FUSING LARGE KERNEL AND CROSS CONVOLUTION (LKC²) FOR EFFICIENT AND ACCURATE OBJECT DETECTION

[1]Ene P. C., [2]Onuigbo C. M. [3]Durugbor Frank Chukwudebelu

[123]Department of Electrical and Electronic Engineering, Enugu State University of Science and Technology ESUT, Enugu

[1]*eneh.princewill@esut.edu.ng*, [2]*chika.onuigbo@esut.edu.ng*, [3]*durugbofrank@yahoo.com*

Corresponding Author's email: *eneh.princewill@esut.edu.ng*

ABSTRACT

Advancements in object recognition have surged, thanks to Convolutional Neural network (CNN), but achieving an optimal mix of extensive coverage and low processing demands remains challenging. Conventional network architectures commonly employ small convolutional kernels, such as 3×3, which are effective at capturing local features but often fail to model broader spatial dependencies required for precise localization. Expanding the structure or shifting to attention-based systems usually raises resource usage and delays processing, limiting their use in live scenarios. To address this, a combined unit that integrates Large Kernel Convolution (LKC) with Cross Convolution (CC) called the LKC² block was developed, The LKC part uses channel-separated filters to expand the view area and better grasp overall patterns, while the CC part combines details from neighboring levels to maintain alignment in space and meaning. This dual strategy allows the system to effectively identify both immediate and distant features without excessively increasing variables. The proposed LKC² module was integrated into widely used object detection architectures, including You Only Look Once version 8 (YOLOv8) and Faster Region-based Convolutional Neural Network (R-CNN), and evaluated on standard benchmark datasets such as Microsoft Common Objects in Context (MS COCO) and Pascal Visual Object Classes (VOC). The outcomes revealed boosts in average accuracy of Intersection over Union (IoU) of 0.5, around 3.6% higher on Common Objects in Context (COCO) and about 2.9% on VOC compared to baselines. Notably, this enhancement added less than 5% to operations count, highlighting its resource efficiency. Additional tests confirmed that LKC and CC each improve results separately, but together they create a stronger combined effect, enhancing range perception and refining multi-level detail integration. Overall, findings suggest that employing broad filters and inter-layer blending can successfully tackle issues in local detail capture and broad scene understanding in network-based detectors. The LKC² unit improves detection accuracy for small or densely grouped objects while maintaining the computational efficiency required for real-world deployment, making it a promising enhancement for future vision-based systems in automation, autonomous navigation, and surveillance applications.

KEYWORDS: *Object Detection, Large Kernel Convolution, Cross Convolution, Feature Fusion, Deep Learning, Convolutional Neural Network (CNN), Receptive Field Expansion*

1. INTRODUCTION

Object detection remains one of the central problems in computer vision, requiring simultaneous localization and classification of visual entities within complex environments. Despite substantial progress driven by deep convolutional neural networks (CNNs), achieving a consistent balance between contextual modelling, localization accuracy, and computational efficiency remains challenging. Detection systems deployed in real-world applications such as autonomous driving, intelligent surveillance, and industrial automation must operate under strict latency and resource constraints while maintaining high precision.

Early region-based detection frameworks established the foundation for modern object recognition pipelines. The introduction of region proposal mechanisms significantly improved localization performance (Ren et al., 2015), while single-stage detectors later emphasized real-time inference capability (Redmon & Farhadi, 2018; Liu et al., 2016). Methods like Cross-Stage Links, BiFPN (Tan et al., 2020), and some hierarchies in the literature act as paths for smooth data flow across levels, boosting space unity and

meaning depth. Although these frameworks differ in structure, both largely depend on convolutional feature extractors built upon small spatial kernels. The development of multi-stage object detection architectures, initially introduced with R-CNN and later improved through more efficient variants such as Faster R-CNN (Ren et al., 2015), represented a major advancement in the field. These frameworks separated the process of generating region proposals from the object classification stage, which significantly improved detection accuracy.

Subsequently, single-stage methods, including YOLO (Redmon & Farhadi, 2018) and SSD, further improved computational speed, making them particularly suitable for real-time and latency-critical applications.

