Responses to Reviewer #2

Q1: *Please provide computational cost of searching networks in Tab4&5.*

A1: In Table 4, NATS on R101 and X101-32×4d take 27 and 36 GPU-days respectively. In Table 5, NATS-A, NATS-B and NATS-C on R50 take 15, 17 and 20 GPU-days respectively. Generally, larger base-networks or more genotypes would require more search time.

Q2: *Please explain about fixation of arch-parameters for the first 10 epochs.*

A2: We find in experiments that searching without fixing architecture for awhile may lead the hyper-net to converge to sub-optimal state, because random paths may take over and prevent other paths to learn in early stage. Fixing arch-parameters for several epochs greatly alleviate the problem. 10 epochs of arch-fixation is an appropriate option.

Q3: *Please provide the outcome of optimization and explain.*

A3: We list one of the optimized network architecture in the supplementary material. As shown in Table1 in the supplementary material, each convolution tends to have various dilation types and large dilations seem to be helpful. We infer that it is because the detector has to deal with objects of large scale variation.

Responses to Reviewer #4

Q1: *Please provide results about efficiency of the discovered models.*

A1: As shown in Table A, B-G16 only takes 6 extra ms but yields 2AP improvement compared to baseline.

Q2: *The novelty of the proposed method is limited.*

A2: To our best knowledge, we are the first to explore this group-wise search space. 1) **Efficiency.** Our search space is compatible with searching and re-training with pre-trained models, which improves the search efficiency to a large extent. 2) **Effectiveness.** Our search space is proved to effectively improve the performance in task of object detection. Besides, a very recent work[2] of Google Brain has also verified the effectiveness of this search space in image classification, while we are studying it **before them** in this more complicated object detection task.

Table A: Inference time of R50 backbones. **B-Gn** means the backbone is searched with group number **n**.

Backbone	B(Baseline)	B-G1	B-G2	B-G4	B-G8	B-G16	B-G32
Inference Time(ms)	42	44	45	47	48	48	48

Responses to Reviewer #5

Q1: *Please improve the presentation in methodology.*

A1: We are sorry for our unclear presentation. 1) L148-149: the meaning of i is index of the i-th channel group; 2) Equation 2: C_{out} means the output channel of a path and C_i^g means the output channel of the g-th genotype in its i-th channel group; 3) We totally agree with your precious suggestion, and ind_i is now defined as $ind_i = \arg\max_g \alpha_i^g$; 4) With the definition of ind_i, intensity of each genotype I^g is defined as equation 4. And the output channel of y^g is obtained as $C^g = C_{out} I^g$. We would further improve our presentation to make our paper better. Thank you.

Q2: *About the influence of pre-training in object detection.*

A2:

1) Pre-training in detection. We does not claim that *object detection must use pre-training* and explain the influence of pre-training in object detection in L38-40. As explored in [1], training from scratch in object detection is **feasible but requires multi-fold extra training time** to reach a comparable performance. We show our results to support this conclusion in Table B. In [3] you mentioned(we would add it in reference), detectors are also trained with multi-fold training time(84.6 vs. 29.7 hours) which is in accordance with our point.

2) Reduce the search time. Learning from scratch in NAS would require even more time while our search space is compatible with searching based on pre-training which greatly accelerates the search of object detector.

Table B: Faster-RCNN with FPN of different training schedules. $\mathbf{n}\times$: $n \times 13$ training epochs. **ft**: finetuning.

Backbone	R50-1x-scratch	R50-1x-ft	R50-2x-scratch	R50-2x-ft	R50-6x-scratch	R50-6x-ft
COCO-AP	33.2	36.4	34.5	37.8	37.9	38.0

Q3: *Please compare channel-level search with path-level search.*

A3: The result of path-level search in listed in the second row of Table 2 in our paper. Please refer to L214-L216. It is shown that path-level search in this setting is less effective. We infer that a single dilation type for each layer might be insufficient to handle the huge scale variation of objects compared to mixed dilation types.

References

[1] K. He, R. Girshick, and P. Dollár. Rethinking imagenet pre-training. *arXiv preprint arXiv:1811.08883*, 2018.

[2] M. Tan and Q. V. Le. Mixnet: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*, 2019.

[3] R. Zhu, S. Zhang, X. Wang, L. Wen, H. Shi, L. Bo, and T. Mei. Scratchdet: Training single-shot object detectors from scratch. In *CVPR*, 2019.