Intelligent Information Management, 2025, 17(6), 225-234
https://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

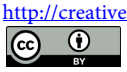
Dual-Dilated Large Kernel Convolution for Visual Attention Network

Kwok-Wai Cheung, Yuk Tai Siu, Ka Lok Sobel Chan

School of Communication, The Hang Seng University of Hong Kong, Hong Kong, China
Email: keithcheung@hsu.edu.hk, ytsiu@hsu.edu.hk, sobelchan@hsu.edu.hk

How to cite this paper: Cheung, K.-W., Siu, Y.T. and Chan, K.L.S. (2025) Dual-Dilated Large Kernel Convolution for Visual Attention Network. *Intelligent Information Management*, **17**, 225-234.
https://doi.org/10.4236/iim.2025.176012

Received: September 17, 2025
Accepted: November 8, 2025
Published: November 11, 2025



Abstract

Visual Attention Networks (VANs) leveraging Large Kernel Attention (LKA) have demonstrated remarkable performance in diverse computer vision tasks, often outperforming Vision Transformers (ViTs) in some cases. LKA strategically combines the strengths of Convolutional Neural Networks (CNNs), such as local structure information, with the long-range dependency and adaptability of self-attention mechanisms, while maintaining linear computational complexity. This paper introduces Dual-Dilated Large Kernel (D2LK), a novel attention mechanism designed to enhance LKA's kernel decomposition. D2LK improves upon LKA by incorporating an additional depth-wise dilation convolution layer, which enables the approximation of larger kernel convolutions with further reduced computational requirements. This decomposition allows for a more efficient representation of larger effective receptive fields. Our experiments demonstrate that D2LK achieves a superior balance between efficiency and performance. For instance, a D2LK module configured with a kernel size of 29 and 32 channels reduces parameters by 11% (3,008 parameters) compared to an LKA module with the same specifications (3,392 parameters). When integrated into the VAN-B0 architecture, D2LK with a larger kernel size of 29 yields a Top-1 accuracy of 85.1% on ImageNet100 classification, a slight improvement over the LKA baseline (kernel size 21), which achieved 85.0%. Critically, this performance gain is accomplished with a marginally reduced overall parameter count (3.8649 million for D2LK vs. 3.8745 million for LKA). These results validate D2LK as an efficient and effective attention mechanism for Visual Attention Networks, enabling enhanced receptive fields at lower computational overhead.

Keywords

Attention, Large Kernel, Dilated Convolution

1. Introduction

Originally designed for natural language processing tasks, self-attention-based transformers have revolutionized computer vision, challenging traditional CNNs and becoming the backbone of state-of-the-art models. However, transformer models have shortcomings of demanding large training datasets, high memory and computational requirements, limiting their use in lightweight edge applications. As a hybrid of CNN and attention-based architecture, Visual Attention Networks (VANs) [1] have emerged as a competitive alternative to CNNs and Vision Transformers (ViTs), balancing efficiency and performance. VANs combine CNN-like locality (via depth-wise convolutions), transformer-like adaptability (via Large Kernel Attention, LKA), and efficiency (with linear complexity instead of quadratic in self-attention).

The advancement of visual attention networks has been significantly influenced by the introduction of large kernel attention mechanisms, which enhance model performance while maintaining computational efficiency. A notable approach in this realm is the Large Kernel Attention (LKA) mechanism, which strategically integrates large separable kernels into neural architecture to effectively capture contextual information over wider receptive fields without disproportionately increasing computational costs [2]. This design philosophy has been shown to enable models to strike a balance between computational demand and performance enhancement, particularly in tasks such as image classification, as evidenced by the application of LKA in a recent study that combined depth-wise and dilated convolutions, outperforming existing backbone networks [3].

LKA has the potential for integration with foundation models. For instance, VAN variants are used as backbones for multimodal models (e.g., CLIP-style architectures) and used in lightweight edge-AI models (e.g., drones, mobile devices). Thus, VANs provide a balance between CNNs and Transformers, offering efficiency, scalability, and strong performance.