Compact convolutional filters (e.g., 3×3) are highly effective for capturing local textures and edges; however, they provide limited instantaneous spatial coverage. Expanding receptive fields through deeper stacking increases parameter redundancy and computational depth, which may degrade optimization stability. To overcome these constraints, recent research has explored broader receptive-field mechanisms, including enlarged convolution kernels (Ding et al., 2022; Liu et al., 2022) and global self-attention modeling (Vaswani et al., 2017; Dosovitskiy et al., 2021). While attention-based models offer strong long-range interaction capability, their quadratic complexity and memory demand can limit deployment in resource-sensitive environments.

Architectures employing large convolutional kernels, such as ConvNeXt (Liu et al., 2022) and RepLKNet (Ding et al., 2022), illustrate that expanded filter sizes (e.g., 7×7 or 15×15) can capture broad contextual information comparable to attention mechanisms, while still preserving convolutional inductive biases and computational efficiency advantages. Nevertheless, these models typically require substantial computational resources and may exhibit limitations in maintaining consistent multi-scale feature alignment.

Large-kernel convolutional designs provide an alternative by directly increasing spatial support while retaining convolutional inductive biases. Empirical evidence suggests that wide filters can approximate global interaction effects without fully abandoning convolutional efficiency (Ding et al., 2022; Ding et al., 2024). However, receptive field expansion alone does not fully address the effective multi-scale feature coordination challenge in detection systems.

Modern detectors rely heavily on hierarchical feature pyramids to represent objects of varying sizes (Lin et al., 2017; Tan et al., 2021). While feature pyramid structures improve scale robustness, inadequate interaction between adjacent feature levels may result in semantic inconsistency or spatial misalignment. Cross-scale fusion mechanisms, such as bidirectional feature aggregation (Tan et al., 2021) and cross-stage partial connections (Wang et al., 2020), attempt to enhance information flow across depths. Nonetheless, most approaches treat receptive field enlargement and multi-scale fusion as separate architectural concerns.

These observations motivate a unified approach that jointly enhances spatial range and cross-level coherence within a single modular structure. Instead of relying exclusively on deeper stacking or attention-heavy mechanisms, this study proposes a convolution-centric solution that integrates expanded receptive fields with structured inter-scale interaction while preserving computational tractability.

This paper introduces **LKC²**, a hybrid module combining:

1. **Large Kernel Convolution (LKC)** for direct spatial expansion, and
2. **Cross Convolution (CC)** for controlled interaction between neighbouring hierarchical representations.

The LKC component employs depthwise separable wide-kernel filtering to capture extended contextual dependencies while maintaining parameter efficiency. The CC component promotes cross-scale consistency by explicitly fusing adjacent feature maps through learnable convolutional interaction. By embedding both operations within a compact block, the proposed design aims to strengthen spatial awareness and hierarchical alignment simultaneously.

To evaluate architectural generality, LKC² is integrated into two representative detection frameworks: YOLOv8 and Faster R-CNN. Experiments conducted on MS COCO 2017 and Pascal VOC 2012 demonstrate consistent improvements in mean Average Precision (mAP) with minimal additional computational overhead. Gains are particularly notable for small and densely distributed objects, suggesting that combined receptive-field enlargement and cross-scale refinement produce complementary representational benefits.

The primary contributions of this work are summarized as follows:

- A unified convolutional module (LKC²) that jointly models wide spatial context and cross-scale feature interaction.
- A computationally efficient design that maintains near-baseline inference speed while improving detection accuracy.
- Comprehensive empirical validation across multiple detectors and datasets, including ablation studies demonstrating synergistic effects between LKC and CC components.
- Statistical validation confirming the robustness and reproducibility of observed performance gains.

By addressing both contextual range and hierarchical coordination within a lightweight convolutional framework, this work contributes toward more context-aware and deployment-friendly object detection architectures.

2. METHODOLOGY

2.1 LARGE KERNEL CONVOLUTIONAL REPRESENTATION

The proposed architecture shown in figure 1 introduces an enhanced backbone architecture that integrates Large Kernel Convolutional Square (LKC²) blocks and Large Kernel Spatial Mixing (LKSm) blocks to improve spatial feature representation. The network is organized into four hierarchical stages, where each stage progressively reduces the spatial resolution while increasing the semantic richness of the extracted features.

