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

Abstract

Object detection represents a critical component in computer vision based on its unique ability to identify the location of one or more objects in an image or video. Given its importance, various approaches were proposed in an attempt to extract meaningful and representative features across different image scales. One such approach would be to vary the receptive fields during the feature extraction process. However, varying and adjusting the receptive field adds complexity to the process of scene understanding by introducing a higher degree of unimportant semantics into the feature maps. To solve this problem, we propose a novel object detection framework by unifying dilation modules (or atrous convolutions) with a vision transformer (DIL-ViT). The proposed model leverages atrous convolutions to generate rich multi-scale feature maps and employs a self-attention mechanism to enrich important backbone features. Specifically, the dilation (i.e., DIL) module enables feature fusions across varying scales from a single input feature map of specific scales. Through this method, we incorporate coarse semantics and fine details into the feature maps by convolving the features with different atrous rates in a multi-branch multi-level structure. By embedding DIL into various object detectors, we observe notable improvements in all of the compared evaluation metrics using the MS-COCO dataset. To further enhance the feature maps produced by the DIL, we then apply channel-wise attention using a vision transformer (i.e., ViT). Crucially, this approach removes unnecessary semantics present in the fused multi-scale feature map. Experimental results of DIL-ViT on the MS-COCO dataset exhibit substantial improvements in all of the compared evaluation metrics.

1 Introduction

In the field of computer vision, object detection represents a fundamental method of detecting instances of semantic objects and its locations in an image or a sequence of images. This is critical in various domains which include human behavioral analytics, autonomous

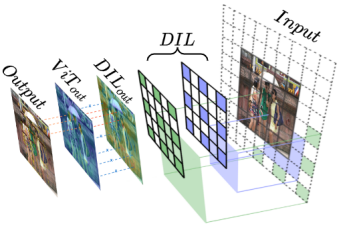

Figure 1: Overall illustration of DIL-ViT, which consists of atrous filters with different dilation rates to capture the varying receptive fields and channel-wise attention using ViT.

vehicles and face recognition. Given these benefits, a multitude of object detectors using deep neural networks has emerged in recent years, which in turn cements their position as an indispensable component in computer vision systems.

In spite of the advances made in the field of object detection, the task of extracting and recognizing features of objects at varying scales has always been a long-standing challenge. This is in part due to the fact that object detectors extract different features at different scales. To tackle this challenge, techniques were proposed by fusing feature maps of varying scales from specific levels of the backbone convolutional neural network (CNN) [13, 22, 27, 31, 35, 36]. These techniques were largely inspired by an image pyramid architecture [1] and therefore rely heavily on a feature fusion process which incorporates features from varying receptive fields. This process uses a fixed sampling grid during feature extraction, limiting the receptive field and, consequently produces features which are less fine-grained. In addition, the process of fusing feature maps of different scales in a feature pyramid structure potentially reduces the ability of a detector to learn scale-specific features contributed from each scale of the feature map [8].

Apart from that, multi-scale feature maps introduce unwanted or unimportant semantics. This raises the risk that the detector head takes into account irrelevant background features that do not contribute to the representation learning of object features. Recently, several works have emerged which utilize transformers for computer vision tasks. Although transformer models are more well-known for solving natural language processing (NLP) problems, researchers have extended and applied the necessary adjustments in order to exploit the strong representational ability of the self-attention mechanism in transformers for object detection. Previous work by [26] proposes to apply self-attention only in local neighbourhoods of an input image to differentiate important features over irrelevant background features. Alternatively, [9] takes a similar approach by proposing Sparse Transformers that employ scalable approximations to global self-attention onto images. However, these specialized attention architectures do not apply global self-attention onto full-sized images or feature maps, which is crucial for amplifying important foreground features.

Nevertheless, some of the most notable approaches of utilizing transformers for the object detection task include using only a pure transformer [11] without relying on a CNN. Even so, the transformer could still be embedded into any traditional deep learning framework by modifying the input to accept a feature map extracted directly from the backbone CNN. However, such a hybrid approach would require the process of splitting and flattening the feature map into a sequence of patches to be fed into the standard transformer module [4]. Despite that, the chosen inputs of these state-of-the-art models lack multi-scale information which can provide rich semantic information to fully utilize the self-attention mechanism.

