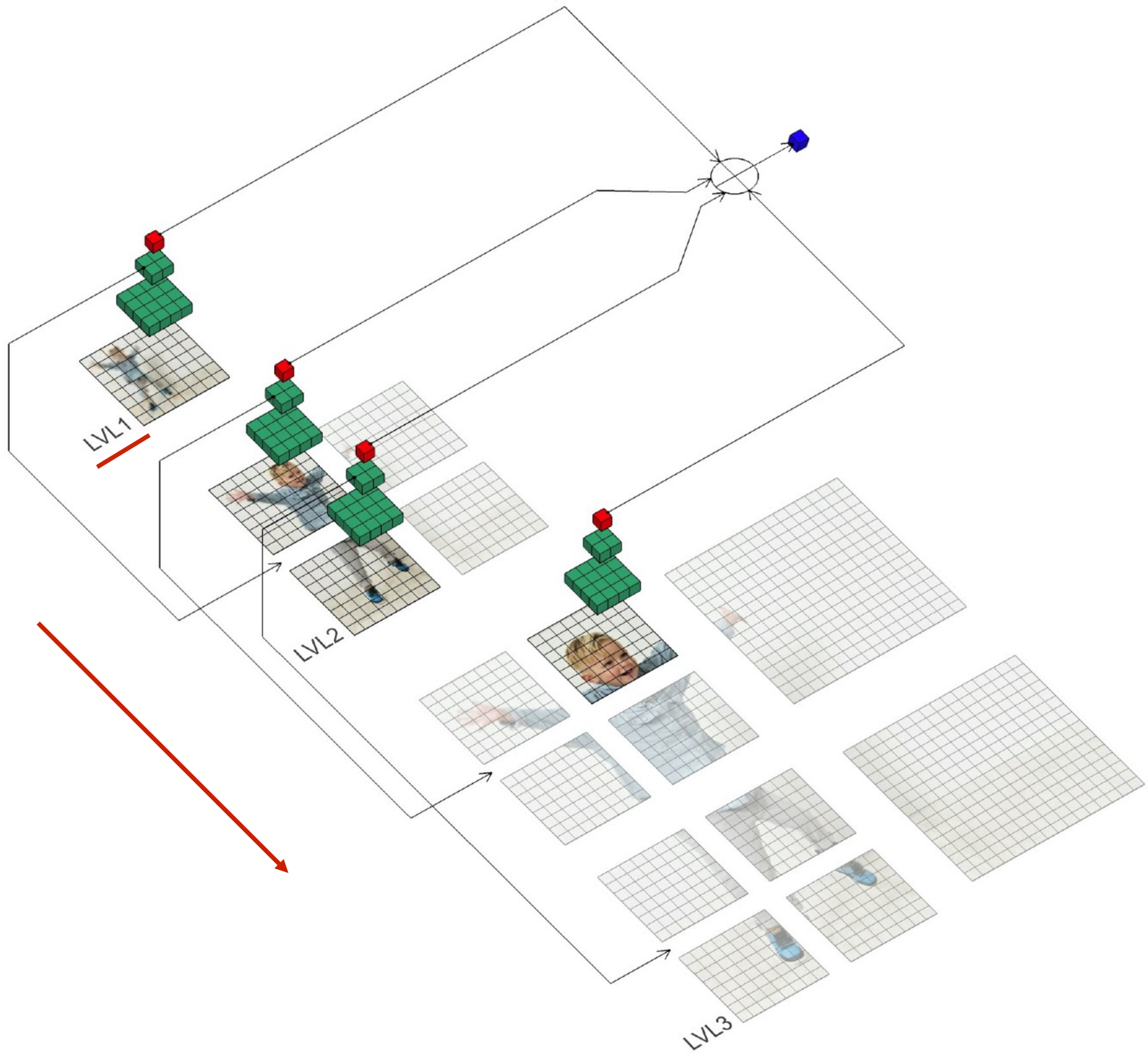
Traversal Network (TNet)



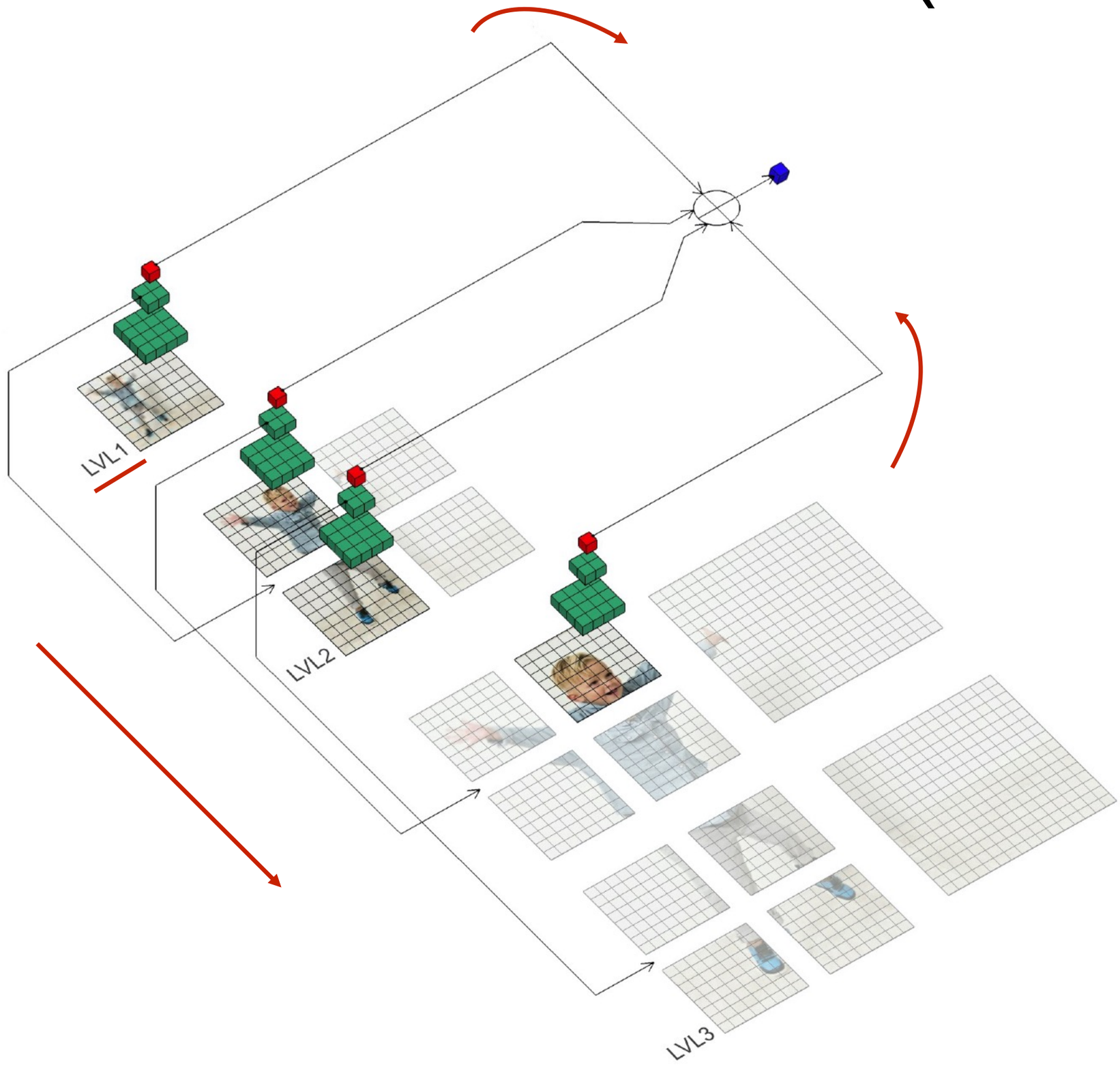
Traversal Network (TNet)



Traversal Network (TNet)