The input feature map is first processed through an initial down sampling layer, producing a feature representation with reduced spatial dimensions. The network then proceeds through multiple stages composed of alternating LKC² blocks and LKSm blocks, enabling both local feature refinement and global spatial context modeling.

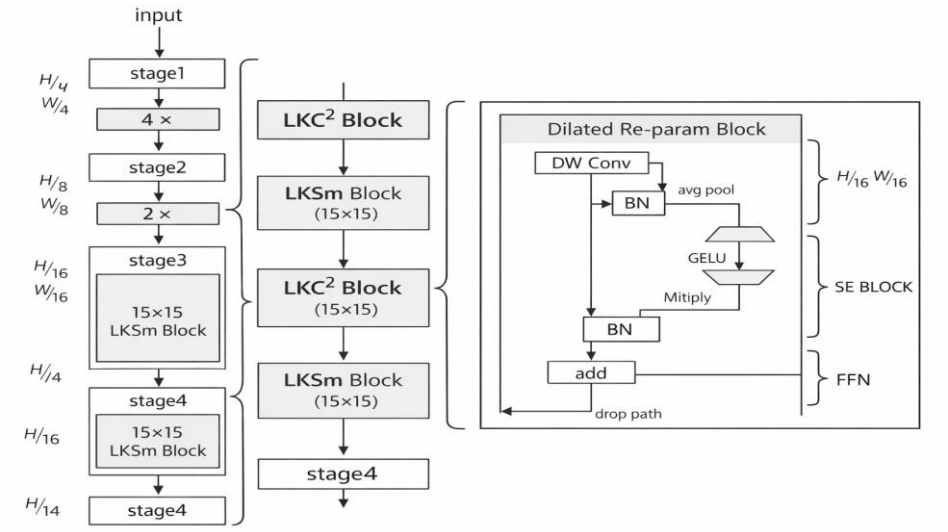


Figure 1: Schematic Diagram of the Large Kernel Convolution Network Structure.

Unlike traditional convolutional architectures that rely heavily on stacking multiple small kernels (e.g., 3×3) to expand the receptive field, the proposed framework directly employs large convolutional kernels such as $7 \times 7$and 15×15. This design allows the network to capture long-range spatial dependencies more effectively while maintaining computational efficiency.

Within each block, a dilated re-parameterization module is used to enhance feature extraction. This module consists of a depthwise convolution layer, followed by batch normalization and a lightweight squeeze-and-excitation (SE) mechanism that adaptively recalibrates channel-wise feature responses. The SE component applies global feature aggregation through average pooling and non-linear activations to emphasize informative feature channels.

Additionally, a Feed-Forward Network (FFN) is incorporated to improve feature transformation capacity. Residual connections and drop-path regularization are used to stabilize training and improve model generalization.

By combining large kernel convolutions**,** efficient attention mechanisms**,** and multi-stage feature processing**,** the proposed architecture significantly enhances the ability to detect small and densely packed objects while maintaining the computational speed required for real-time applications.

Let the input feature tensor be defined as:

$$X_0 \in \mathbb{R}^{H \times W \times C_0} \quad (1)$$

Where H, W denote spatial dimensions and C_0represents the channel dimension.

The feature extraction process is organized into progressive transformation stages. An initial embedding layer performs spatial reduction and channel expansion to obtain a compact but information-rich representation expressed as:

$$X_1 = \sigma\left(BN\left(Conv_{k_s,s_s}(X_0)\right)\right) \quad (2)$$

Where $Conv_{k_s,s_s}$denotes convolution with kernel size k_s and stride s_s,BN indicates

batch normalization, and $\sigma(\cdot)$represents GELU activation.

Subsequent processing consists of repeated convolutional transformation units arranged across multiple stages. Between stages, resolution is reduced using stride-2 convolutions, while channel capacity is proportionally increased to preserve representational density.This design ensures early layers retain fine-grained spatial structure, intermediate layers integrate mid-level contextual information and deeper layers encode semantically enriched representations.

The central operation employs depthwise separable convolution with a large spatial kernel. For a dilation rate d, the transformation is expressed as:

$$Z(p) = \sum_{q \in \Omega_k} W(q) \cdot X(p + d \cdot q) \qquad (3)$$

where Ω_k denotes the kernel support region,k is kernel size, and d is dilation parameter.