While not yet as dominant as ViTs in large-scale systems, they are gaining traction in real-world applications where hardware constraints matter. As the core of VAN, Large Kernel Attention (LKA) offers efficiency advantages over standard self-attention. However, LKA has a major limitation—large kernels (e.g., 21×21) require storing more weights than standard 3×3 convolutions, limiting its use in memory-constrained devices (mobile/edge). Thus, LKA is implemented as decomposed kernels of depth-wise and dilated convolutions.

The efficacy of LKA is demonstrated in lightweight models designed for real-time applications. For example, multi-scale convolutional networks incorporating LKA have shown promising results in optimizing detection accuracy while maintaining reduced computational complexity, particularly for small target detection in visual inspections [4]. The large receptive fields enabled by LKA techniques contribute significantly to the model's ability to retain critical information while minimizing resource load, making them well-suited for environments with constrained computational resources.

The integration of large kernel attention mechanisms into visual attention networks has significantly enhanced the efficiency and effectiveness of image processing tasks. This paper introduces and explores the use of mixed dilated convolution in LKA decomposition to further reduce computational requirements while maintaining performance.

2. Background

2.1. Large Kernel Attention

Large Kernel Attention (LKA), as shown in Figure 1, is an attention mechanism designed for computer vision tasks that combines the benefits of large receptive fields (like in CNNs) and adaptive feature selection (like in attention models). It improves upon traditional attention mechanisms by using large convolutional kernels to capture long-range dependencies efficiently.

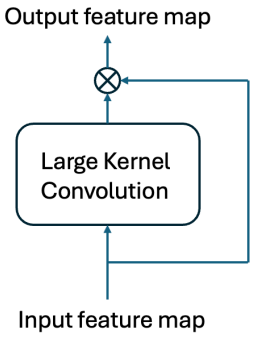


Figure 1. Large kernel attention.

To reduce the computational burden, LKA is implemented by decomposing the large kernel convolution into Depth-wise Convolution (DW-Conv), Depth-wise Dilated Convolution (DW-D-Conv), and Point-wise Convolution (1 × 1 Conv).

Given input feature map X, LKA computes attention weights as:

$$\text{Attention-map} = \text{Conv}_{1\times1}(\text{DW-D-Conv}(\text{DW-Conv}(X)))$$

$$\text{Output} = \text{Attention-map} \otimes X$$

2.2. Large Kernel Convolution in Visual Attention Network

The Visual Attention Network (VAN) [1] is used as a backbone for computer vision tasks, including classification, detection, and segmentation. VAN follows a hierarchical multi-stage architecture, similar to many modern vision backbones (e.g., CNNs or vision transformers). It combines local (CNN) and global (attention) modeling for better feature learning.

VAN utilizes a 4-stage hierarchy to process the input image through four sequential stages, with each stage progressively transforming the feature representation. The structure of the processing stages is illustrated in Figure 2.

Progressive downsampling is applied through the stages. For instance, the output size and number of output channels of VAN baseline model [1] is shown in

Table 1. The spatial resolution is decreased via strided convolution, while channel dimension is increased for richer feature representation.

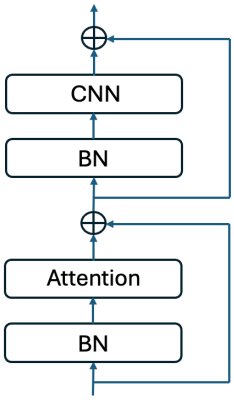


Figure 2. A stage of VAN.

Table 1. The setting for VAN baseline configuration.

VAN baseline	Output size	No. of output channels
Stage 1	H/4 × W/4	32
Stage 2	H/8 × W/8	64
Stage 3	H/16 × W/16	160
Stage 4	H/32 × W/32	256

2.3. Advantages of Large Kernel Attention Over Self-Attention

As an adaptive selection process according to input feature, attention for vision tasks can be categorized into channel attention for recalibrating channel-wise feature responses [5] [6], spatial attention for generating a spatial mask to highlight key areas [7] [8], temporal attention for weighting frames dynamically in videos/time-series [9], transformer-based self-attention for global pairwise affinity between image patches [10] [11], and hybrid attention such as combining CNN and transformer for local and global context [1] [12].

