Dynamic Head: Unifying Object Detection Heads with Attentions

Xiyang Dai Yinpeng Chen Bin Xiao Dongdong Chen Mengchen Liu
Lu Yuan Lei Zhang
Microsoft
Redmond, USA

발표자 : 윤희원

DAVIAN
Data and Visual Analytics Lab

Introduction

- *The challenges in developing a good object detection head can be summarized into three categories*
  1. scale-aware.
  2. spatial-aware.
  3. task-aware.

Introduction

- *The challenges in developing a good object detection head can be summarized into three categories*
  1. scale-aware.
     - feature pyramid
     - path augmentation

  2. spatial-aware.
     - increasing the model capability
     - expensive data augmentations
     - new convolution operators

  3. task-aware.
     - Two-stage vs One-Stage
     - various representations of objects could potentially improve the performance

In this paper, we propose a novel detection head, called ***dynamic head****, to unify scale-awareness, spatial-awareness, and task-awareness all together*

Approach

Motivation

$$\mathcal{F} \in \mathcal{R}^{L \times H \times W \times C}$$

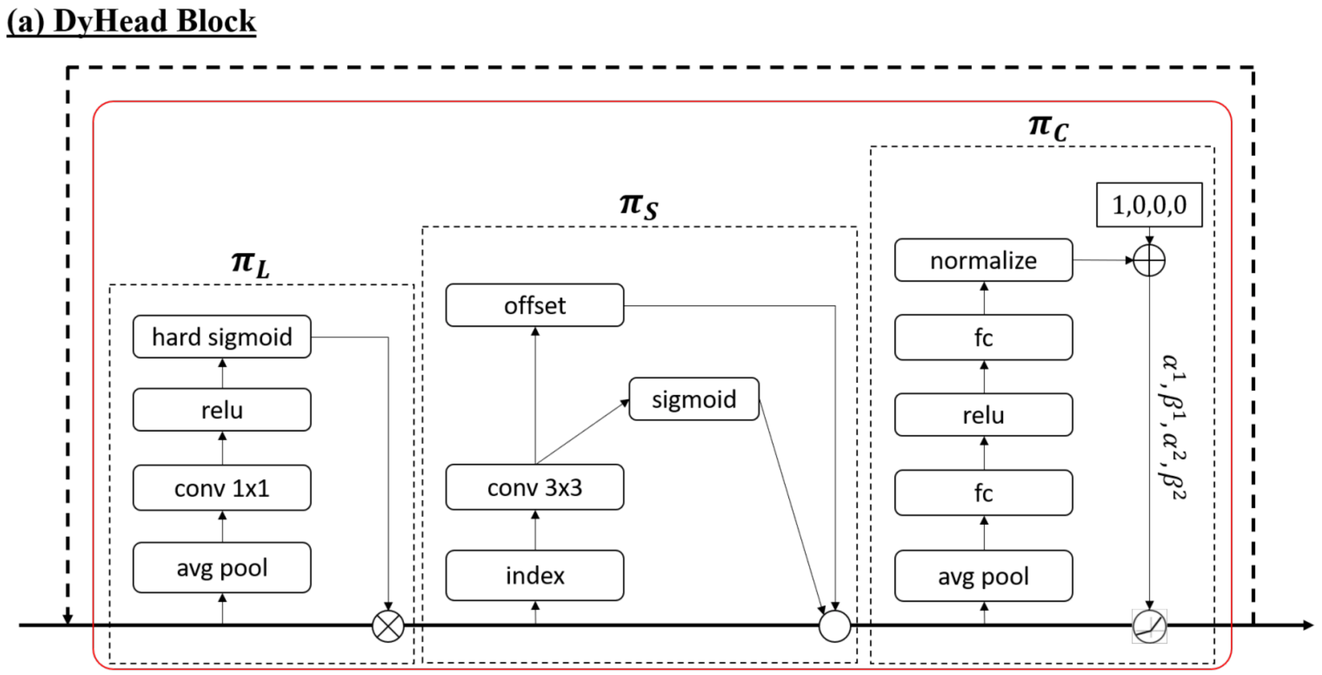
$$\mathcal{F} \in \mathcal{R}^{L \times S \times C}$$