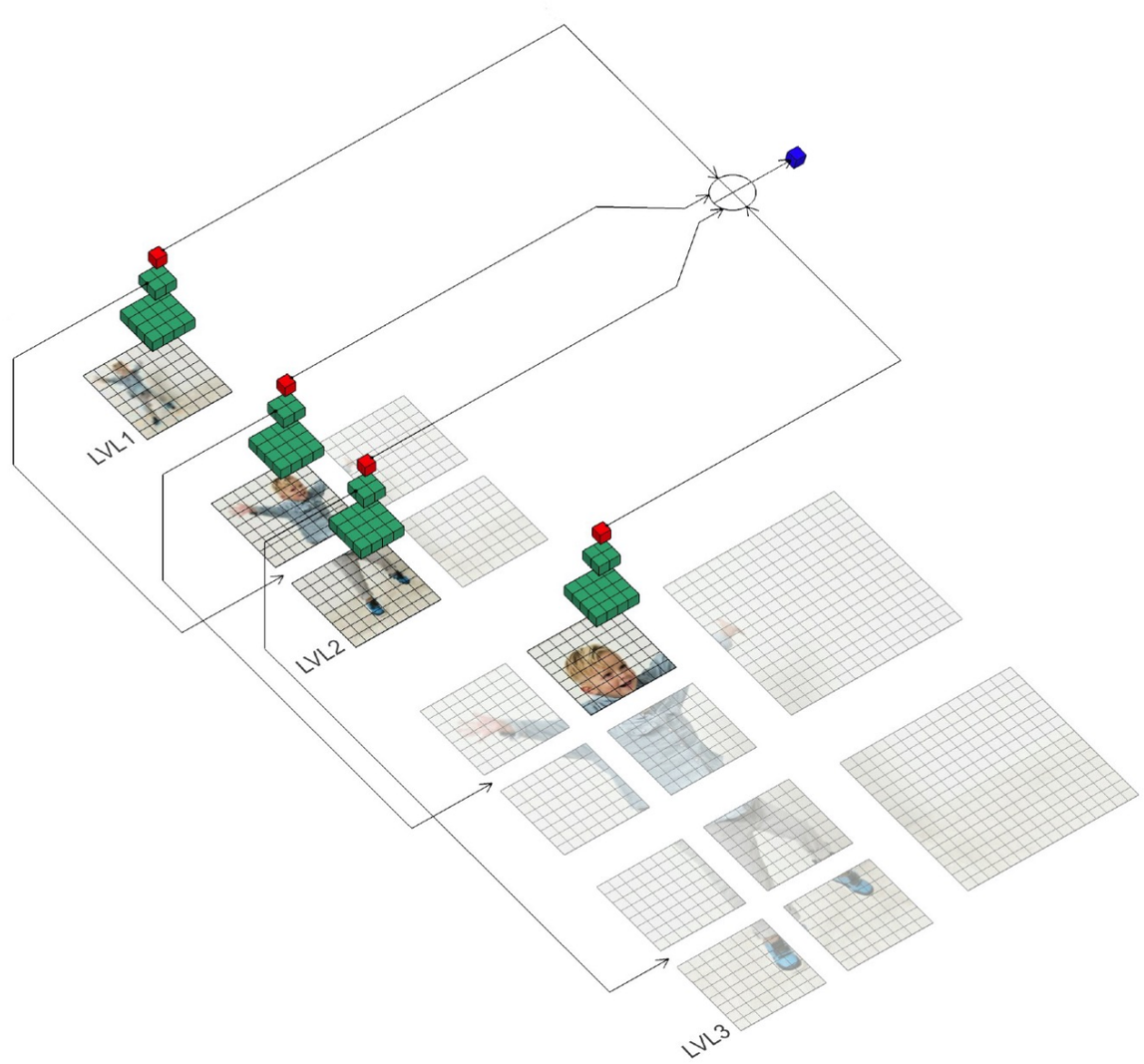


Traversal Network (TNet)

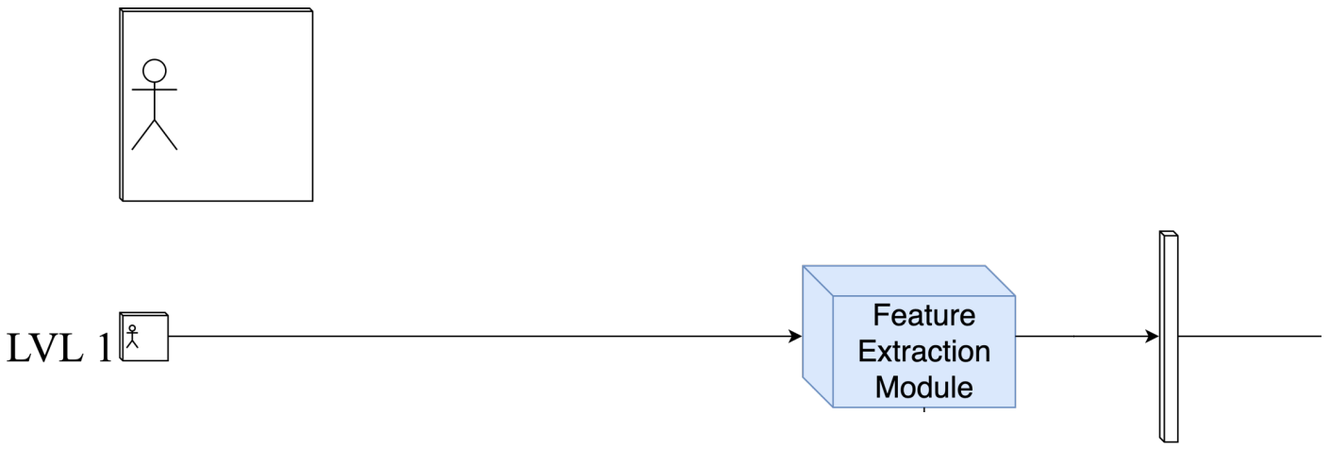


Traversal Network (TNet)

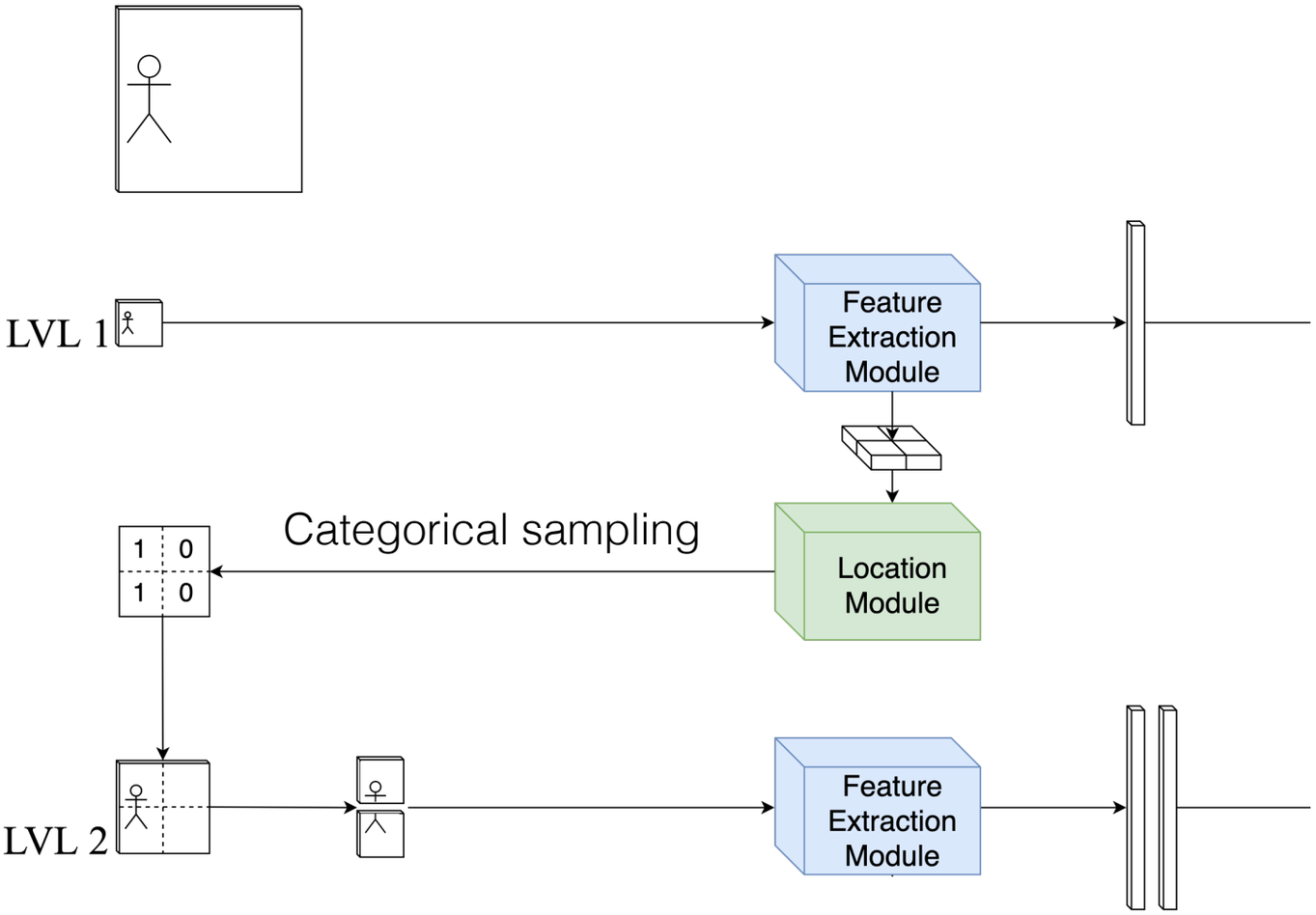
- Complexity linear to number of locations
- Dynamic inference
- Can virtually be applied to any resolution
- Interpretable predictions without extra cost
- End-to-end trainable, only class labels
- Compatible with top-performing models