To solve the aforementioned problems, in this paper, we propose a novel object detection

framework by unifying dilation modules (or atrous convolutions) with a vision transformer (DIL-ViT), as illustrated in Fig. 1. Here, the proposed model leverages atrous convolutions to generate rich multi-scale feature maps and employs a self-attention mechanism to enrich important backbone features. Specifically, the dilation (i.e., DIL) module enables feature fusions across varying scales from a single input feature map of specific scales. Through this method, we incorporate coarse semantics and fine details into the feature maps by convolving the features with different atrous rates in a multi-branch multi-level structure. To further enhance the feature maps produced by DIL, we apply channel-wise attention using a vision transformer (i.e., ViT). Crucially, this approach removes unnecessary semantics present in the fused multi-scale feature map. The contributions in this paper are summarized as follows:

- We first put forward a novel feature fusion module which produces fused multi-scale feature maps from a single feature map of specific scale. The feature fusion module, DIL possesses a multi-branch, multi-level architecture which aggregates multi-scale features through a series of branched atrous convolutions of varying dilation rates. We also demonstrate the modularity of DIL to improve baseline performances by integrating it into different architectures as the model neck.

- Next, we produce a variant of the Recursive Feature Pyramid (RFP) by integrating vision transformers (i.e., RFP-ViT) as the connecting module to recursively apply channel-wise attention onto the spatially distant semantic information present in the multi-scale fused feature maps produced in DIL.

- Then, we propose a model architecture, DIL-ViT which leverages the ability of atrous convolutions to vary the receptive field and a self-attention mechanism to effectively produce fused multi-scale feature maps with strict attention on important semantic information for better detection performance.

2 Related Works

Object Detection. Deep learning baselines for object detection can be mainly divided into two types of approaches, namely one-stage approaches and two-stage approaches. Firstly, one-stage approaches such as SSD [25] and YOLO [28] treat object detection as both a regression and classification task [37]. They perform object detection by learning the probabilities of object classes and the coordinates of the bounding boxes for a given input image [17]. Two-stage approaches utilize an additional mechanism called the Region Proposal Network (RPN) to generate regions of interests before the regression and classification step. Two-stage approaches such as the RCNN [14] and Faster-RCNN [29] achieve competitive results that outperform most one-stage detectors. Motivated by two-stage approaches, our proposed model adopts a two-stage approach with Cascade-RCNN [2] as the baseline architecture to achieve competitive results. A table summarizing the state-of-the-art approaches of object detection using deep learning is available in the supplemental material.

Multi-scale Features. To overcome the challenge of extracting and processing multi-scale features, early solutions utilize an image pyramid formed by the resized input image of varying scales. Although this approach improves detection accuracy, it is computationally expensive. [22] proposes a top-down pathway to enhance the pyramidal feature hierarchy created from backbone networks. Based on this approach, several works [2, 13, 24, 30, 36, 38] emerged which leverage multi-scale feature maps. However, most of these approaches inherently use backbone features for predictions [3, 25] or simply concatenate features without

attention [22, 24, 36]. To tackle this issue, [31] proposes a Bi-directional Feature Pyramid Network (BiFPN) which introduces learnable weights between cross-scale connections that penalize insignificant features. The drawback here is that these approaches rely on multiple feature maps with varying scale, which impacts the computational speed of the model during the feature fusion process.

To efficiently produce features which are more robust and representative of the foreground objects, parallel branching schemes have been adopted which utilize varying atrous rates or kernel sizes. Atrous Spatial Pyramid Pooling (ASPP) [6] proposes to employ several atrous convolutional layers with different sampling rates in parallel as a segmentation module to extract features. DetectoRS [27] takes a similar approach by applying dynamic convolutions to the network using switchable atrous convolution (SAC). However, the SAC module potentially disregards key semantics present in larger contexts as only a single extra dilation filter is applied [8]. Hence, we apply a greater emphasis on varying the receptive field using the DIL method proposed in this paper to capture complex object features.