Approach

Dynamic Head: Unifying with Attentions

$$W(\mathcal{F}) = \pi(\mathcal{F}) \cdot \mathcal{F} \tag{1}$$

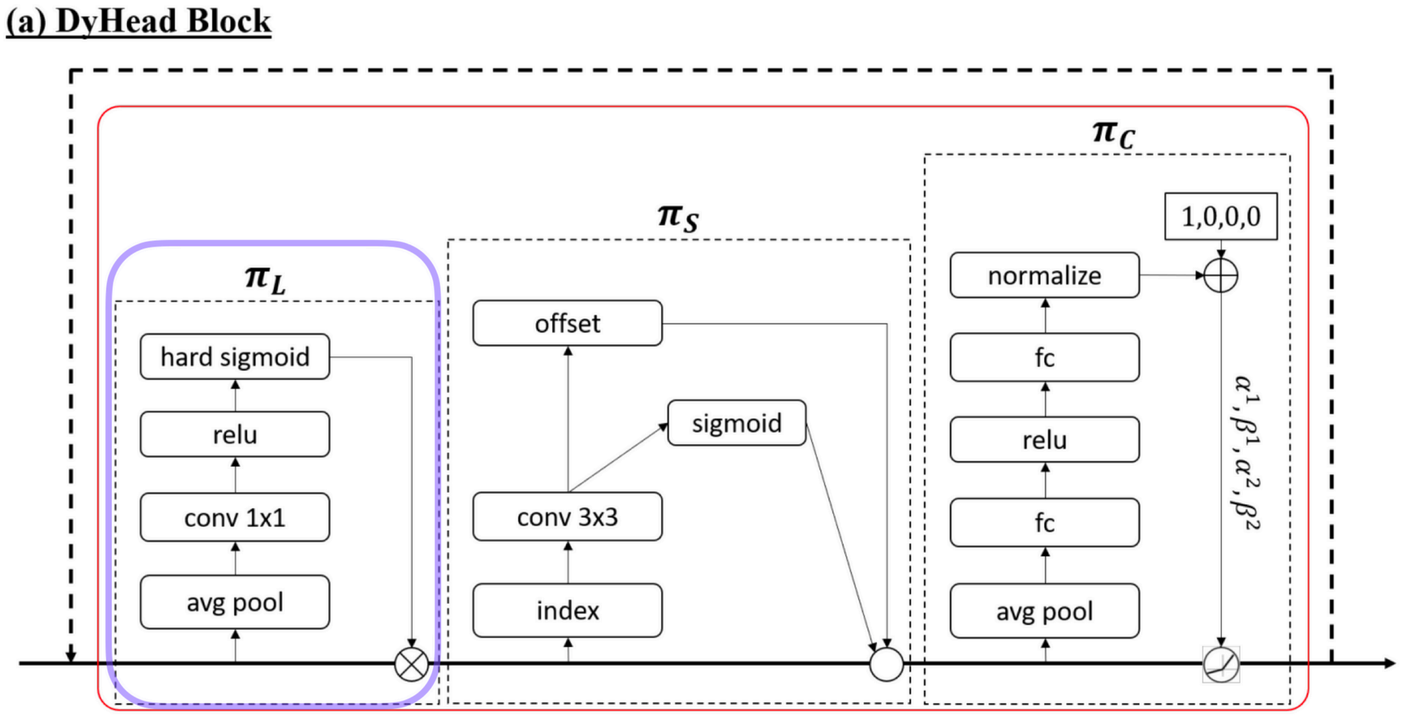
$$W(\mathcal{F}) = \pi_C\left(\pi_S\left(\pi_L(\mathcal{F}) \cdot \mathcal{F}\right) \cdot \mathcal{F}\right) \cdot \mathcal{F}, \tag{2}$$



Approach

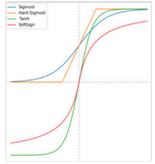
Dynamic Head: Unifying with Attentions

Scale-aware Attention



$$\pi_L(\mathcal{F}) \cdot \mathcal{F} = \sigma\left(f\left(\frac{1}{SC}\sum_{S,C}\mathcal{F}\right)\right) \cdot \mathcal{F} \qquad (3)$$