Originally designed for natural language processing tasks [13], self-attention has gained popularity in vision tasks [10] [11] [14] due to its ability to model long-range dependencies, capture global context, scale with data, and adapt dynamically to input content—addressing key limitations of traditional CNN.

However, self-attention's superior performance comes at a cost: it requires quadratically growing computation and memory as input size increases. Additionally, by treating images as 1D sequences of patches, it loses the inherent 2D spatial structure, weakening local context modeling—a strength of CNNs.

VAN employs LKA as its attention mechanism. As an alternative to standard self-attention mechanisms in vision models, LKA offers several advantages in terms of efficiency, locality, and performance. LKA uses depth-wise convolutions with large kernels (e.g., 21 × 21) to capture long-range dependencies. Its linear

complexity is more scalable for high-resolution inputs. The built-in locality of large-kernel convolutions naturally focuses on nearby pixels first, then expands to global context, addressing the limitation of self-attention.

3. Methodology

3.1. Decomposition of Large Kernel Convolution

Given the input and output feature maps $X, Y \in \mathbb{R}^{H\times W\times C}$ where C is the number of input channels, H and W represent the height and width of the feature maps respectively, large kernel convolution using a kernel of size $k \times k$ can be expressed as

$$Y_{(i,j,c')} = \sum_{m}\sum_{n}\sum_{c} W^{LK}_{(m,n,c,c')} \cdot X_{(i+m,j+n,c)} \tag{1}$$

where $i \in [1,H]$, $j \in [1,W], c \in [1,C]$, $m,n \in [1,k]$. The kernel weight is denoted as $W^{LK} \in \mathbb{R}^{k\times k\times C\times C}$. For instance, convolution using 21 × 21 kernel with 32 input and output channels uses 451,584 weight parameters.

Efficient computation of large kernel attention [1] decomposes a $K \times K$ large kernel convolution into a $(2d-1) \times (2d-1)$ DW-Conv, a $\lfloor K/d \rfloor \times \lfloor K/d \rfloor$ DW-D-Conv and a $\text{Conv}_{1\times1}$. Given $Y^{dw}, Y^{dwd} \in \mathbb{R}^{H\times W\times C}$, $W^{dw}, W^{dwd} \in \mathbb{R}^{k\times k\times C}$, $W^{p} \in \mathbb{R}^{1\times1\times C\times C}$, the computation of DW-Conv, DW-D-Conv, and $\text{Conv}_{1\times1}$ are expressed as:

$$\text{DW-Conv:}\quad Y^{dw}_{(i,j,c)} = \sum_{m=0}^{2d-2}\sum_{n=0}^{2d-2} W^{dw}_{(m,n,c)} \cdot X_{(i+m,j+n,c)} \tag{2}$$

$$\text{DW-D-Conv:}\quad Y^{dwd}_{(i,j,c)} = \sum_{m=0}^{\lfloor K/d \rfloor-1}\sum_{n=0}^{\lfloor K/d \rfloor-1} W^{dwd}_{(m,n,c)} \cdot Y^{dw}_{(i+dm,j+dn,c)} \tag{3}$$

$$\text{Conv}_{1\times1}:\quad Y_{(i,j,c')} = \sum_{c=0}^{C-1} W^{p}_{(1,1,c,c')} \cdot Y^{dwd}_{(i,j,c)} \tag{4}$$

where d is the dilation rate. **Figure 3(a)** shows the depth-wise kernel locations of this decomposed convolution for K=21 and d = 4.