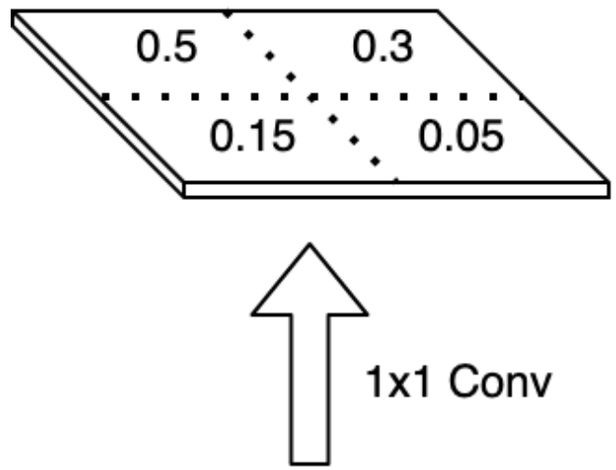


Architecture

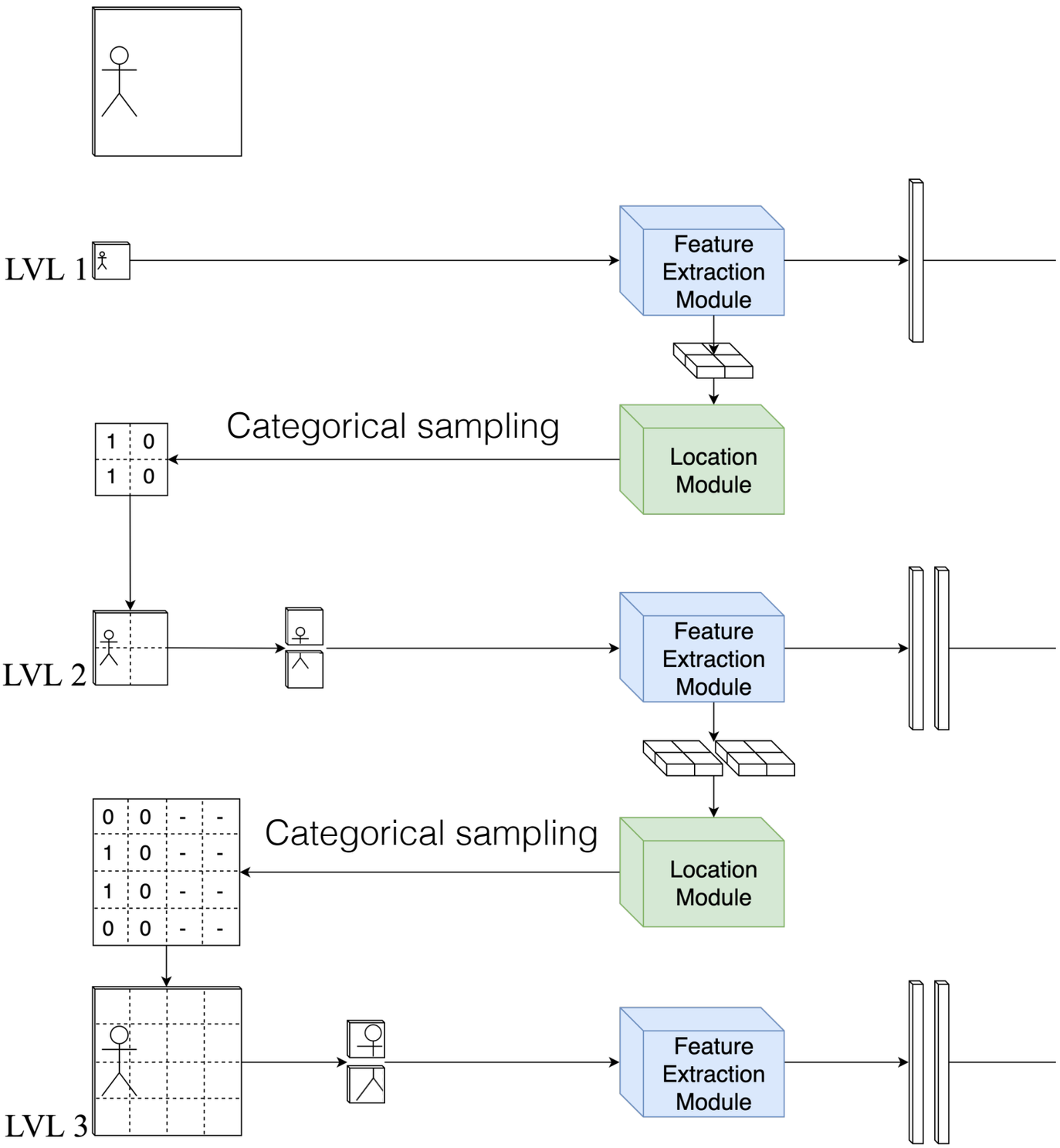


Architecture

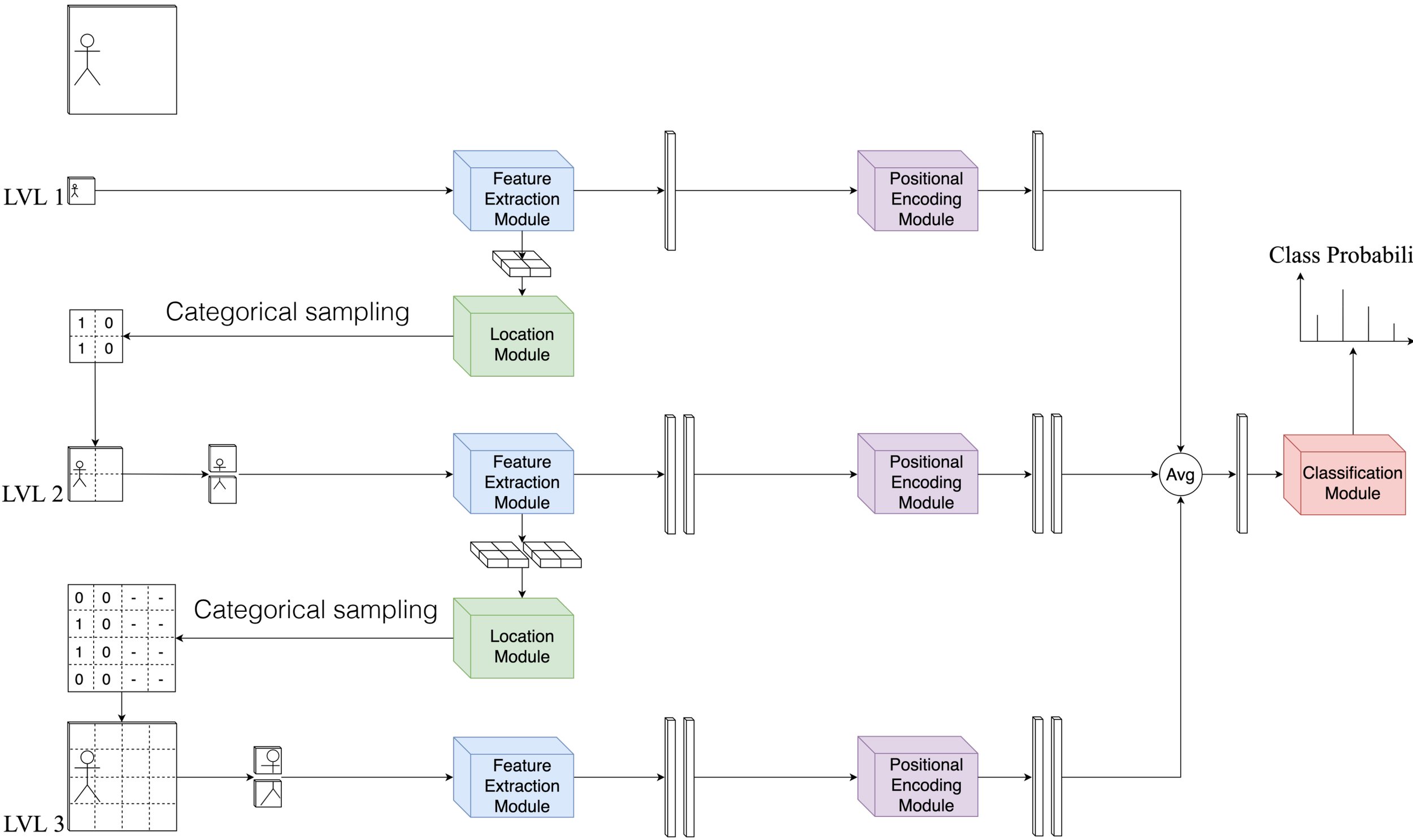




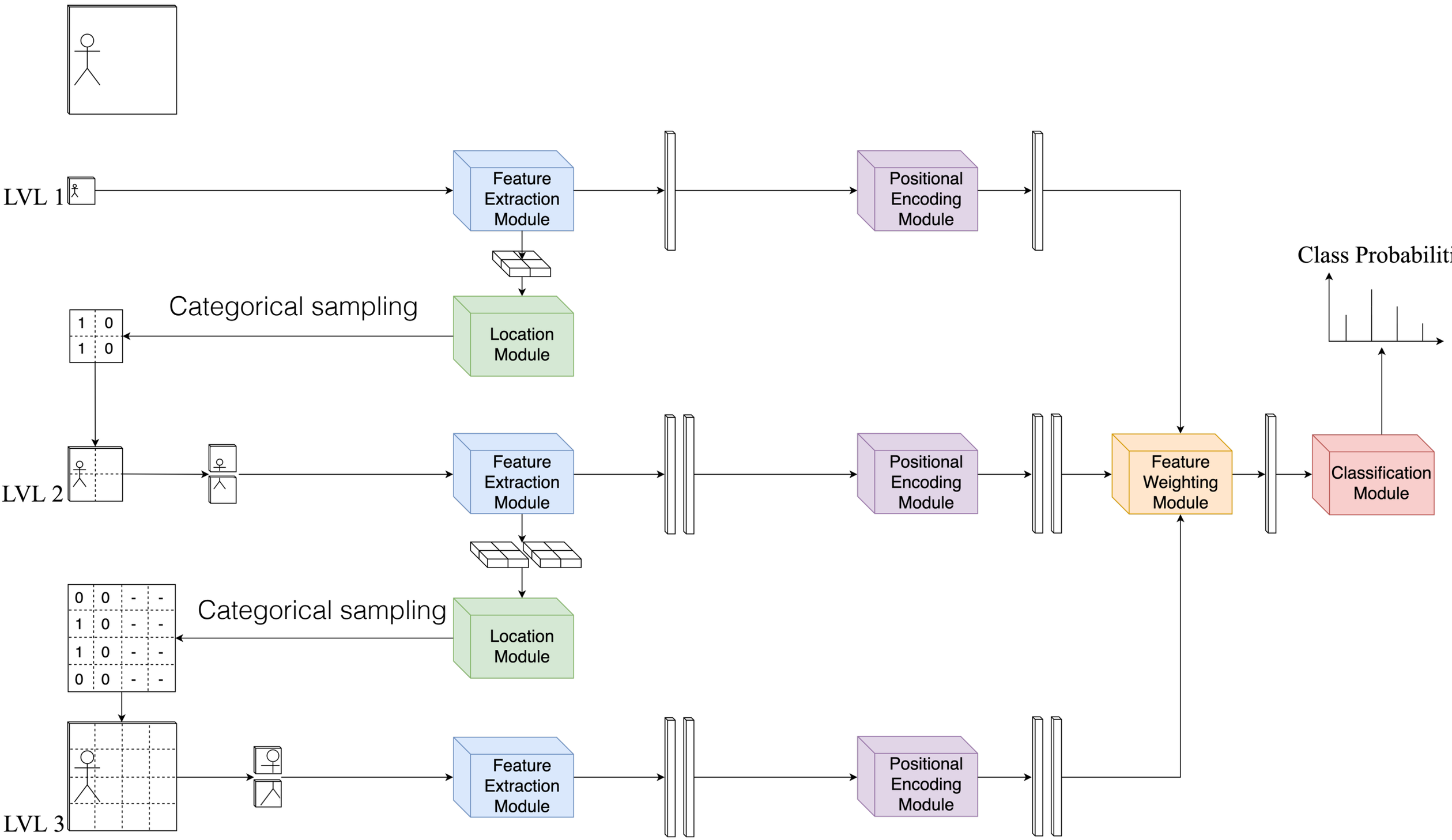
Architecture



Architecture



Architecture

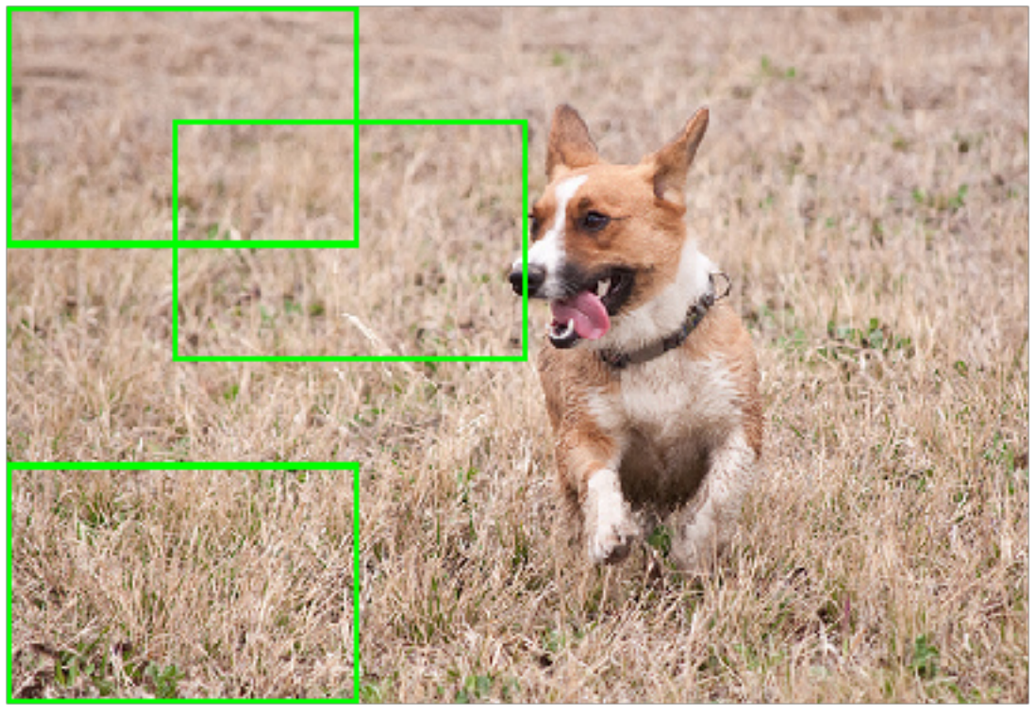


Training

- Because of sampling, not end-to-end differentiable, but with reinforcement learning end-to-end trainable

$$L_F = \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{\partial \log