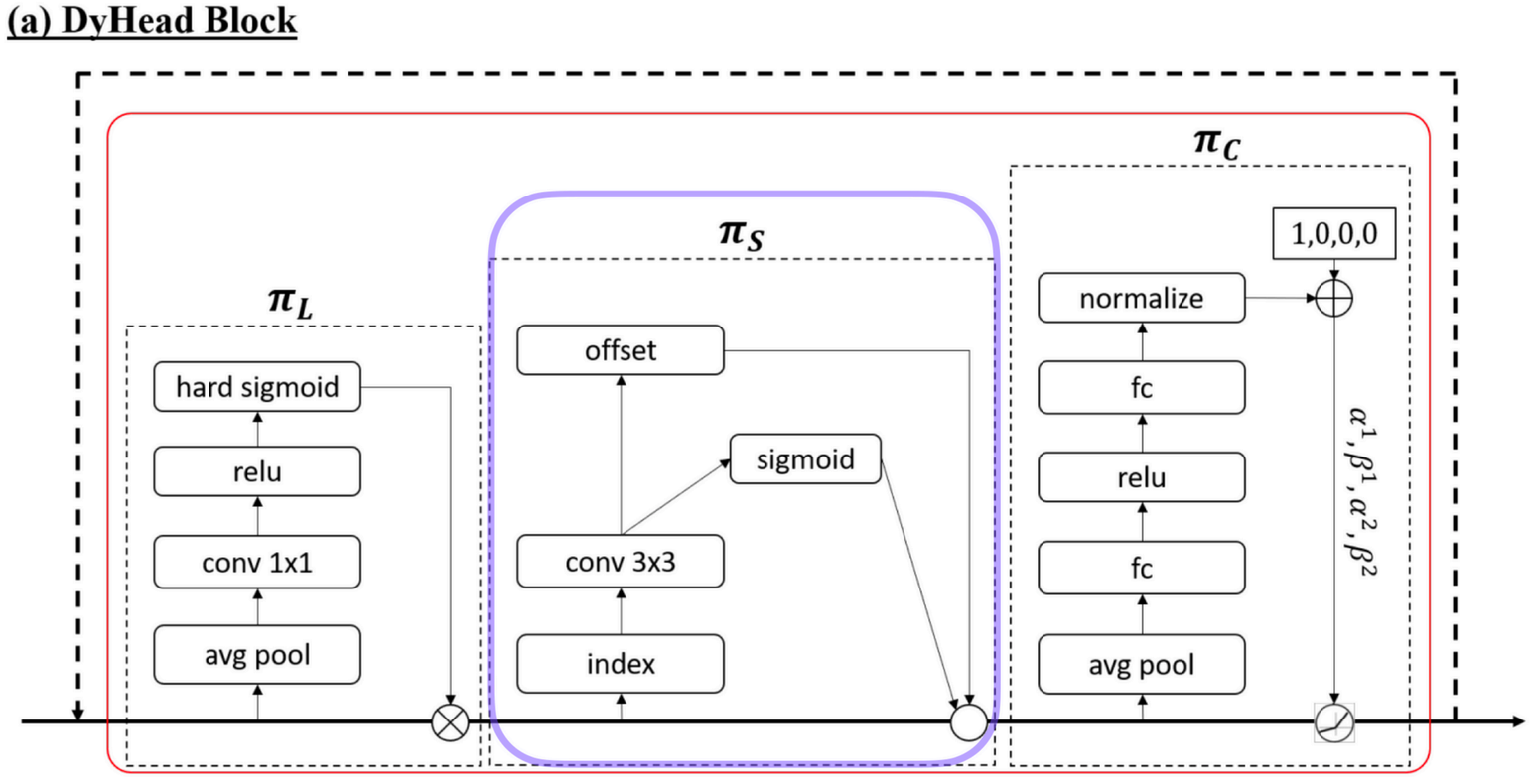
$$\sigma(x) = max(0, min(1, \tfrac{x+1}{2}))$$