3.2. Dual-Dilated Large Kernel (D2LK)

To enhance the mid and long range support without increasing the kernel sizes, we introduce dual dilated depth-wise convolution in the decomposition. For decomposition with double dilated depth-wise convolution layers. The decomposition can be expressed as:

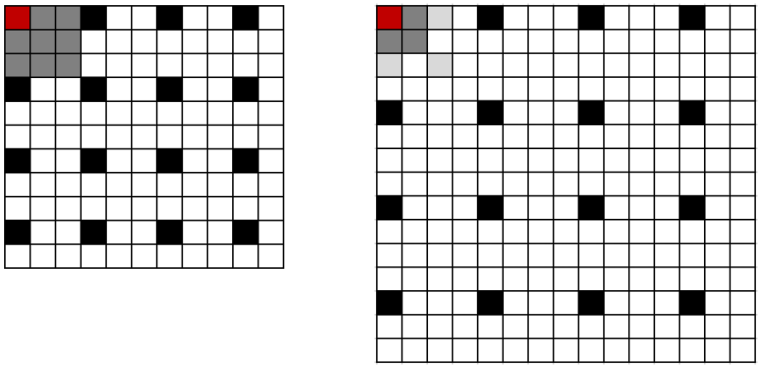
$$\text{DW-Conv:}\quad Y^{dw}_{(i,j,c)} = \sum_{m=0}^{k-1}\sum_{n=0}^{k-1} W^{dw}_{(m,n,c)} \cdot X_{(i+m,j+n,c)} \quad \text{for} \quad k = 2\left\lfloor \frac{d_1}{2} \right\rfloor + 1 \tag{5}$$

$$\text{DW-D-Conv1:}\quad Y^{dwd1}_{(i,j,c)} = \sum_{m=0}^{2}\sum_{n=0}^{2} W^{dwd1}_{(m,n,c)} \cdot Y^{dw}_{(i+d_1m,j+d_1n,c)} \quad \text{for} \quad d_1 = \left\lfloor \frac{d_2}{2} \right\rfloor \tag{6}$$

$$\text{DW-D-Conv2:}\quad Y^{dwd2}_{(i,j,c)} = \sum_{m=0}^{\lfloor K/d_2 \rfloor-1}\sum_{n=0}^{\lfloor K/d_2 \rfloor-1} W^{dwd2}_{(m,n,c)} \cdot Y^{dwd1}_{(i+d_2m,j+d_2n,c)} \tag{7}$$

$$\text{Conv}_{1\times1}:\quad Y_{(i,j,c')} = \sum_{c=0}^{C-1} W^{p}_{(1,1,c,c')} \cdot Y^{dwd2}_{(i,j,c)} \tag{8}$$

where d_2 has the minimum value of 4 for dual dilation layers design. Figure 3(b) shows the depth-wise kernel locations of double dilated depth-wise convolution for K=29 and d_2 = 4.



(a) LKA for K = 21, d = 3 (b) D2LK for K = 29, d_2 = 4

Figure 3. Decomposition of large kernel convolution. The grey grid and black grid represent the location of the convolution kernels. The red grid means the center point.

3.3. Complexity Analysis

This subsection details the computational cost, in Floating-Point Operations (FLOPs), and the number of parameters for the proposed D2LK and LKA modules, alongside the large kernel convolution. For simplicity, the bias term is omitted from all calculations. We assume the input and output feature maps for all modules share identical dimensions ($H\times W\times C$). The FLOPs and parameters for large kernel convolution are calculated as follows:

$$\text{Parameters} = \left(K\times K\times C\right)\times C \tag{9}$$

$$\text{FLOPs} = \left(K\times K\times C\right)\times C\times H\times W \tag{10}$$

The FLOPs and parameters for the original LKA module are calculated as follows:

$$\text{Parameters} = \left(\left(2d-1\right)^2+\left\lfloor\frac{K}{d}\right\rfloor^2+C\right)\times C \tag{11}$$

$$\text{FLOPs} = \left(\left(2d-1\right)^2+\left\lfloor\frac{K}{d}\right\rfloor^2+C\right)\times C\times H\times W \tag{12}$$

where d_2 is the dilation rate of DW-D-Conv2 layer.

According to Equation (9) & Equation (11), LKA can significantly reduce the computation requirement of large kernel convolution by decomposing the convolution into three layers—depth-wise convolution, depth-wise dilation convolution, and channel convolution. D2LK can approximate larger kernel convolution with reduced computation requirement compared with LKA by introducing an additional depth-wise dilation convolution layer. For example, the number parameters of LKA for d = 3, K=21 and C=32 is 3,392 while that of D2LK for d_2 = 4, K=29 and C = 32 is 3,008, *i.e.* 11% reduction.