p(y_i|l^i, x_i, w)}{\partial w} + \lambda_f(R_i - b)\frac{\partial \log p(l^i|x_i, w)}{\partial w}\Big]$$

Per-Feature Regularization

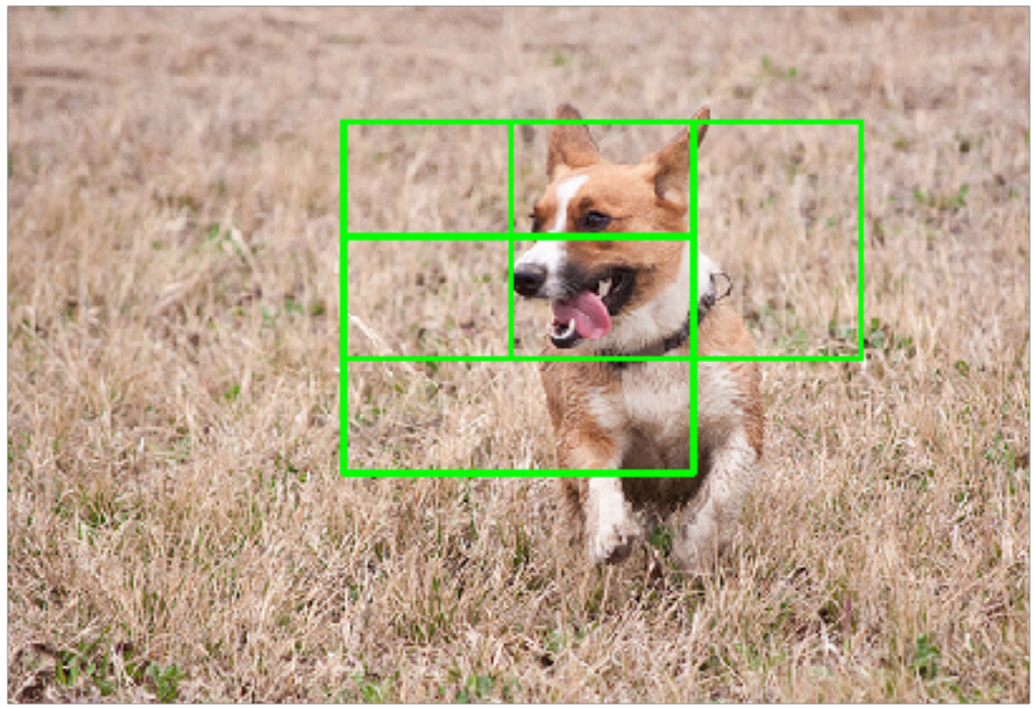


Training

- Because of sampling, not end-to-end differentiable, but with reinforcement learning end-to-end trainable

$$L_F = \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{\partial \log p(y_i|l^i, x_i, w)}{\partial w} + \lambda_f(R_i - b)\frac{\partial \log p(l^i|x_i, w)}{\partial w}\Big]$$