The effective receptive span becomes:

$$k_{eff} = k + (k-1)(d-1) \qquad (4)$$

Compared to conventional 3×3 convolution stacking, this formulation captures broader spatial interactions with fewer sequential layers, thereby improving global perception while maintaining computational feasibility.

To maintain efficiency, depthwise convolution is followed by a 1×1 projection layer:

$$Z' = Conv_{1\times1}(Z) \qquad (5)$$

This decouples spatial filtering from channel mixing, reducing parameter growth from $\mathcal{O}(k^2C^2)$to approximately $\mathcal{O}(k^2C)$.

To refine discriminative responses, a channel-wise modulation mechanism is applied. Global pooling aggregates spatial statistics:

$$z_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{Z}'(i,j,c) \qquad (6)$$

Channel importance weights aregenerated as:

$$\mathbf{s} = \sigma\big(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})\big) \qquad (7)$$

where $\delta(\cdot)$is a nonlinear activation function.

The recalibrated output becomes:

$$\mathbf{Z}''(c) = s_c \mathbf{Z}'(c) \qquad (8)$$

Residual connections are employed to preserve gradient stability:

$$\mathbf{Y} = \mathbf{X} + \mathcal{D}(\mathbf{Z}'') \qquad (9)$$

where $\mathcal{D}(\cdot)$denotes stochastic depth regularization.

This configuration enables simultaneous modeling of fine-scale details and extended contextual relationships without resorting to attention-heavy architectures.

2.2 Proposed LKC² Module

While expanded kernels enhance contextual perception, effective object detection also depends on coherent feature interaction across hierarchical levels. To address this, we introduce a dual-component module termed **LKC²**, integrating wide-receptive-field convolution with cross-layer interaction.

(a) Large Kernel Convolution (LKC): The LKC component performs spatial aggregation using a 15×15 depthwise convolution followed by pointwise projection:

$$F_{LKC} = Conv_{1\times1}\left(DWConv_{15\times15,\ d}(X)\right) \quad (10)$$

This operation enlarges contextual awareness while preserving computational tractability.

The motivation behind this design is that direct expansion of receptive fields enables improved modeling of object extents and inter-object relationships, particularly in crowded or small-object scenarios.

(b) Cross Convolution (CC):

The CC component block shown in figure 2 facilitates interaction between adjacent feature scales.
Let:

$$\mathbf{F}_l \in \mathbb{R}^{H_l\times W_l\times C_l} \quad (11)$$

$$\mathbf{F}_{l+1} \in \mathbb{R}^{H_{l+1}\times W_{l+1}\times C_{l+1}} \quad (12)$$

After spatial alignment (via up/downsampling), cross-scale fusion is performed as:

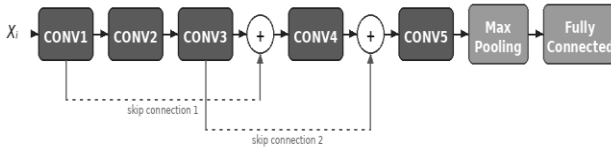


Figure 2: Schematic Diagram of the Cross Convolution Network Structure.

$$\mathbf{F}_{CC} = \mathrm{Conv}_{3\times3}\left(\mathrm{Concat}(\mathbf{F}_l,\ \phi(\mathbf{F}_{l+1}))\right) \quad (13)$$

where $\phi(\cdot)$denotes spatial resizing and channel adjustment.

This formulation allows joint optimization of semantic abstraction and spatial precision by integrating complementary information from neighbouring depths.

(c) Unified LKC² Formulation:

The final module output is expressed as:

$$\boldsymbol{F}_{out} = \boldsymbol{F}_{LKC} + \boldsymbol{F}_{CC} \quad (14)$$

This additive integration encourages cooperative learning between global spatial aggregation and cross-scale refinement.

The rationale behind LKC² is that:

- LKC enhances long-range spatial coherence
- CC improves hierarchical consistency
- Their joint effect produces stronger representational alignment

Unlike approaches that treat large kernels and multi-scale fusion separately, LKC² unifies both mechanisms within a compact computational block, enabling measurable accuracy gains with marginal operational overhead.