Transformer Models. More recently, the transformer model [32], which is the de-facto network for natural language processing, has been adopted for computer vision tasks such as object classification, segmentation and detection. Here, ViT [11] applies a pure transformer model by splitting and flattening an input image into a sequence of smaller image patches before applying self-attention. Subsequently, [4] proposes an end-to-end transformer pipeline for object detection tasks (i.e., DETR) by using set predictions. The aforementioned approach further improves object localization, by supplementing flattened image features with additional positional encodings before being processed in the encoder stage. Nevertheless, DETR comes with several limitations such as longer training times and poor detection performance on small objects. Unlike previous works, our proposed vision transformer module accepts a multi-scale fused feature map as the input to the transformer to improve the performance of the self-attention mechanism on varying object scales. Furthermore, we produce a variant of the RFP architecture by embedding the transformer module onto the lateral connections. This way, the transformer module is able to recursively apply channel-wise attention onto the input feature maps to further enrich and produce fine-grained semantics.

3 Methodology

In this section, we first discuss the formulation of DIL and ViT, followed by how each of these components are embedded onto the overall network architecture, DIL-ViT.

3.1 Dilation Module

3.1.1 Atrous Convolutions

Although deeper networks with different kernel sizes are encouraged for accuracy and generating multi-scale receptive fields, it is inevitable that these repeated convolutions require more parameters to be trained. Recently, dilated convolutions have proven to be an effective method to resample features for enhancing multi-scale representation. To elaborate, atrous convolutions are able to effectively enlarge the receptive fields of filters and provide dense feature extraction [16] with no additional cost of increasing the number of network parameters. To illustrate the process, the output $y[i][j]$ of atrous convolution of a 2-D input signal $x[i][j]$ and $f[h][w]$ filter of size (H,W) could be formulated as in (1).

$$y[i][j] = \sum_{h=1}^{H} \sum_{w=1}^{W} x[i + r*h][j + r*w] f[h][w] \tag{1}$$

An additional parameter r denotes the dilation rate, which corresponds to the stride with which the input signal is sampled. By specifying r to be greater than 1, the algorithm stretches filters by a factor of r and introduces zeros in between filter values.

3.1.2 DIL Structure

Our proposed DIL module is constructed using parallel branches consisting of atrous convolutions with different atrous rates of r, repeated through multiple levels of l depending on the desired scale and complexity as shown in Fig. 2. First, the module starts by applying a 1×1 convolution to split up the input feature map X of arbitrary size into four parts (channel-wise, D_X) where the channels of the input to each branch is $1/4$ the number of channels ($D_X/4$) of the input to the module. Next, these inputs are fed into the respective branches where each branch b consist of a single 3×3 convolution with different atrous rates. The parameters for each convolution such as the dilation rate r_{bl} are set accordingly to craft the effective receptive field to improve learning of scale-specific features. To elaborate, the parameters that are set in the default setting of each branch is $rate = [1,2,3,6]$. Following that, the outputs of each parallel branch y_{bl} are then activated using a ReLU function before concatenation to form an intermediary feature map Y_i on each specific level. According to the number of levels repeated, multiple intermediary feature maps that contain multi-level semantics are obtained with very minimal computational cost [36]. Finally, to produce the final output feature map Y, the level-wise feature maps are concatenated and proceeded by a 1×1 convolution to compress and resize the features before the addition of a shortcut connection used to effectively propagate the gradients.

To summarise, the entire module structure can be expressed as shown in (2) where the output feature map X_{out} can be produced from the element-wise addition of the input to the ResNet block X_{in} and the output of DIL X_{DIL} where $Concat$ corresponds to a channel-wise concatenation and f_{bi} represents the corresponding atrous convolution operation for the particular branch and level, i.e., $\forall_b = 1,...,4$ and $\forall_i = 1,...,i$.

$$X_{DIL} = \forall_i Concat(\forall_b Concat(f_{bi}(X_i))), \tag{2}$$

To demonstrate the flexibility and modularity of DIL, we compare its performance as a neck across various object detection architectures. Additionally, we perform an empirical analysis on the scalability of DIL. Tabulation and analysis of these results are available in the supplemental material.