Approach

Dynamic Head: Unifying with Attentions

Spatial-aware Attention

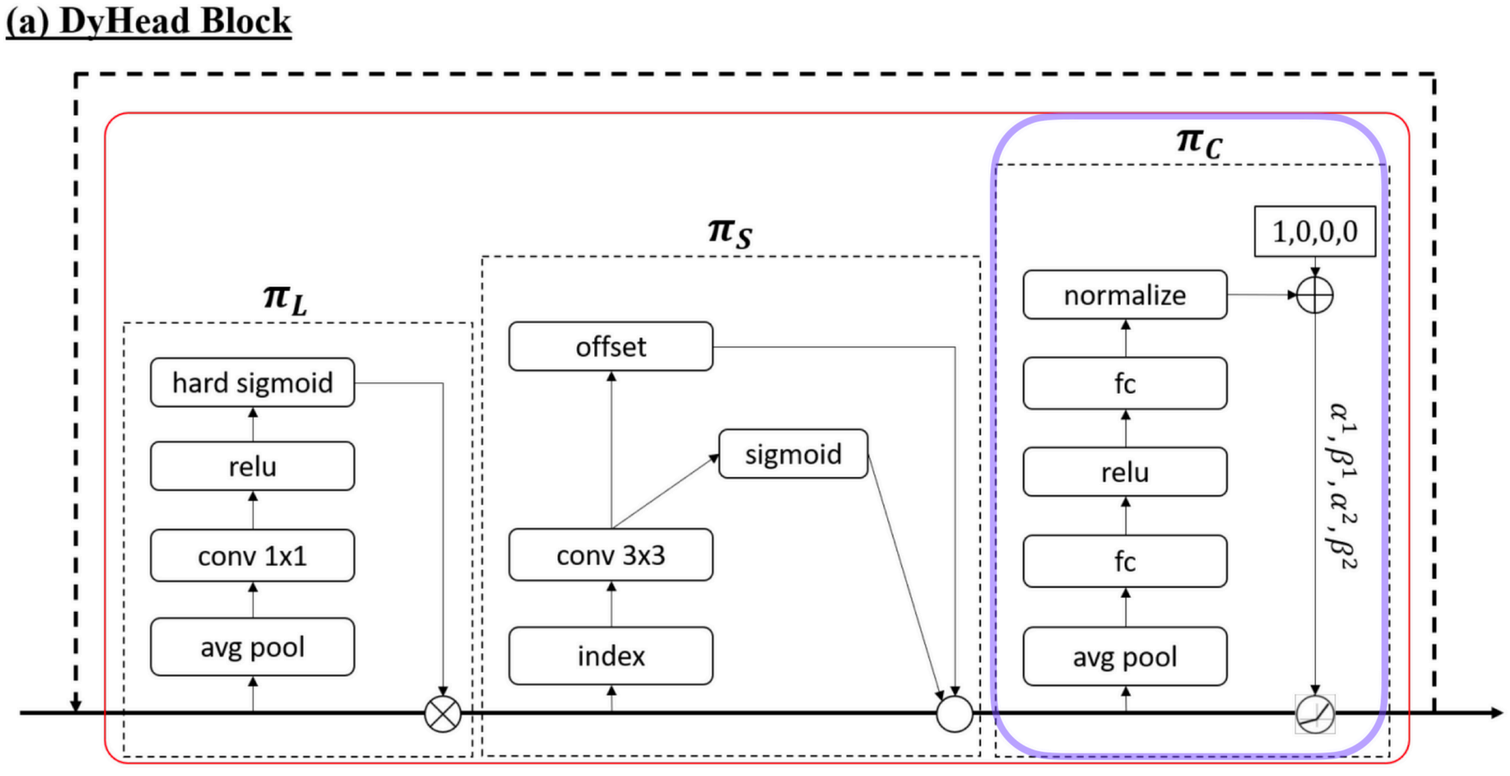


$$\pi_S(\mathcal{F}) \cdot \mathcal{F} = \frac{1}{L} \sum_{l=1}^{L} \sum_{k=1}^{K} w_{l,k} \cdot \mathcal{F}(l; p_k + \Delta p_k; c) \cdot \Delta m_k, \quad (4)$$

Approach

Dynamic Head: Unifying with Attentions

Task-aware Attention



$$\pi_C(\mathcal{F})\cdot\mathcal{F} = max\left(\alpha^1(\mathcal{F})\cdot\mathcal{F}_c+\beta^1(\mathcal{F}), \alpha^2(\mathcal{F})\cdot\mathcal{F}_c+\beta^2(\mathcal{F})\right) \tag{5}$$