Per-Feature Regularization

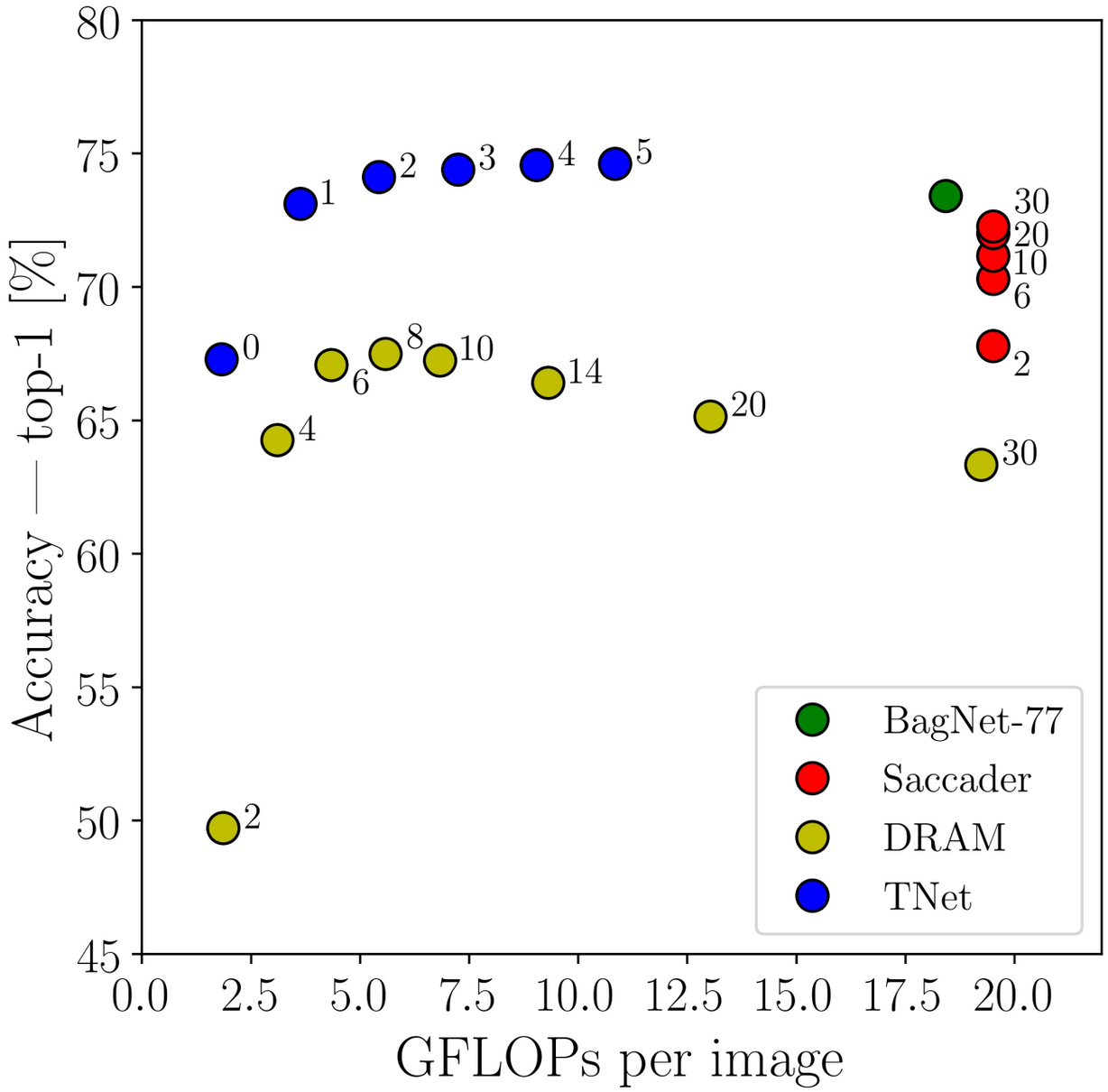


Experimental Evaluation

Effectiveness of Hard-Attention

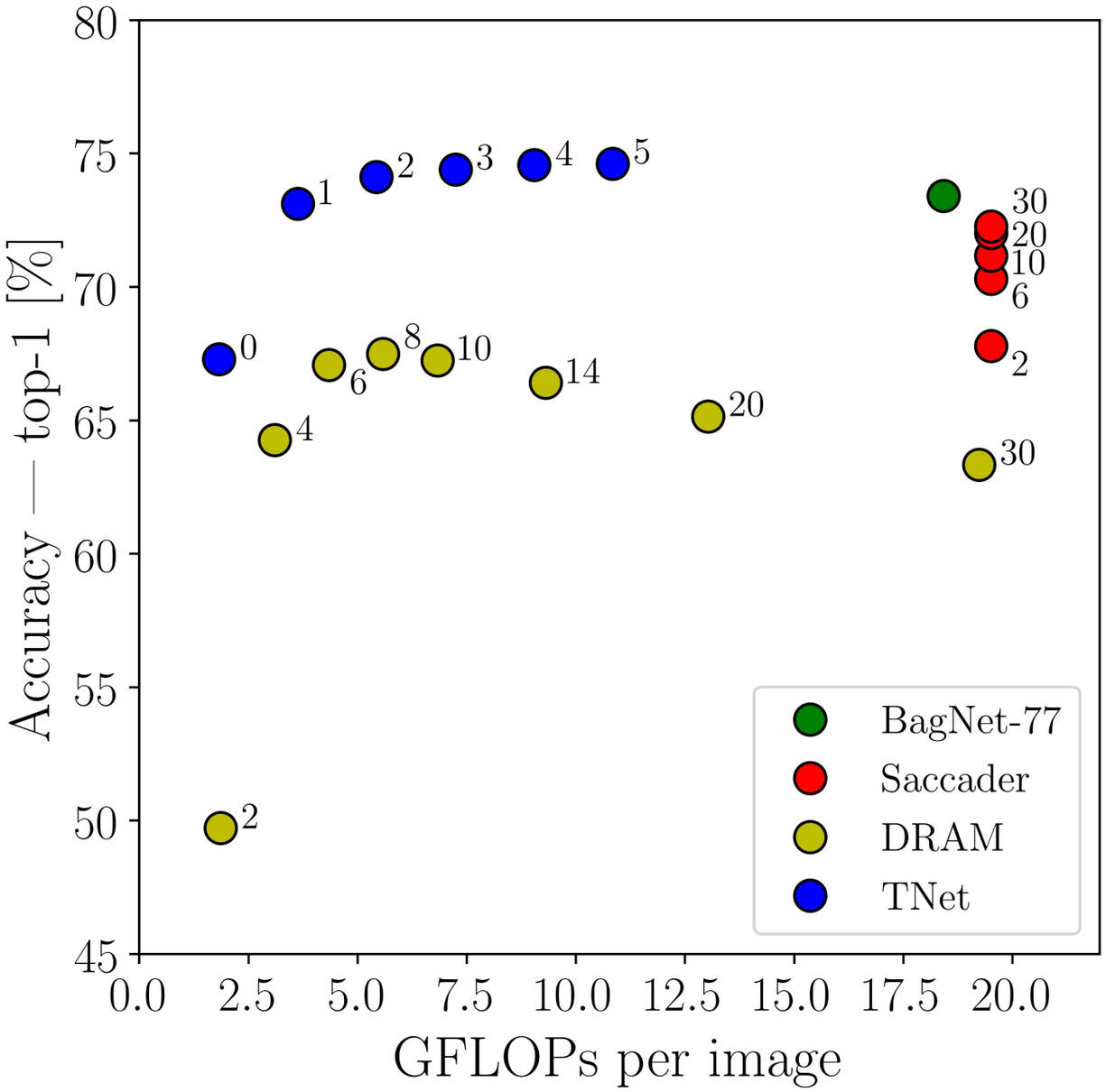
- **Data:** ImageNet, 224x224 px inputs
- **Baselines:** Saccader, DRAM, BagNet-77
- **TNet:** base resolution 77x77 px, 5x5 grid, 2 levels