3.1.3 Dilation Module In The Backbone

For the overall architecture of our proposed dilation network, we employ the ResNet backbone and make several changes to its design to incorporate DIL. As seen in Fig. 2, the inputs for the first block of each ResNet stage are fed into both the original pipeline with shortcut connections and DIL. Next, the outputs of each stage are combined by element-wise addition followed by a ReLU activation function. This essentially transforms the original bottleneck block to combine feature maps from three different connections. Equation (3) presents the aforementioned process of feature combination, where y is the respective output of the corresponding bottleneck block, x is the input, $F(x)$ represents the series of convolutions employed in the original bottleneck block implementation and $DIL(x)$ represents the convolutions applied in DIL.

$$y = x + F(x) + DIL(x) \tag{3}$$

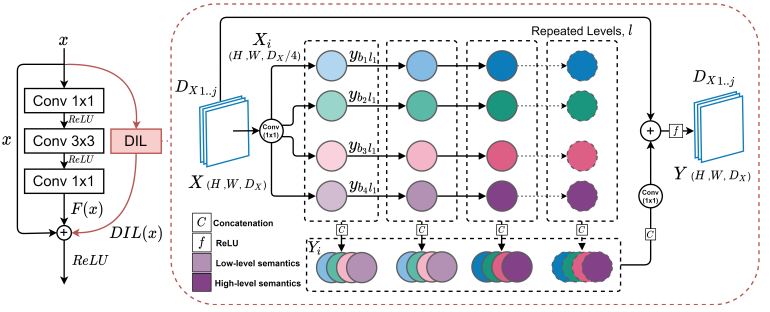

Figure 2: Left: The first bottleneck block of each stage in ResNet is replaced with a version which incorporates DIL in the residual connection. Right: The proposed DIL consists of multiple repeatable levels of atrous convolutions with adjustable rates (denoted by different colors) in a multi-branch scheme to enhance multi-level and multi-scale features into the chosen input feature map. Each level of DIL consists of four branches of 3×3 convolutions. The lighter and darker colors represent low-level and high-level semantics respectively.

The purpose of placing DIL at the first block of every stage is to allow the feature maps to be enriched with features sourced from varying receptive fields. As such, the following ResNet blocks in the current stage are able to learn distinctive representations of objects with different aspect ratios to further enhance the backbone feature maps.

3.2 Transformer Module

3.2.1 Vision Transformer

In the transformer module, the output feature maps from the convolution operations are fed directly into the vision transformer to further process the high-level information as shown in Figure 4. Similar to how it can relate distant semantics in a sentence, the self-attention module can be applied to relate long-range semantics across different receptive fields [□]. Therefore, we apply the transformer module after the backbone to enhance the ability of the model to focus on high-level features across scales. The self-attention mechanism extracts the spatial relationship from high-level features by first transforming the input into a set of linearized patch embeddings. This is done to avoid the use of a fixed pixel array representation, which would increase the computational complexity. Each patch then undergoes a 2D-convolution to a fixed number of channels followed by a linear projection into a fixed hidden representation to generate the query q, key k and value v vectors. Then, the softmax function is applied to the final output to obtain a probability distribution across the channels. Finally, the embeddings are projected back onto the input feature map X via a channel-wise multiplication to obtain the channel-wise attended feature map Y as shown in (4).

$$Y = \underbrace{softmax(softmax(\frac{QK^T}{\sqrt{d_k}})V)}_{A} \cdot X \tag{4}$$

where, Q, K and V represent the query, key and value matrices, respectively, d_k represents the dimensions of the key, $X \in \mathbb{R}^{h \times w \times c}$ denotes the input feature map, and A represents the self-attention mechanism. Specifically, we allocate a total of 768 channels for the output embeddings.

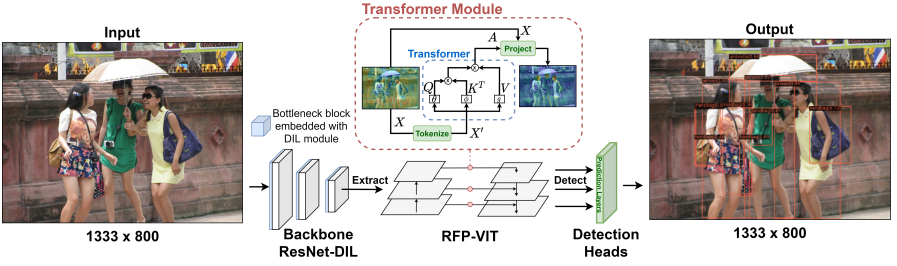

Figure 3: An overview of the proposed architecture. Our architecture utilizes the ResNet backbone as its baseline with the addition of DIL in the first bottleneck in each ResNet stage. The backbone features are then extracted and used to build a RFP [27] which integrates the transformer module into the lateral connections between the top-down and bottom-up pathways. The features are then subsequently fed into the prediction layers to produce and compute bounding box and category scores.