Approach

Dynamic Head: Unifying with Attentions

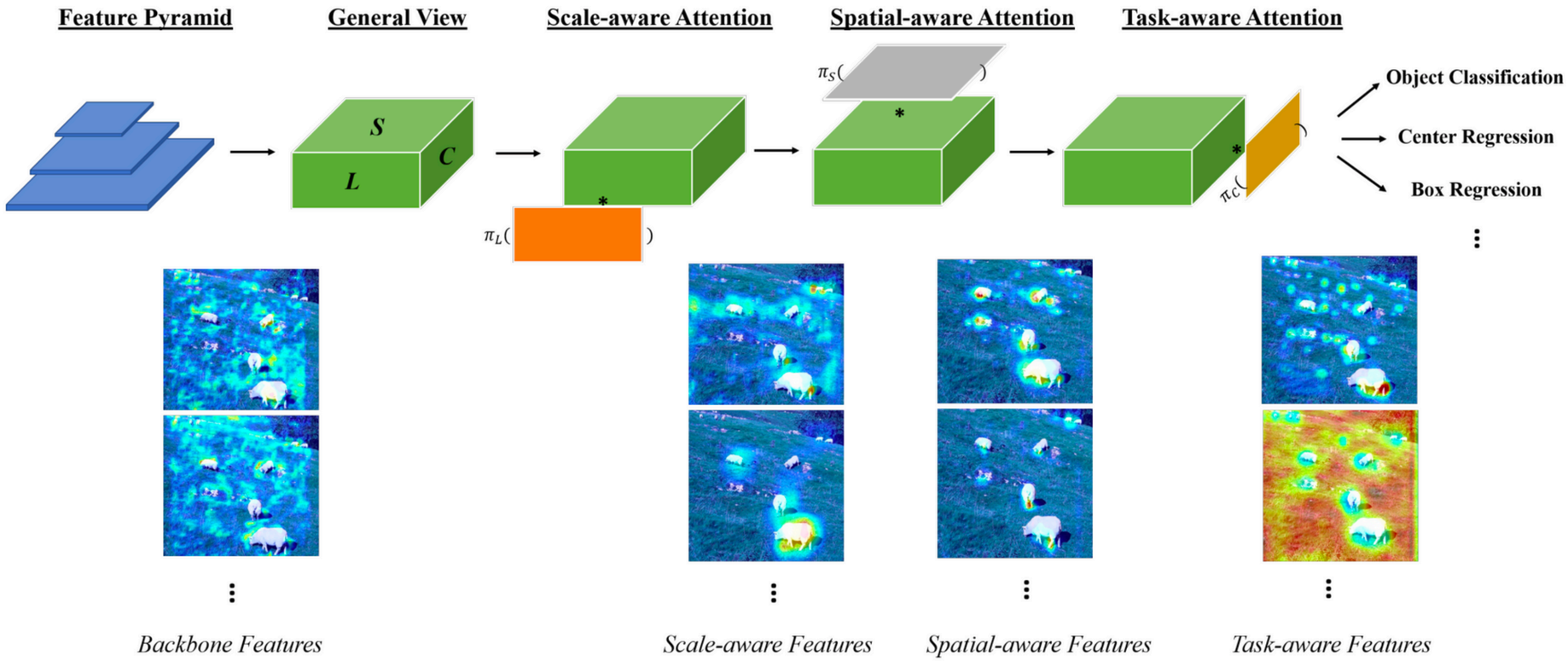
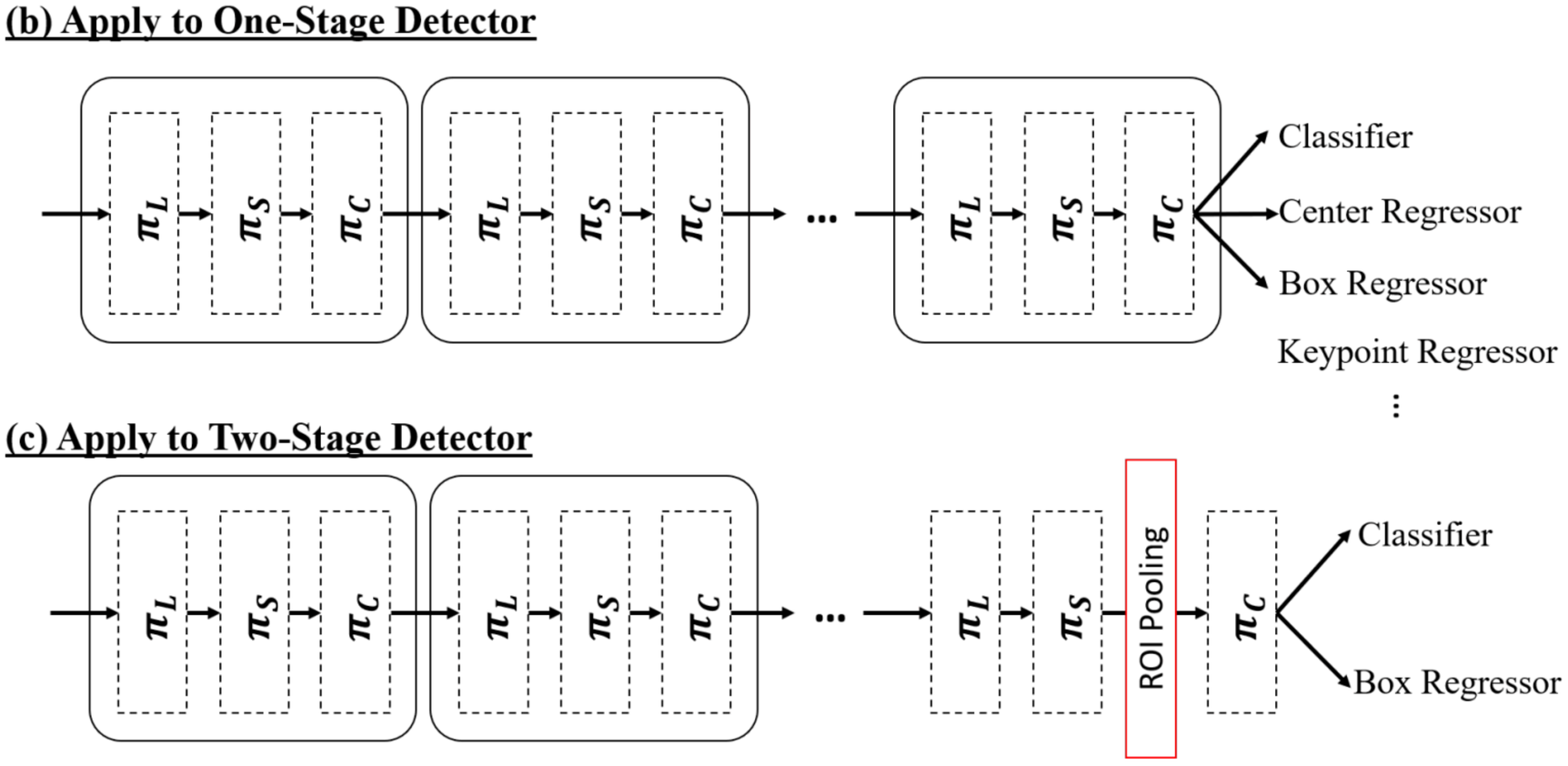