4. Experiments

This section details the experimental evaluation of the proposed Dual-Dilated Large Kernel (D2LK) attention mechanism within the Visual Attention Network (VAN) architecture. The primary objective was to assess D2LK's performance and efficiency compared to the standard Large Kernel Attention (LKA) design. The experiments utilized the VAN-B0 architecture as a baseline model. Comparisons were performed across different kernel sizes (K) and included both standard and separable versions of the attention modules for image classification. To reduce the computation requirement, ImageNet100 [15], a subset of ImageNet-1k Dataset is used in the experiments. Specifically, VAN-B0 models employing LKA were tested with a kernel size of 21, while the proposed D2LK models were evaluated with a kernel size of 29. The "separable" versions refer to configurations that decompose the 2D convolutional kernel into cascaded horizontal and vertical 1D kernels, a strategy also explored with LSKA [2] in other contexts to mitigate quadratic increases in computational and memory footprints.

4.1. Performance on ImageNet100 Classification

The comparative results of the performance on ImageNet-100 classification for Top-1 accuracy and model parameters are shown in **Table 2**. The baseline VAN-B0 model, integrating the standard LKA with a kernel size of 21, achieved a Top-1 accuracy of 85.0% with a parameter count of 3.8745 million. The proposed D2LK model, employing a larger kernel size of 29, demonstrated a slight improvement in performance, yielding a Top-1 accuracy of 85.1%. This was achieved with a marginally reduced parameter count of 3.8649 million. For separable configurations, the VAN-B0 (separable) with a kernel size of 21 attained a Top-1 accuracy of 84.6% with 3.7977 million parameters. The D2LK (separable) model, which also has a larger kernel size of 29, achieved an identical Top-1 accuracy of 84.6%, although with a slightly higher parameter count of 3.8041 million compared to the VAN-B0 separable baseline.

Table 2. Comparison of VAN baseline and D2LK.

Model	Large kernel size (K)	Parameters (M)	Top-1 Acc (%)
VAN-B0	21	3,874,500	85.0
VAN-B0 (separable)	21	3,797,700	84.6
D2LK	29	3,864,900	85.1
D2LK (separable)	29	3,804,100	84.6

4.2. Computational Efficiency and Parameter Reduction

Beyond empirical accuracy, a theoretical complexity analysis highlighted D2LK's efficiency advantages. D2LK is designed to approximate larger kernel convolutions with reduced computational requirements compared to LKA by incorporating an additional depth-wise dilation convolution layer. As an example, for a fixed

kernel size of $K = 21$ and 32 channels ($C = 32$), the LKA design had 3,392 parameters. In contrast, the D2LK design, with a d_2 (dilation rate for DW-D-Conv2) of 4 and $K = 29$, demonstrated an 11% reduction in parameters, totaling 3,008 parameters. This reduction in individual attention module parameters contributes to the overall efficiency.

4.3. Discussion

The experimental findings indicate that the proposed D2LK attention mechanism, even with a larger kernel size of 29, can achieve comparable or slightly superior Top-1 accuracy compared to the baseline VAN-B0 with LKA (kernel size 21). Notably, this performance is maintained or slightly improved while either reducing or keeping the overall model parameter count very similar. The theoretical analysis further supports D2LK's design philosophy of enhancing mid and long-range support without a substantial increase in computational cost. This is achieved by introducing a dual dilated depth-wise convolution within the decomposition, which allows for larger effective receptive fields at a lower parameter and FLOPs cost compared to standard large kernel convolutions or even LKA. This balance of performance and efficiency underscores D2LK's potential for applications where both are critical.