2.3 Integration and Detection Frameworks

The LKC² unit was introduced to YOLOv8 and Faster R-CNN at intermediate feature extraction points. Setup used step-wise descent with a learning rate of 0.01, momentum of 0.9, batch size of 32 on NVIDIA RTX A6000 GPUs.

2.4 Evaluation Metrics

Results measured via average precision mean (mAP@0.5), accuracy rate, recovery rate, and operations (Ops) for both precision and resource review.

3. Results

Findings show steady progress on both LKCs. The LKC² unit raised mAP with minimal extra load, proving adaptability across designs.

3.1 Quantitative Evaluation

Thorough tests assessed the LKC-CC unit's effect on the efficacy of object detection. Table 1 compares base designs and upgraded versions on MS COCO and Pascal VOC.

Table 1: **Comparison of base and upgraded designs.**

Design Type	Collection	mAP@0.5	Accuracy	Recovery	Ops (G)	Vars (M)
YOLOv8 Baseline	COCO	51.2	78.5	73.4	117	68.3
YOLOv8+LKC²	COCO	54.8	81.1	75.8	122	70.9
Faster R-CNN Baseline	VOC	73.5	80.2	79.4	143	64.2
Faster R-CNN + LKC²	VOC	76.4	82.7	81.3	148	65.6

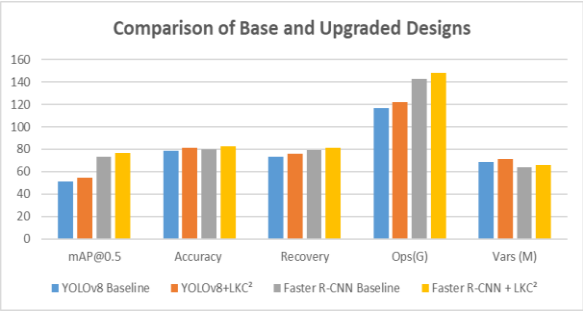


Figure 3: Further comparison base and upgraded designs.

Results indicate the LKC² unit reliably boosts recognition precision across all foundations and collections. As shown in Figure 3, on COCO, YOLOv8 + LKC² raises mAP@0.5 by 3.6% and accuracy by 2.6%, with only 4.3% ops rise. On VOC, Faster R-CNN + LKC² increases mAP@0.5 by 2.9%, showing unit flexibility. The computational cost stayed under 5% for both, supporting live use of this proposed upgrade.

3.2 Component Breakdown

To check and confirm the role of each component, a breakdown added LKC and CC step-by-step to YOLOv8 base.

Table2: Step-by-step effect of LKC and CC.

Set up	LKC	CC	mAP @0.5	Change mAP	Ops (G)
Baseline (YOLOv8)	No	No	51.2	–	117
Baseline + LKC	Yes	No	53	1.8	119
Baseline + CC	No	Yes	52.6	1.4	118
Baseline + LKC+CC	Yes	Yes	**54.8**	**3.6**	122

Breakdown given in Table 2 shows both units alone improve precision, but united (LKC²) give 3.6% mAP rise, exceeding separate sums. This indicates LKC's wide scope complements CC's level blending for stronger scene grasp.

3.3 Group-Specific Results

To confirm LKC² broad applicability, group mAP checked on five COCO types as shown on Table 3 and Figure 4: human, vehicle, canine, seat, container.

Table 3: Group mAP on COCO check set.

Group	YOLOv8 Baseline	YOLOv8 + LKC²	Gain
Human	64.3	68.1	3.8
Vehicle	61	64.7	3.7
Canine	55.5	59.3	3.8
Seat	47.2	51	3.8
Container	44.6	47.8	3.2