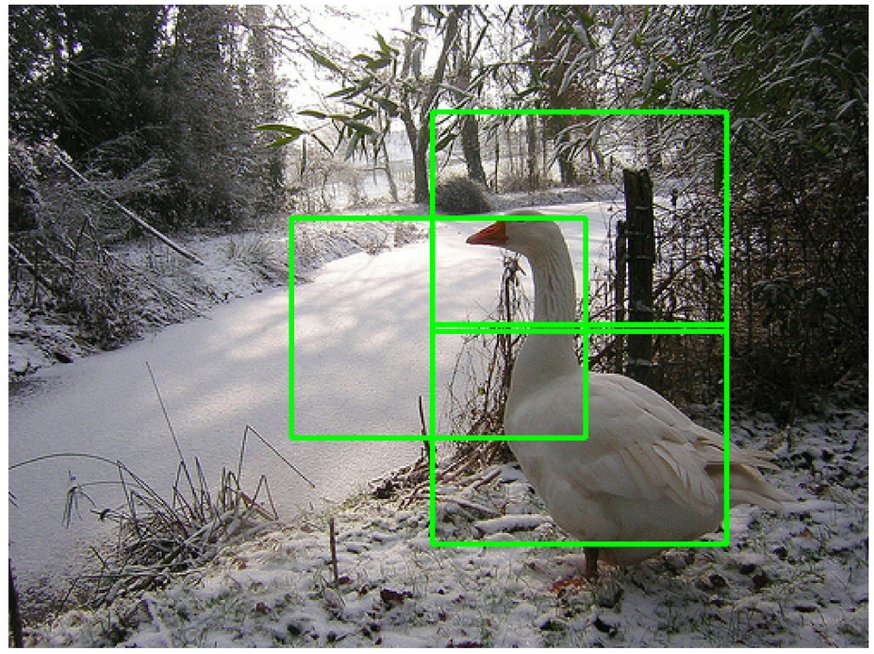
Effectiveness of Hard-Attention



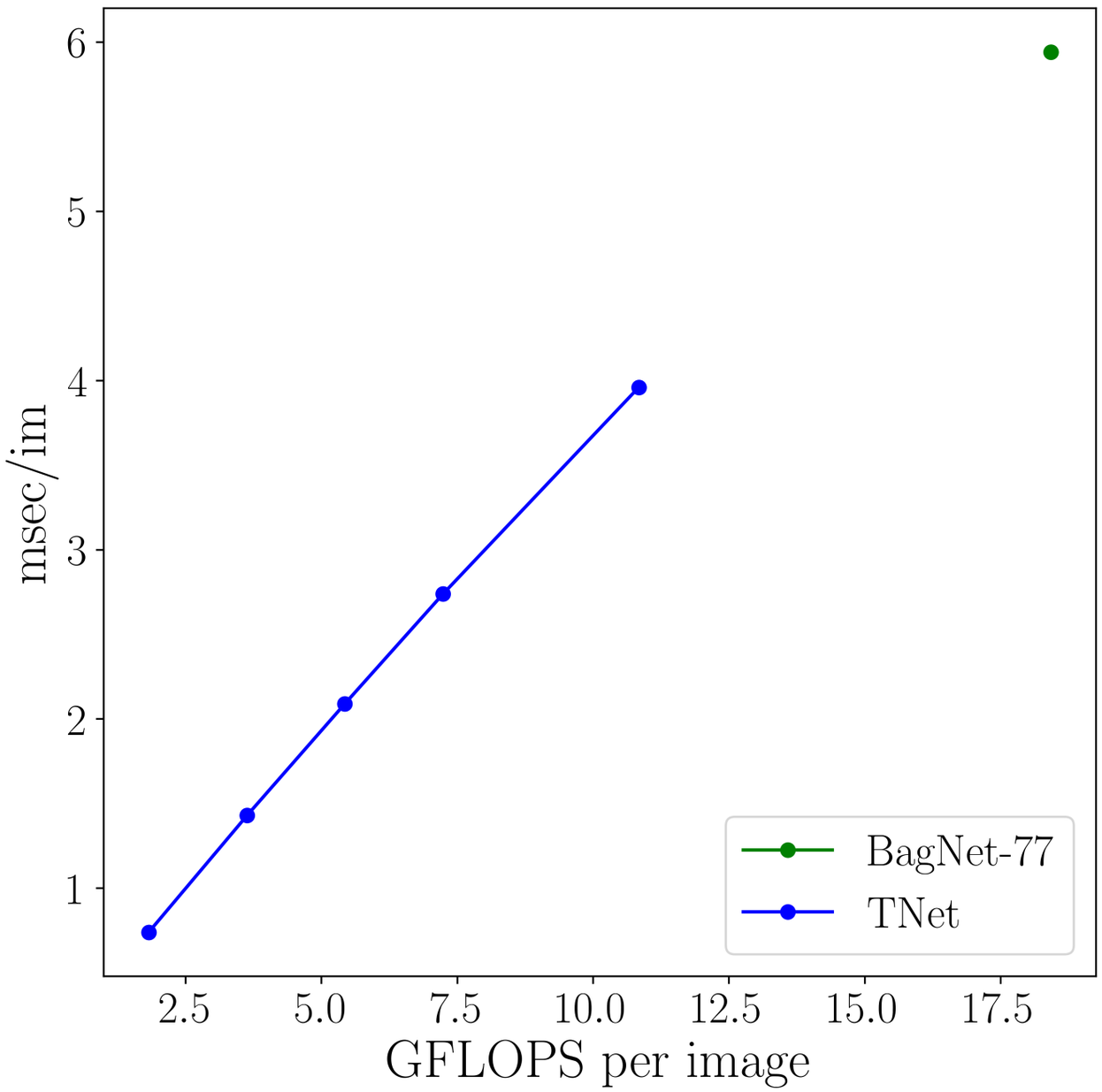
- Multi-crop data augmentation
- Less co-adaptation

Effectiveness of Hard-Attention





Effectiveness of Hard-Attention



Effectiveness of Hard-Attention

Model	#Locs	Top-1 Acc.	Top-5 Acc.	FLOPs (B)	#Params (M)	Time (msec/im)	Memory (GB)
Saccader	2	67.79%	85.42%	19.5	35.58	7.31 ± 1.55	2.58
	6	70.31%	87.8%	19.5		7.32 ± 1.55	2.52
	20	72.02%	89.51%	19.51		7.36 ± 1.50	2.53
	30	72.27%	89.79%	19.51		7.36 ± 1.49	2.51
DRAM	2	49.72%	73.27%	1.86	45.61	3.43 ± 1.57	0.45
	4	64.26%	84.84%	3.1		3.92 ± 1.56	0.44
	8	67.5%	86.6%	5.58		4.61 ± 1.59	0.45
	20	65.15%	84.58%	13.03		7.58 ± 1.53	0.46
BagNet-77	-	73.42%	91.1%	18.42	20.55	5.94 ± 0.09	2.62
TNet	0	67.29%	87.38%	1.82	21.86	0.74 ± 0.01	0.46
	1	73.12%	90.56%	3.63		1.43 ± 0.01	0.57
	2	74.12%	91.18%	5.43		2.09 ± 0.02	0.69
	3	74.41%	91.4%	7.24		2.74 ± 0.03	0.95
	5	74.62%	91.35%	10.84		3.96 ± 0.04	1.47

