Supplementary Material: Paying Attention to Varying Receptive Fields: Object Detection with Atrous Filters and Vision Transformers

Arthur Lam[1]
alam0015@student.monash.edu
JunYi Lim[2]
jun.lim@monash.edu
Ricky Sutopo[2]
ricky.sutopo@monash.edu
Vishnu Monn Baskaran[1]
vishnu.monn@monash.edu

[1] School of Information Technology
Monash University Malaysia
Selangor, Malaysia

[2] School of Engineering
Monash University Malaysia
Selangor, Malaysia

1 Summary of Related Works

We categorise and highlight the key contributions of past works in Table 1. Nevertheless, there still exists several common drawbacks that can be improved. To elaborate, most works attempt to alleviate the scale variation problem by convolving feature maps of different scales instead of varying and adjusting the receptive field on a feature map of specific scale which is less computationally expensive. In addition, feature fusion operations adds complexity to the process of scene understanding by introducing a higher degree of unimportant semantics such as background information or noise into the feature maps which can deteriorate detection performance. Thus, we propose a novel object detection architecture which unifies a multi-scale feature fusion module utilising atrous convolutions with a self-attention mechanism to enhance backbone features by applying strict attention onto important semantics.

2 Experiments

In this section, we include additional experiments that have been carried out to assess the performance of the proposed approach which we could not fit in the original paper.

Modularity of DIL. Firstly, we analyse the impact of utilising atrous filters (or DIL modules) in a multi-branch and multi-level scheme. Here, we performed extensive tests on several state-of-the-art object detectors by comparing the results obtained with and without the integration of the DIL module as a neck (i.e. between the backbone and detection layers). Table 2 tabulates the results before and after integrating DIL using the COCO evaluation metrics on the MS COCO val2017 validation set. From these tabulated results, we can observe

Category	Author	Framework	Highlights
One-stage	Liu et al. [8]	SSD	- Uses a set of default anchor boxes with different aspect ratios and scales to discretize the output space of bounding boxes
	Redmon et al. [10]	YOLO	- Predicts both confidences for multiple categories and bounding boxes from the whole topmost feature map in one evaluation
Two-stage	Girshick et al. [5]	RCNN	- Region proposals are extracted using selective search before detection in a CNN-based network
	Ren et al. [11]	Faster-RCNN	- Region proposals are extracted using a CNN-based network
	Cai & Vasconcelos [1]	Cascade-RCNN	- Uses a sequence of detectors trained with increasing IoU thresholds
Multi-scale	Lin et al. [6]	FPN	- Uses a pyramidal architecture with bottom-up and top-down pathways with lateral connections for feature aggregation
	Zhao et al. [13]	M2Det	- Proposes a multi-level feature pyramid network using feature fusion and thinned u-shaped modules
	Tan et al. [12]	EfficientDet	- Proposes a weighted bi-directional feature pyramid network
	Chen et al. [3]	DeepLab	- Uses atrous spatial pyramid pooling (ASPP) to resample a given feature map at varying atrous rates prior to convolution
	Liu et al. [7]	RFBNet	- Proposed a convolutional block consisting of several multi-branch convolution layers with different kernels and dilated pooling layers
	Qiao et al. [9]	DetectoRS	- Proposes switchable atrous convolutions (SAC) in the backbone and a recursive feature pyramid (RFP) at the model neck
Transformer	Dosovitskiy et al. [4]	ViT	- Inputs a sequence of patches of an image into a transformer module for image classification
	Carion et al. [2]	DETR	- Uses a set-based global loss that forces unique predictions via bipartite matching and a transformer encoder-decoder architecture for box predictions

Table 1: An overview of the prominent works of object detection with deep learning

that the DIL module acting as the model neck potentially improves features of the backbone network before these features are channeled into the detection heads. In detail, the addition of the DIL module increases the average precision scores of the analysed object detectors anywhere between 0.7 and 2.1 percentage points. Although the addition of the DIL module increases the model complexity, results from Table 2 reveal that its impact on the computational overheads were minimal as the differences in inferencing speed (i.e., frame rate per second or FPS) with and without the DIL module were trivial. This showcases the advantages of DIL in generating substantially effective features and simultaneously maintaining computation efficiency with simple modularity across various object detectors.