Figure 1. An illustration of our Dynamic Head approach. It contains three different attention mechanisms, each focusing on a different perspective: scale-aware attention, spatial-aware attention, and task-aware attention. We also visualize how the feature maps are improved after each attention module.

Approach

Generalizing to Existing Detectors



Experiments

Ablation Study

Effectiveness of Attention Modules.

L.	S.	C.	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
×	×	×	39.0	57.2	42.4	22.1	43.1	50.2
✓	×	×	39.9	57.8	43.5	25.4	44.0	52.4
×	✓	×	41.4	58.5	45.2	26.8	45.2	54.3
×	×	✓	40.3	58.3	43.9	24.2	44.6	53.7
×	✓	✓	42.0	59.5	45.5	25.5	46.1	55.2
✓	×	✓	40.6	58.6	44.4	24.6	44.8	53.3
✓	✓	×	41.9	59.2	45.6	24.8	46.1	54.5
✓	✓	✓	**42.6**	**60.1**	**46.4**	**26.1**	**46.8**	**56.0**

Table 1. Ablation study on the effectiveness of each attention module in our dynamic head block.

Experiments

Ablation Study

Effectiveness on Attention Learning.

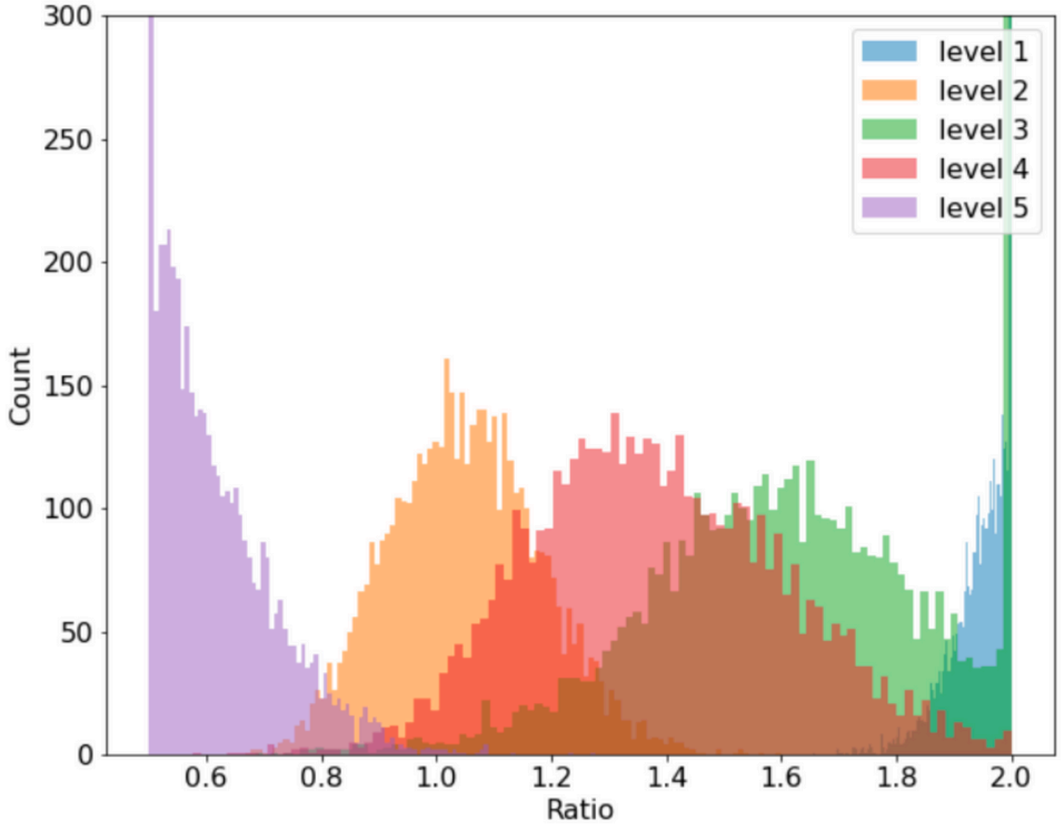


Figure 3. Ablation study on the effectiveness of our scale-aware attention module.

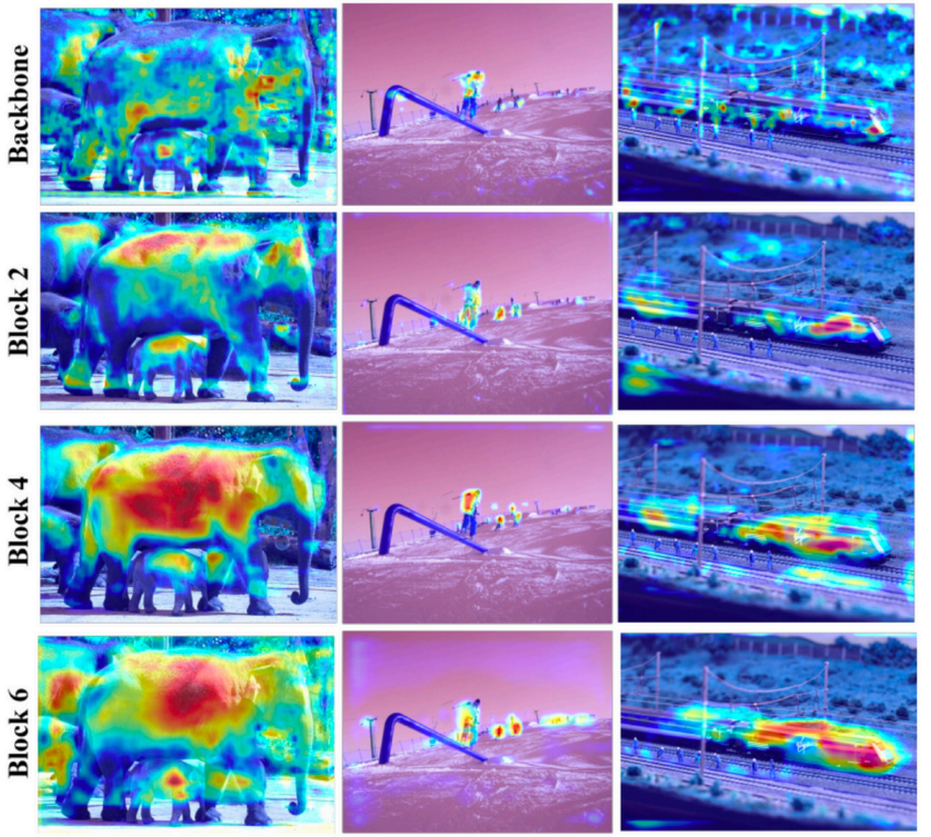


Figure 4. A visualization on the effectiveness of our spatial-aware attention module.

Experiments

Ablation Study

Efficiency on the Depth of Head.

#Block	GFLOPs	AP	AP_{50}	AP_{75}
Baseline	254.39	39.0	57.2	42.4
1	-84.69	36.7	55.5	40.0
2	-63.45	39.5	57.8	43.1
4	-20.97	42.0	59.9	45.9
6	+21.50	**42.6**	**60.1**	**46.4**
8	+63.98	42.5	59.6	46.1
10	+106.46	42.3	59.4	45.9

Table 2. Ablation study on the efficiency and effectiveness of stacking different number of dynamic head blocks.

Experiments

Ablation Study

Generalization on Existing Object Detectors.