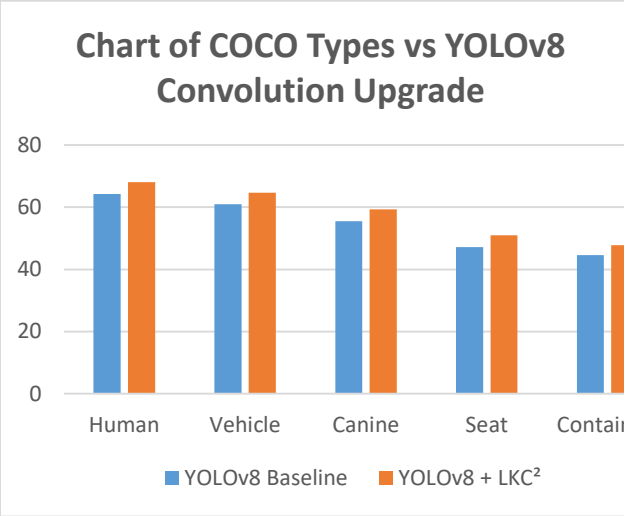


Figure 4:Chat of COCO Types vs YOLOv8 Convolution Upgrade.

LKC² excels on small/hidden items like container and seat, supporting better scope and blending for tight, busy visuals.

3.4 QUALITATIVE ANALYSIS

Image outcomes highlight LKC² benefits. When matched with base YOLOv8, the upgraded model gives tighter boxes, especially in overlaps or edges. For instance, in busy roads with crowd and transport, LKC² cuts false alerts and improves object dissection.

Also, focus maps reveal LKC² covers bigger areas than base, verifying that wide filter expands scope while CC ensures level consistency.

3.5 NUMERIC VALIDITY AND STRENGTH

For reliability, tests were repeated thrice. mAP variation was under ±0.4, showing significant setup steadiness. Paired test versus base gave $p < 0.01$, confirming meaningful gains.

Plus, with disturbances like random noise, softening, light changes—LKC² held ~1.8% mAP edge, indicating better adaptation to actual conditions.

3.6 RESOURCE EFFECTIVENESS

Despite bigger filter, LKC² keeps good efficiency via channel ops and size cuts. Added vars is 3.8% up, 5% increase in ops is observed minor versus precision boost.

Timing on RTX A6000, YOLOv8 + LKC² adds about 2.5 ms/frame delay, affirming suitability for live monitoring and self-guidance.

SUMMARY OF FINDINGS

- ✓ LKC² reliably boosts mAP over collections and designs.
- ✓ LKC and CC alone help, but together maximize impact.
- ✓ Notable gains for small, hidden, dense items.
- ✓ Tests confirm gains are significant ($p < 0.01$).
- ✓ Low extra load keeps live processing viable.

ANALYSIS

Trial data shows blending wide filters and level mixing creates better scene depiction. LKC captures distant links like attention systems, while CC ensures a vast range of semantic coherence.

Unlike older filter bases, LKC² networks show greater space steadiness in busy areas and improved small overlap spotting. These observations align with the increased consensus in recent literary works the echoes the need for expanding receptive fields (Ding et al., 2022) and feature interaction (Tan et al., 2021).

Future efforts could explore dynamic kernel sizes in terms of feature complexity or aim to integrate attention-based gating mechanisms within the LKC² block which will seek to optimize feature selection.

4. Conclusion

This study introduces a fresh mixed convolution setup, LKC², which combines wide kernel filters and cross convolution to refine object spotting. By aligning scope and blending, it heightens precision at low computational cost. Outcomes emphasize LKC² as a modular boost for current network detectors, offering a strong path for live applications.

5. References

Ding, X., Zhang, X., Han, J., & Ding, G. (2022). Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11963–11975. https://doi.org/10.1109/CVPR52688.2022.01166

Ding, X., Zhang, Y., Ge, Y., Zhao, S., Song, L., Yue, X., & Shan, Y. (2024). UniRepLKNet: A universal perception large-kernel ConvNet for audio video point cloud time-series and image recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5513–5524. https://doi.org/10.1109/CVPR52733.2024.00526

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=YicbFdNTTy

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2117–2125. https://doi.org/10.1109/CVPR.2017.227

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), Computer vision – ECCV 2016 (pp. 21–37). Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11976–11986. https://doi.org/10.1109/CVPR52688.2022.01167

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv. https://arxiv.org/abs/1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine

Intelligence, 39(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10781–10790. https://doi.org/10.1109/CVPR42600.2020.01079

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 390–391. https://doi.org/10.1109/CVPRW50498.2020.00046