5. Conclusions

This paper introduced and explored Dual-Dilated Large Kernel (D2LK), a novel attention mechanism designed to address the computational inefficiencies associated with increasing kernel sizes in the Large Kernel Attention (LKA) modules of Visual Attention Networks (VANs). While LKA offers advantages over standard self-attention by combining CNN-like locality and transformer-like adaptability, its depth-wise convolutional layer incurs a quadratic increase in computational and memory requirements as kernel sizes grow, restricting the use of extremely large kernels. To address this limitation, D2LK enhances LKA's kernel decomposition by incorporating an additional depth-wise dilation convolution layer, enabling the approximation of larger kernel convolutions with further reduced computational demands. This innovative design allows for a more efficient representation of larger effective receptive fields, providing enhanced mid and long-range support without a substantial increase in computational overhead.

D2LK can reduce computational costs and parameter counts. For instance, a D2LK module with a kernel size of 21 and 32 channels achieved an 11% reduction in parameters (3,008 parameters) compared to a standard LKA module (3,392 parameters). When integrated into the VAN-B0 architecture, D2LK with a larger kernel size of 29 yielded a Top-1 accuracy of 85.1% on ImageNet100 classification, slightly outperforming the LKA baseline (kernel size 21), which achieved 85.0%. This performance gain was accomplished with a marginally reduced overall parameter count (3.8649 million for D2LK vs. 3.8745 million for LKA). D2LK's separable configurations also maintained comparable accuracy with competitive pa-

rameter counts.

The design of D2LK allows for a good trade-off between kernel size, parameter size, and speed, maintaining comparable or better performance than original LKA in image classification evaluations. This indicates D2LK's potential for scalability to larger kernels without performance saturation. D2LK provides an efficient and effective attention mechanism for Visual Attention Networks, enabling enhanced receptive fields at lower computational overhead, thus offering a strong baseline for future research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Guo, M., Lu, C., Liu, Z., Cheng, M. and Hu, S. (2023) Visual Attention Network. *Computational Visual Media*, **9**, 733-752. https://doi.org/10.1007/s41095-023-0364-2

[2] Lau, K., Po, L. and Rehman, Y. (2023) Large Separable Kernel Attention: Rethinking the Large Kernel Attention Design in CNN. https://doi.org/10.2139/ssrn.4463661

[3] Liu, S., Wei, J., Liu, G. and Zhou, B. (2023) Image Classification Model Based on Large Kernel Attention Mechanism and Relative Position Self-Attention Mechanism. *PeerJ Computer Science*, **9**, e1344. https://doi.org/10.7717/peerj-cs.1344

[4] Wang, J., Wang, Y., Sun, A. and Zhang, Y. (2024) A Lightweight Network FLA-Detect for Steel Surface Defect Detection. https://doi.org/10.21203/rs.3.rs-4581669/v1

[5] Hu, J., Shen, L., Albanie, S., Sun, G. and Wu, E. (2019) Squeeze-and-Excitation Networks. https://doi.org/10.48550/arXiv.1709.01507

[6] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W. and Hu, Q. (2020) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 11531-11539. https://doi.org/10.1109/cvpr42600.2020.01155

[7] Wang, X., Girshick, R., Gupta, A. and He, K. (2018) Non-Local Neural Networks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7794-7803. https://doi.org/10.1109/cvpr.2018.00813

[8] Woo, S., Park, J., Lee, J. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. In: Ferrari, V., *et al.*, Eds., *Computer Vision—ECCV* 2018, Springer International Publishing, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[9] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., *et al.* (2016) Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: Leibe, B., *et al.*, Eds., *Computer Vision—ECCV 2016*, Springer International Publishing, 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., *et al.* (2021) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929

[11] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al.* (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/iccv48922.2021.00986

[12] Dai, Z., Liu, H., Le, Q.V. and Tan, M. (2021) CoAtNet: Marrying Convolution and

Attention for All Data Sizes. https://doi.org/10.48550/arXiv.2106.04803

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010. https://arxiv.org/abs/1706.03762

[14] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S. (2020) End-to-End Object Detection with Transformers. In: Vedaldi, A., *et al.*, Eds., *Computer Vision—ECCV* 2020, Springer International Publishing, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[15] Shekhar, A. (2021) ImageNet100: A Sample of ImageNet Classes. https://www.kaggle.com/datasets/ambityga/imagenet100