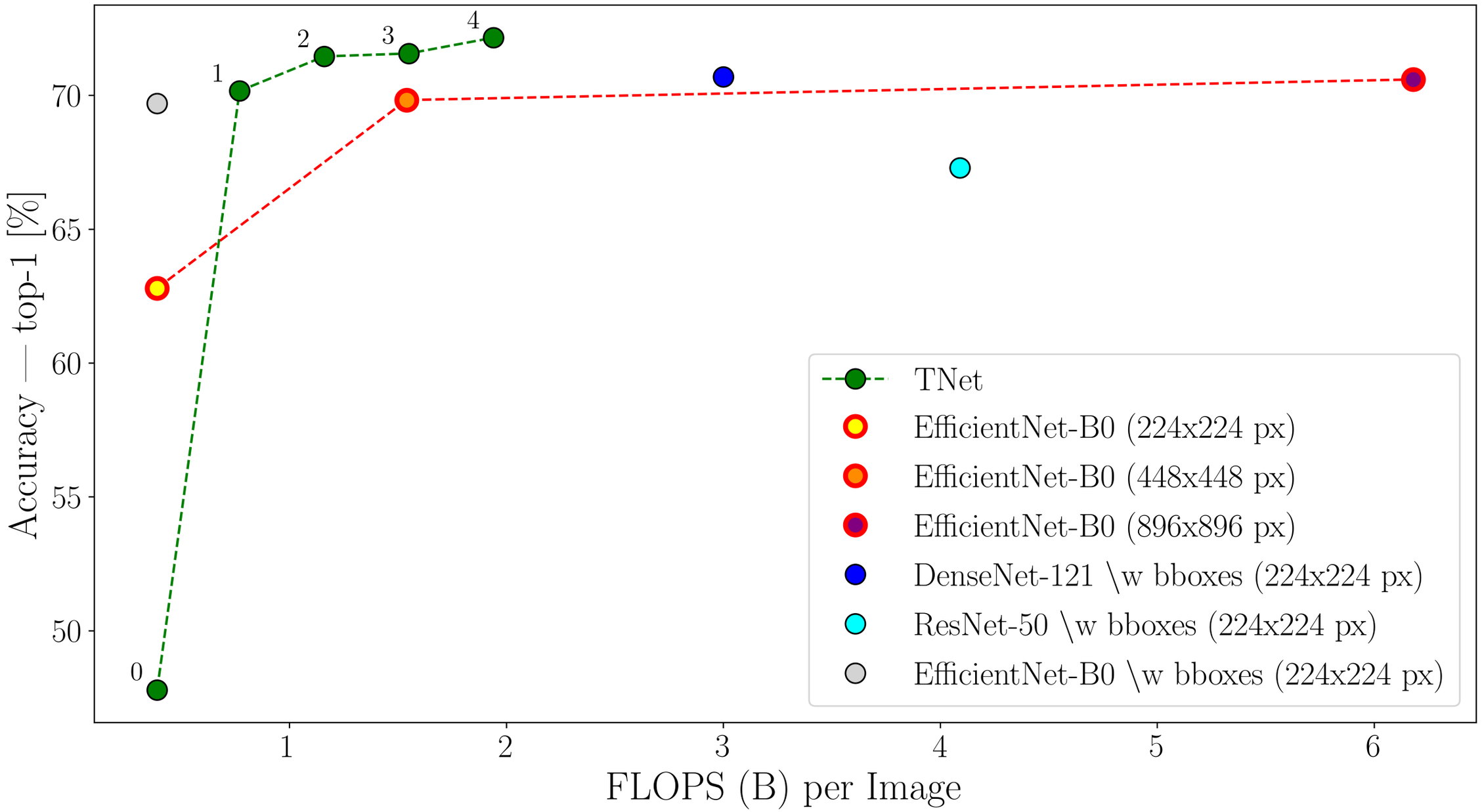
Effectiveness of Hard-Attention

Model	#Locs	Top-1 Acc.	Top-5 Acc.	FLOPs (B)	#Params (M)	Time (msec/im)	Memory (GB)
Saccader	2	67.79%	85.42%	19.5	35.58	7.31 ± 1.55	2.58
	6	70.31%	87.8%	19.5		7.32 ± 1.55	2.52
	20	72.02%	89.51%	19.51		7.36 ± 1.50	2.53
	30	72.27%	89.79%	19.51		7.36 ± 1.49	2.51
DRAM	2	49.72%	73.27%	1.86	45.61	3.43 ± 1.57	0.45
	4	64.26%	84.84%	3.1		3.92 ± 1.56	0.44
	8	67.5%	86.6%	5.58		4.61 ± 1.59	0.45
	20	65.15%	84.58%	13.03		7.58 ± 1.53	0.46
BagNet-77	-	73.42%	91.1%	18.42	20.55	5.94 ± 0.09	2.62
TNet	0	67.29%	87.38%	1.82	21.86	0.74 ± 0.01	0.46
	1	73.12%	90.56%	3.63		1.43 ± 0.01	0.57
	2	74.12%	91.18%	5.43		2.09 ± 0.02	0.69
	3	74.41%	91.4%	7.24		2.74 ± 0.03	0.95
	5	74.62%	91.35%	10.84		3.96 ± 0.04	1.47

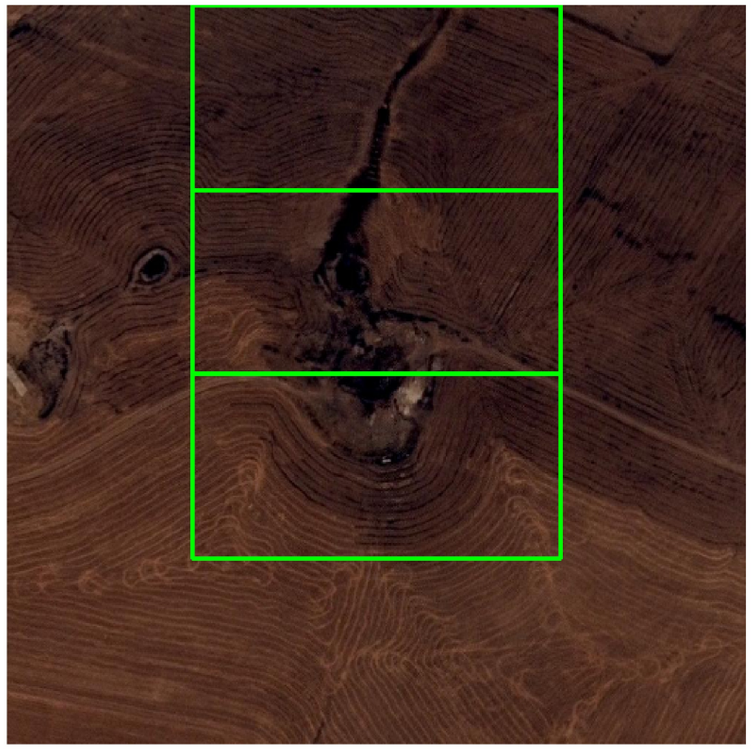
Scalability

- **Data:** fMoW – satellite images, bboxes
- **Baselines:**
    - EfficientNet-B0 + bboxes, 224x224 px
    - EfficientNet-B0, 224x224 px
    - EfficientNet-B0, 448x448 px
    - EfficientNet-B0, 896x896 px
- **TNet:** base resolution 224x224 px, 3x3 grid, 3 levels