3.2.2 Vision Transformers in the Neck

In order to maximise the benefit of the multi-scale feature maps produced from the backbone, we employ RFP [27] as the model neck and replace the ASPP connecting module with the proposed vision transformer module. The removal of ASPP module greatly improves the speed of the network and enables the RFP to focus on capturing semantically meaningful information to enrich the feature maps. Specifically, we extract the feature map outputs from the last residual block of each stage. Thus, each transformer module receives a fused multi-scale feature map as the input to aggregate spatially distant features using cross-scale information. Finally, the embeddings are normalized by applying the softmax function onto the outputs followed by a projection back onto the input feature maps. Therefore, this process applies attention onto the image contexts at varying receptive fields. By mapping the long range dependencies across the channels of the feature map for feature refinement, the attention module is able to further emphasise important semantics on each scale of the feature maps in the feature pyramid. Subsequently, the detector heads can easily capture object-relevant features across multiple scales to improve detection accuracy.

3.3 DIL-ViT

Additionally, we propose a hybrid convolutional-transformer architecture, namely, DIL-ViT, which combines DIL and ViT into a single component. The illustration in Fig. 3 shows the overall architecture of DIL-ViT. We employ ResNet as the backbone network and add DIL to the first block of each stage. We then extract the feature map outputs from each stages last residual block from the backbone network. These feature maps of varying scales are then passed into a modified RFP embedded with ViT to recursively apply channel-wise attention onto the fused multi-scale feature maps generated from the backbone. The final output feature resulting from the RFP top-down pathway are then fed into the prediction layers. The prediction layers consist of a Cascade ROI head followed by standard classification and regression heads to produce and compute bounding box and category scores respectively.

4 Experiments

A series of extensive experiments were carried out to evaluate the performance of the proposed architectures against various object detectors. These experiments were conducted using the MS COCO dataset [21] which contains a total of 80 object classes varying over multiple scales and orientations. For training, the train2017 set was used whereas test-dev was used to consolidate results for comparisons with other state-of-the-art object detectors. Additionally, the val2017 dataset was also used to evaluate the results for analysis study. The COCO metric of Average Precision (AP) was chosen to evaluate the overall performance, by measuring over multiple intersection-over-union (IoU) thresholds and over small, medium and large objects. Following common practice, the backbone network used in the proposed architecture is first pre-trained on the ImageNet dataset and subsequently fine-tuned on the MS COCO dataset.

Training and Implementation Details. The implementation for the proposed architecture was based on the MMDetection PyTorch framework [5]. The training and testing settings that were employed are common to the Cascade R-CNN baseline, utilizing the same bounding box and classification loss functions, detection heads, and different data augmentation techniques, to improve training.

We trained the proposed architecture on a compute platform configured with an Intel Core i9-10920X processor, 64 GBytes of memory and an NVIDIA RTX 8000 graphics processor unit (GPU). We employed a single-scale training pipeline which first resizes input images to a fixed scale of 1333×800 followed by horizontal or vertical flipping with a flip ratio of 0.5. An SGD optimiser is used with a learning rate of 0.02, weight decay of $1e-4$ and momentum of 0.9 alongside a batch size of 8 images per GPU. The learning policy defined for training was a linear warmup strategy of 500 iterations with a ratio of $1e-3$ and trained for a total of 12 epochs. Experiments were conducted with test-time augmentation (TTA) enabled which included multi-scale testing through resizing and horizontal flipping.

State-of-the-art Comparison. In this section, we analyse the performance of the proposed DIL-ViT module using ResNet-50, ResNet-101 and ResNeXt-101-32×4d, as the backbone network to act as the baseline model. The performance of these architectures was compared against a range of one-stage and two-stage object detectors which serves as benchmark. Table 1 tabulates the results of the aforementioned proposed and benchmarked object detectors tested on the MS COCO test-dev using the COCO evaluation metrics. Based on these tabulated results, our proposed DIL-ViT architecture achieves competitive AP scores of 46.5%, 48.3% and 50.1% using the ResNet-50, ResNet-101 and ResNeXt-101-32×4d backbones, respectively. We can also observe that the AP across objects of different sizes exhibits improved accuracy. These improvements are a result of using the atrous filters and a vision transformer to incorporate and effectively relate multi-scale spatial information into the feature maps.