Method	AP	AP_{50}	AP_{75}
anchor-based two-stage:			
Faster R-CNN [23]	36.4	57.9	39.4
+ DyHead	**38.9**	**57.6**	**42.0**
anchor-based one-stage:			
RetinaNet [16]	35.7	54.3	37.9
+ DyHead	**38.4**	**57.5**	**41.3**
anchor-free box-based:			
ATSS [35]	39.4	57.5	42.9
+ DyHead	**42.6**	**60.1**	**46.4**
anchor-free center-based:			
FCOS [28]	38.8	57.3	41.9
+ DyHead	**40.0**	**58.2**	**43.4**
anchor-free keypoint-based:			
RepPoints [33]	38.2	59.7	40.7
+ DyHead	**39.6**	**59.8**	**42.8**

Table 3. Ablation study on the generalization of our dynamic head when applying to popular object detection methods.

Experiments

Comparison with the State of the Art

Cooperate with Different Backbones

Method	Backbone	Iteration	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
two-stage detector:								
Mask R-CNN[12]	ResNet-101	2x	38.2	60.3	41.7	20.1	41.1	50.2
Cascade-RCNN[1]	ResNet-50	3x	40.6	59.9	44.0	22.6	42.7	52.1
Cascade-RCNN[1]	ResNet-101	3x	42.8	62.1	46.3	23.7	45.5	55.2
one-stage detector:								
FCOS[28]	ResNet-101	2x	41.5	60.7	45.0	24.4	44.8	51.6
FCOS[28]	ResNeXt-64x4d-101	2x	43.2	62.8	46.6	26.5	46.2	53.3
ATSS[35]	ResNet-101	2x	43.6	62.1	47.4	26.1	47.0	53.6
ATSS[35]	ResNeXt-64x4d-101	2x	45.6	64.6	49.7	28.5	48.9	55.6
BorderDet[21]	ResNet-101	1x	43.2	62.1	46.7	24.4	46.3	54.9
BorderDet[21]	ResNet-101	2x	45.4	64.1	48.8	26.7	48.3	56.5
BorderDet[21]	ResNeXt-64x4d-101	2x	46.5	65.7	50.5	29.1	49.4	57.5
DyHead	ResNet-50	1x	43.0	60.7	46.8	24.7	46.4	53.9
DyHead	ResNet-101	2x	46.5	64.5	50.7	28.3	50.3	57.5
DyHead	ResNeXt-64x4d-101	2x	47.7	65.7	51.9	31.5	51.7	60.7

Table 4. Comparison with results using different backbones on the MS COCO test-dev set

Experiments

Comparison with the State of the Art

Compared to State-of-the-Art Detectors.

Method	Backbone	Iteration	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
multi-scale training:								
ATSS[35]	ResNeXt-64x4d-101-DCN	2x	47.7	66.5	51.9	29.7	50.8	59.4
SEPC[31]	ResNeXt-64x4d-101-DCN	2x	50.1	69.8	54.3	31.3	53.3	63.7
BorderDet[21]	ResNeXt-64x4d-101-DCN	2x	48.0	67.1	52.1	29.4	50.7	60.5
RepPoints v2[4]	ResNeXt-64x4d-101-DCN	2x	49.4	68.9	53.4	30.3	52.1	62.3
RelationNet++[5]	ResNeXt-64x4d-101-DCN	2x	50.3	69.0	55.0	32.8	55.0	65.8
DETR[2]	ResNet-101	~25x	44.9	64.7	47.7	23.7	49.5	62.3
Deformable DETR[38]	ResNeXt-64x4d-101-DCN	~4x	50.1	69.7	54.6	30.6	52.8	64.7
EfficientDet[27]	Efficient-B7	~50x	52.2	71.4	56.3	–	–	–
SpineNet[8]	SpineNet-190	~40x	52.1	71.8	56.5	35.4	55.0	63.6
DyHead	ResNeXt-64x4d-101-DCN	2x	**52.3**	**70.7**	**57.2**	**35.1**	**56.2**	**63.4**
multi-scale training and multi-scale testing:								
ATSS[35]	ResNeXt-64x4d-101-DCN	2x	50.7	68.9	56.3	33.2	52.9	62.4
BorderDet[21]	ResNeXt-64x4d-101-DCN	2x	50.3	68.9	55.2	32.8	52.8	62.3
RepPoints v2[4]	ResNeXt-64x4d-101-DCN	2x	52.1	70.1	57.5	34.5	54.6	63.6
Deformable DETR[38]	ResNeXt-64x4d-101-DCN	~4x	52.3	71.9	58.1	34.4	54.4	65.6
RelationNet++[5]	ResNeXt-64x4d-101-DCN	2x	52.7	70.4	58.3	35.8	55.3	64.7
DyHead	ResNeXt-64x4d-101-DCN	2x	**54.0**	**72.1**	**59.3**	**37.1**	**57.2**	**66.3**

Table 5. Comparison with the state-of-the-art results on the MS COCO test-dev set

Thank you !