Author Rebuttal for NeurIPS 2020 Submission #3770

We thank all the reviewers for their valuable comments and suggestions. To improve readability, we will add a section of the related work. Besides, we will carefully revise the manuscript and describe the idea more clear according to the suggestions. We respond to the main concerns as follows. All the source code will be released to the community soon.

Q: *The authors should report the actual inference time/latency of their models.*

A: We evaluate the proposed method on a Tesla V100 GPU. Since our method only modifies the FPN head, we report the latency and the computational complexity of that part in Tab. 5. The result demonstrates the efficiency of our method on GPUs even with unfriendly sparse operations. Besides, our method has great potential to further improve efficiency by leveraging a sophisticated optimization or using specialized hardware accelerators for sparse operations (*e.g.*, SCNN, Cambricon-X, EIE and Eyeriss V2). We will clarify it and report the empirical runtime in the final version.

Table 5: The latency and computational complexity of the FPN heads on a Tesla V100 GPU. 'DH' is the dynamic head.

Model	DH	mAP(%)	Latency$_{avg}$(ms)	Latency$_{max}$(ms)	Latency$_{min}$(ms)	FLOPs$_{avg}$(G)
FCOS-D6 Baseline	✗	40.3	46.8	-	-	298.1
Ours@Large	✓	**41.0**	52.3	63.6	42.9	104.3
Ours@Small	✓	40.3	**34.9**	46.5	28.9	**70.0**

Response to Reviewer #1

Q1: *Some confusions on statements of symbols.*

A1: Sorry for the confusion. Specifically, K, $f^{l,k}$, and y_i^l represent the number of adjacent FPN scales, an adjacent feature map of node l and the feature vector of a pixel in the output of node l, respectively. We promise to revise the manuscript according to your suggestions carefully.

Q2: *What is the dynamic depth path in Fig.3a? Is it the same as the purple line in Fig.2?*

A2: Yes. The dynamic depth path in Fig.3a is the same as the purple line in Fig.2, whose spatial gate allows the network to just process a subset of locations. We will further explain this point in the final version.

Response to Reviewer #2

Q1: *The authors should discuss and compare with the "PointRend".*

A1: Thanks for your suggestion. Although the motivation has some similarities, "PointRend" mainly focuses on the boundary refinement in the instance segmentation task without high-level semantic enhancement. The reported performance (refer to the repo in GitHub) on object detection is also inferior to ours. We will add more discussions.

Q2: *It would be better to show results of FCOS by using SPConv+GN as the head.*

A2: The SPConv is identical to the regular convolution when enabling all the locations, and the GN is adopted by default. Please refer to Tab.1, Tab.2 and Sec.3.1 for more details. We will clarify it in the final version.

Response to Reviewer #3

Q1: *Some concerns about the novelty.*

A1: Sorry for the misunderstanding. Different from the dense prediction in the semantic segmentation task, the prediction of object detection is encouraged to be sparse in space, *e.g.*, the number of foreground samples (1.17% for FCOS on COCO dataset) is much less than that of background. Therefore, a large amount of computation in space is redundant and can be eliminated by the pixel-level dynamic routing. Meanwhile, pixel-level aggregation can handle the small but representative sub-regions of an instance, which is crucial to high-quality object detection. However, *these properties could not be achieved by the feature-level routing methods, e.g., [20].* This paper is the first to introduce the pixel-level dynamic routing mechanism into object detection, which brings stronger multi-scale representation with less computational complexity. Besides, the pixel-level extension is not straightforward. It needs to consider some extra aspects, *e.g.*, the change of the receptive field and the effect of regional connectivity on efficiency (refer to Sec.2.3).

Q2: *Some concerns about the FCOS baselines.*

A2: Our implementation for FCOS does not use the centerness-weighted trick for the regression loss, which can obtain 0.5 mAP absolute gains. We will update the results with this trick in the final version. The baseline head of FCOS-D{2,4,6,8} adopts the same architecture as dynamic head with depth and scale paths, but without spatial gates.

Q3: *The improvement of the proposed activation function is small compared to Restricted Tanh.*

A3: Please refer to "A1" in the section "Response to Reviewer #4".

Response to Reviewer #4

Q1: *The value of the new activation function as in Table 2 is weak.*

A1: Thanks for this comment. Our major contribution is to introduce the pixel-level dynamic routing mechanism into object detection, which enhances sub-region features of an instance with less computational complexity. The new activation function can provide a better and more generic solution to realize end-to-end training for spatial gates.