Ablation Study. In this section, we have included additional experiments on the COCO val2017 set to further evaluate the impact of each component on the overall architecture. We validate the overall model architecture using the baseline Cascade-RCNN architecture with ResNet-50+FPN as its backbone and alternating the addition of DIL and ViT. Compared to the baseline architecture, the addition of DIL successfully improved the AP by 3.3%, whereas ViT increased the AP by 2.5%, as shown in Table 2. Furthermore, it is clear that

Table 1: Performance comparison of our proposed DIL-ViT model against one- and two-stage detectors on the MS COCO test-dev subset.

Method	Backbone	Avg. Precision, IoU			Avg.Precision, Area		
		0.5:0.95	0.5	0.75	S	M	L
one-stage:							
DETR [8]	ResNet-50	40.1	60.6	42.0	18.3	43.3	59.5
AutoAssign [43]	ResNet-50	40.4	59.6	43.7	22.7	44.1	52.9
RetinaNet [24]	ResNeXt-101	40.8	61.1	44.1	24.1	44.2	51.2
CornerNet [18]	Hourglass-104	42.1	57.8	45.3	20.8	44.8	56.7
GFL [19]	ResNet-50	42.9	61.2	46.5	27.3	46.9	53.3
M2Det [42]	VGG-16	44.2	64.6	49.3	29.2	47.9	55.1
Deformable-DETR [44]	ResNet-50	44.5	63.2	48.9	28.0	47.8	58.8
CentripetalNet [10]	Hourglass-104	44.6	62.3	47.7	25.9	48.2	59.8
CentreNet [9]	Hourglass-104	44.9	62.4	48.1	25.6	47.4	57.4
two-stage:							
Mask R-CNN [15]	ResNet-101	39.8	62.3	43.4	22.1	43.2	51.2
SABL [32]	Cascade + ResNet-50	41.6	59.1	44.6	23.1	45.0	55.1
Cascade R-CNN [1]	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2
SCNet [31]	ResNet-50	43.5	62.9	47.4	26.0	46.9	56.8
DetectoRS [27]	Cascade + ResNet-50	45.0	64.3	49.1	26.4	49.3	59.7
DCN-v2 [40]	ResNet-101	46.0	67.9	50.8	27.8	49.1	59.5
TridentNet [20]	ResNet-101-Deformable	46.8	67.6	51.5	28.0	51.2	60.5
PANet [25]	ResNeXt-101	47.4	67.2	51.8	**30.1**	51.7	60.0
DIL-ViT (Ours)	Cascade + ResNet-50	46.5	65.1	50.6	26.7	49.6	59.4
DIL-ViT (Ours)	Cascade + ResNet-101	48.3	66.8	53.1	27.8	52	61.9
DIL-ViT (Ours)	**Cascade + ResNeXt-101-32×4d**	**50.1**	**68.1**	**54.8**	29.1	**53.4**	**63.2**

Table 2: Ablation study of the proposed architecture on COCO val2017.

Cascade	DIL	ViT	Avg. Precision, IoU			Avg.Precision, Area		
			0.5:0.95	0.5	0.75	S	M	L
✓			40.3	58.6	44.0	22.5	43.8	52.9
✓	✓		43.6	61.4	47.2	25.9	46.9	56.6
✓		✓	42.8	59.9	46.8	25.1	46.2	56.1
✓	✓	✓	**46.2**	**64.6**	**50.0**	**27.2**	**50.2**	**60.9**

DIL has contributed more to the increase in detection accuracy, which can be associated with its ability to learn better representations of the objects by varying the receptive field. Therefore, we verify that DIL performs better as compared to a traditional multi-scale pyramidal architecture such as the FPN where the receptive field is limited to a fixed set of feature map scales. Thus, DIL is shown to be capable of improving detection performance across all object scales. Additionally, experimental results shown in Table 3 suggest that additional atrous branches in DIL would increase the receptive field which in turn increases the accuracy of the model. Nevertheless, we do observe a substantially increased trade-off of complexity from the additional atrous branches for marginally higher detection accuracy.