Scalability



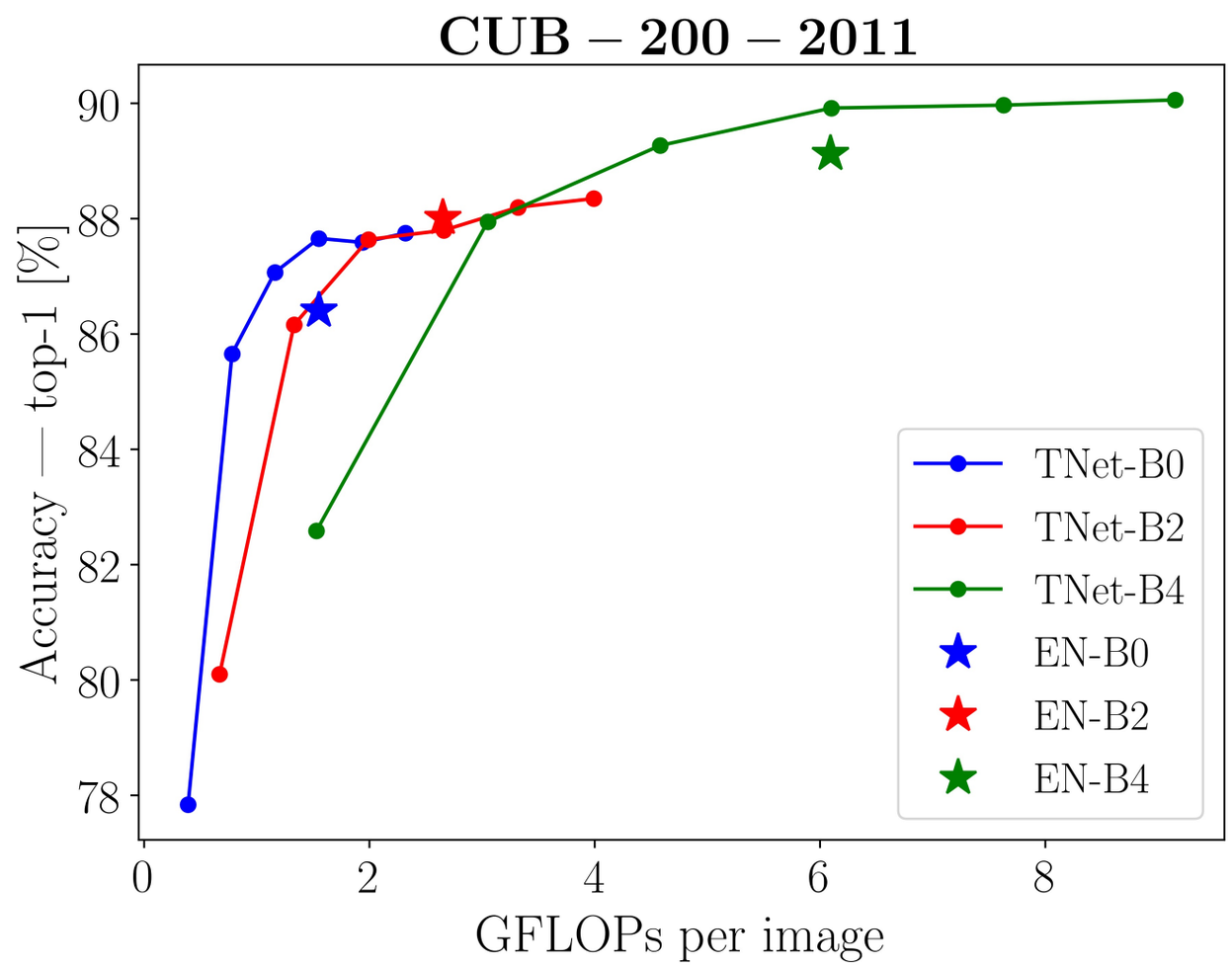
Scalability

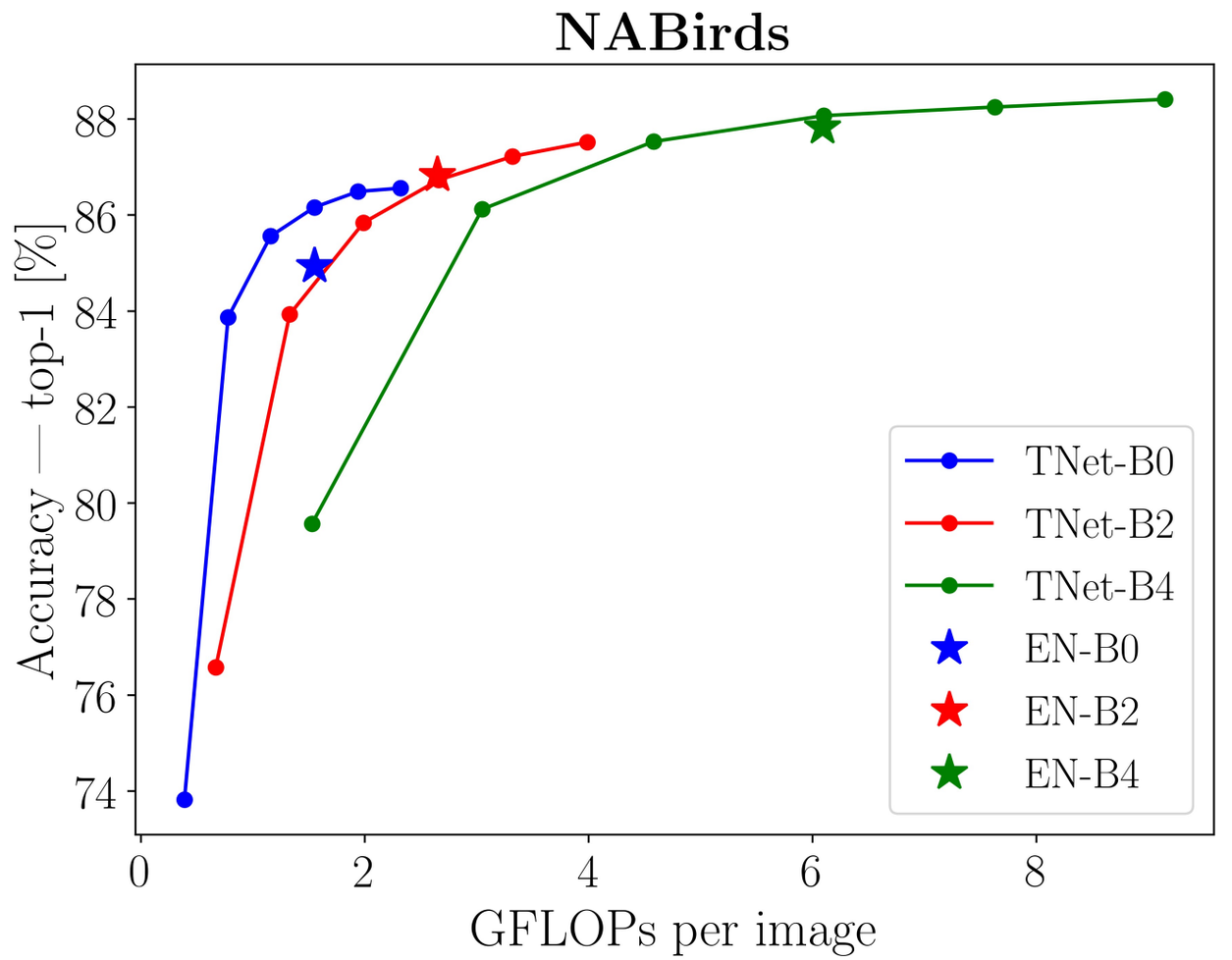


Modularity and Fine-Tuning

- **Data:** CUBirds (6K), NABirds (25K), 448x448 px
- **Baselines:** EfficientNet-B0 – B4
- **TNet:** base resolution 224x224 px, 5x5 grid, 2 levels, feature weighting module

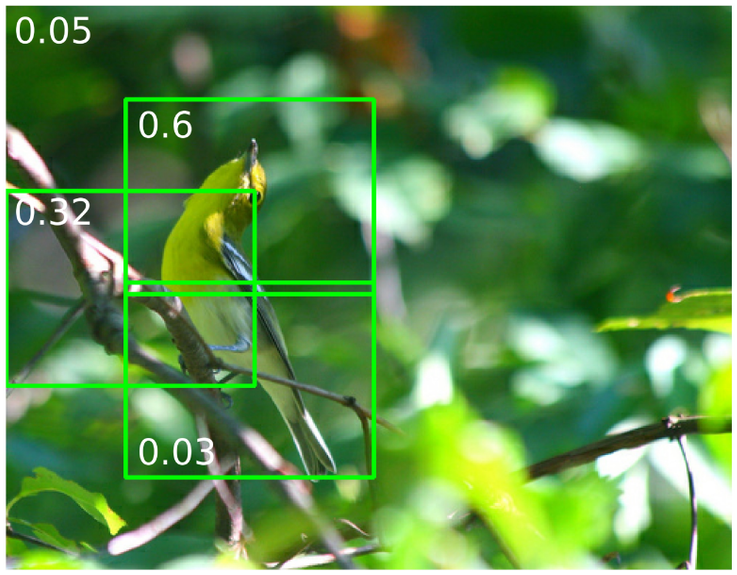
Modularity and Fine-Tuning

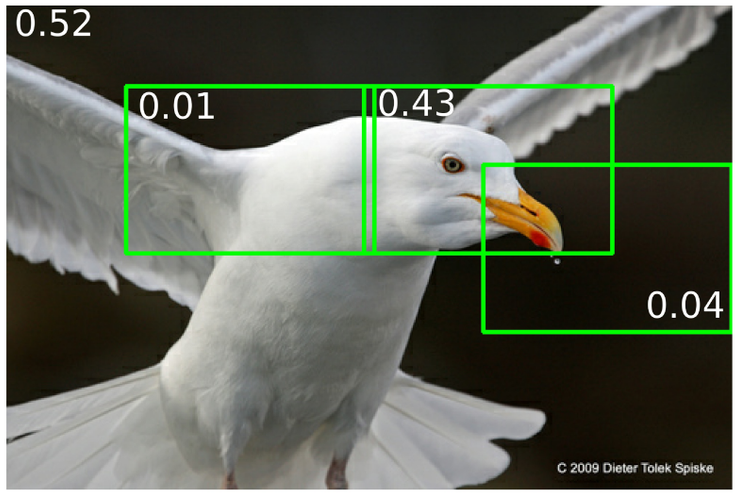




Interpretability

- Qualitative analysis





Interpretability

- Quantitative analysis
  - No context

Model	# Locs	Top-1 Acc.	Top-1 Acc. \w context
TNet	2	67.95%	74.12%
	3	69.66%	74.41%
	5	71.05%	74.62%
Saccader	2	67.79%	-
	4	69.51%	-
	8	70.80%	-

Limitations

- Design choices like attention grid size, are task-specific
- Strictly top-down processing
- Not learned policy on the number of locations
- Each image region is processed independently

Contributions

- We designed TNet, a novel multi-scale hard-attention architecture with adjustable processing
- We demonstrated the effectiveness of the hard-attention mechanism on ImageNet
- We demonstrated TNet's scalability on fMoW
- We showed that TNet is compatible with top-performing models, effectively learns from pre-trained weights
- Predictions inherently have a degree of interpretability, without extra cost beyond inference

Thank You!

- Code: https://github.com/Tpap/TNet

- Contact: Thanos Papadopoulos – tpapadop@nyu.edu