Method	Backbone	DIL	FPS	Avg. Precision, IoU			Avg.Precision, Area		
				0.5:0.95	0.5	0.75	S	M	L
SSD300	VGG-16	No	41	25.6	43.8	26.3	6.8	27.8	42.2
SSD300	**VGG-16**	**Yes**	**27**	**27.4**	**45.4**	**28.4**	**8.8**	**29.3**	**43.9**
YOLOv3	DarkNet-53	No	57	27.9	49.2	28.3	10.5	30.1	43.8
YOLOv3	**DarkNet-53**	**Yes**	**41**	**29.4**	**51.1**	**29.8**	**11.5**	**32.4**	**45.1**
EfficientDet-D2	EfficientNet	No	15	42.1	61.2	44.6	23.7	47.4	58.4
EfficientDet-D2	**EfficientNet**	**Yes**	**9**	**44.2**	**64.6**	**47.2**	**24.9**	**50.1**	**60.6**
Faster-RCNN	ResNet-50-FPN	No	11	37.4	57.7	40.8	21.7	40.4	48.4
Faster-RCNN	**ResNet-50-FPN**	**Yes**	**6**	**38.9**	**59.6**	**42.3**	**22.7**	**42.7**	**49.7**
RetinaNet	ResNet-50-FPN	No	11	36.5	55.4	39.1	20.4	40.3	48.1
RetinaNet	**ResNet-50-FPN**	**Yes**	**6**	**38.4**	**57.0**	**40.4**	**20.7**	**41.3**	**49.8**
Mask-RCNN	ResNet-50-FPN	No	8	39.2	59.6	42.8	22.9	42.6	51.2
Mask-RCNN	**ResNet-50-FPN**	**Yes**	**6**	**40.0**	**60.6**	**43.6**	**23.2**	**43.8**	**51.8**
DetectoRS	Cascade + ResNet-50	No	4	47.4	65.6	51.5	29.0	51.5	62.4
DetectoRS	**Cascade + ResNet-50**	**Yes**	**3**	**48.1**	**66.7**	**52.1**	**29.4**	**52.3**	**63.3**

Table 2: Performance comparison of various models with the addition of our proposed DIL module on the MS COCO val2017 subset.

Scalability of DIL-ViT. As discussed in the paper, we propose repeatable layers in the DIL module to allow model scaling based on different resource contraints. Table 3 compares and illustrates the model accuracy and complexity for models with varying DIL layers. By simply increasing the number of DIL layers by one from the baseline (three layers), we observe that the resulting AP improves to 47.4%, delivering a gain of 0.9%. Therefore, this verifies that DIL can be effectively scaled up with minimal increase to the model parameters.

Layers	Params (M)	FLOPs (G)	Param ratio	FLOPs ratio	Avg. Precision, IoU			Avg.Precision, Area		
					0.5:0.95	0.5	0.75	S	M	L
2	411.56	292.25	0.99x	0.98x	45.1	64.3	49.3	28.3	49.1	58.5
3	412.19	296.73	1x	1x	46.5	65.1	50.6	26.7	49.6	59.4
4	412.83	301.21	1.01x	1.02x	47.4	65.9	52.2	27.1	51.3	61.2

Table 3: Comparisons of the model with varying DIL layers where the param ratio and FLOPs ratio are calculated with respect to the baseline of 3 layers.

Qualitative Analysis. To provide deeper insights on the performance of the network, we compare the proposed model against other one-stage and two-stage detectors by visualizing the inferred bounding box outputs. Furthermore, we also visualize the heatmaps produced by the self-attention mechanism in ViT. This is to illustrate how the model is able to focus on the important features of objects and put less emphasis on irrelevant information.

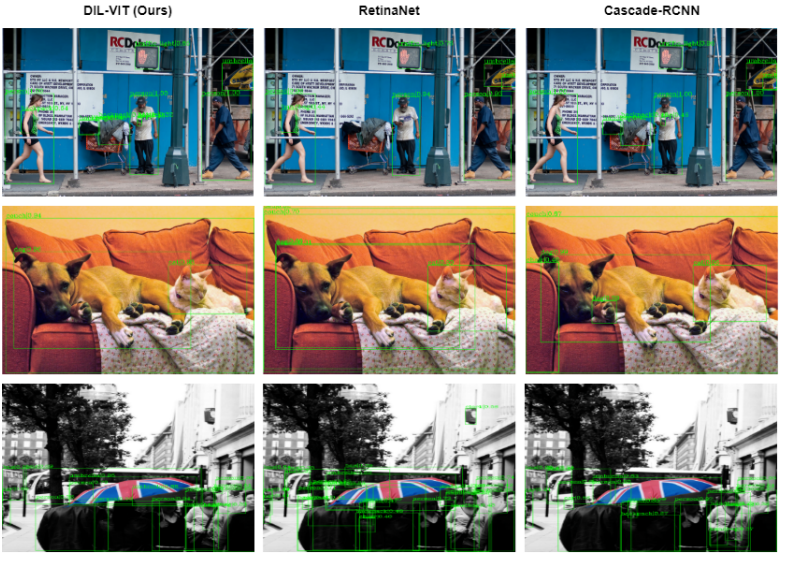

Figure 1: Qualitative results and comparisons between RetinaNet and Cascade-RCNN against DIL-ViT on the MS-COCO val2017 subset.

Figure 1 illustrates the ability of DIL-ViT to accurately localize and identify objects of varying sizes in the image. Specifically, the first and third rows of Fig. 1 depict the improved capability of the proposed model in detecting smaller objects in cluttered scenes. Furthermore, it can be observed that DIL-ViT produces fewer false positives when compared to other object detectors. These results demonstrate that varying the receptive field of the model allows the model to learn better representations of objects with different aspect ratios and ignore irrelevant semantic information.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[7] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 385–400, 2018.

[8] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[9] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. *arXiv preprint arXiv:2006.02334*, 2020.

[10] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[12] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[13] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9259–9266, 2019.