Qualitative Analysis. To better understand the impact of the self-attention mechanism in ViT, we visualise the heatmaps produced before and after employing ViT. Figure 4 clearly illustrates that prior to the application of the self-attention mechanism (first row), the heatmap produced by the model is less refined. This can be explained by the high activation on both foreground and background features. After the application of the self-attention mechanism (second row), more distinctive feature maps are produced by applying feature importance onto the foreground objects (as indicated by the red regions) and less emphasis on irrelevant semantic information (as indicated by the blue regions). Therefore, it can be inferred that

Figure 4: Visualisations of the heatmaps produced by the self-attention mechanism in ViT. For example, in the first column of images the attention (red) towards the persons in the image is shown and other unimportant semantic information (blue) are less emphasized. Top row: Before applying ViT. Bottom row: After applying ViT.

Table 3: Ablation study on the number of atrous branches in DIL using the SSD-300 baseline on COCO val2017. * indicates the SSD-300 baseline with no atrous branches.

Branches	Avg. Precision, IoU			Avg.Precision, Area		
	0.5:0.95	0.5	0.75	S	M	L
*	25.6	43.8	26.3	6.8	27.8	42.2
2	26.3	44.7	27.2	7.3	28.4	43.1
4	27.4	45.4	28.4	8.8	29.3	43.9
8	28.9	46.8	29.1	9.4	30.2	44.8

ViT enhances the feature extraction process using the self-attention mechanism by relating spatially distant features present in the feature map and focusing solely on the important features. The supplemental material includes additional analysis in visualizing the inferred bounding box outputs of images with varying degrees of scene complexity.

5 Conclusion

In this paper, we put forward a novel method of unifying the dilation module with a vision transformer (i.e., DIL-ViT). This method addresses the issue of the high degree of less important semantics, which was caused by the variations of receptive fields during the feature extraction process. The proposed model leverages atrous convolutions to generate rich multi-scale feature maps and employs a self-attention mechanism to enhance important backbone features. By integrating DIL into various object detectors, we observe notable improvements in detection accuracies with minimal impact to the inferring speed. As a whole, the ability of DIL-ViT to apply attention onto specific receptive fields has the potential to yield beneficial impacts in object detection. For future research, we could extend the applicability of our work into the domain of spatio-temporal object detection.

Acknowledgements

This work was funded by the Malaysian Ministry of Education's Fundamental Research Grant Scheme (Grant Number:FRGS/1/2018/ICT02/MUSM/02/3) and the Advanced Engineering Platform's CyberSecurity AI (ψ^2) Cluster Funding (Account Number: AEP-2020-Cluster-04), Monash University Malaysia.

References

[1] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984.

[2] 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.

[3] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016.

[4] 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.

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[6] 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.

[7] Xiaoyu Chen, Wei Li, Qingbo Wu, and Fanman Meng. Adaptive multi-scale information flow for object detection. In *BMVC*, page 83, 2018.

[8] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2590–2600, 2017.

[9] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[10] Zhiwei Dong, Guoxuan Li, Yue Liao, Fei Wang, Pengju Ren, and Chen Qian. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] 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.

[12] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019.

[13] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7036–7045, 2019.

[14] 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.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[16] Matthias Holschneider, Richard Kronland-Martinet, Jean Morlet, and Ph Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*, pages 286–297. Springer, 1990.

[17] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. *IEEE Access*, 7:128837–128868, 2019.

[18] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[19] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *arXiv preprint arXiv:2006.04388*, 2020.

[20] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6054–6063, 2019.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[22] 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.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[24] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[25] 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.

[26] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[27] 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.

[28] 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.

[29] 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.

[30] Zhiqiang Shen, Honghui Shi, Jiahui Yu, Hai Phan, Rogerio Feris, Liangliang Cao, Ding Liu, Xinchao Wang, Thomas Huang, and Marios Savvides. Improving object detection from scratch via gated feature reuse. *arXiv preprint arXiv:1712.00886*, 2017.

[31] 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.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[33] Thang Vu, Kang Haeyong, and Chang D Yoo. Scnet: Training inference sample consistency for instance segmentation. In *AAAI*, 2021.

[34] Jiaqi Wang, Wenwei Zhang, Yuhang Cao, Kai Chen, Jiangmiao Pang, Tao Gong, Jianping Shi, Chen Change Loy, and Dahua Lin. Side-aware boundary localization for more precise object detection. In *ECCV*, 2020.

[35] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017.

[36] 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.

[37] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[38] Peng Zhou, Bingbing Ni, Cong Geng, Jianguo Hu, and Yi Xu. Scale-transferrable object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 528–537, 2018.

[39] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*, 2020.

[40] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.

[41] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.