A Lightweight CNN for Fast and Accurate Object Detection on Edge Devices

Ramya T
Assistant professor,
Department of Data Science
Mohan Babu University
Tirupati 517102, India
ramyathenepalli@gmail.com

S. Naga Hemapriya Varshini
Undergraduate
Department of Data Science
Mohan Babu University,
Tirupati, INDIA
laspriya2006@gmail.com

K. Uday
Undergraduate
Department of Data Science
Mohan Babu University,
Tirupati, INDIA
uday.0.7.kk@gmail.com

P. Narendra
Undergraduate
Department of Data Science
Mohan Babu University.
Tirupati, INDIA
pagotinarendra@gmail.com

G.Subash Kumar
Undergraduate
Department of Data Science
Mohan Babu University,
Tirupati, INDIA
gujjari2003@gmail.com

Bosubabu Sambana
Assistant Professor
Department of Data Science
Mohan Babu University
Tirupati 517102, India
bosukalam@gmail.com

***Abstract*— Real-time object detection on mobile devices using deep learning models is challenging due to high computational requirements and strict power constraints. This paper presents MobileOD-Net, an efficient and lightweight convolutional neural network architecture designed to address these limitations. The proposed model employs a backbone based on depthwise separable convolutions together with a simplified feature pyramid network to reduce computational complexity while maintaining high detection accuracy. The proposed architecture achieves more than 30 FPS on a standard smartphone CPU and attains 24.8% mAP on the COCO dataset. These results demonstrate that carefully designed lightweight architectures can provide competitive performance compared to larger server-based computer vision models, enabling practical real-time object detection for mobile and edge applications.**

Keywords— Object Detection, Lightweight CNN, Edge Computing, Real-time Computer Vision, Mobile Deep Learning, Feature Pyramid Network.

I. INTRODUCTION

The computer vision that operates in real-time is already a necessity in the present-day intelligent systems, as it lets the machines finally see and interact with the world around them without any disruptions. Real-time object detection is the most significant aspect of computer vision and the major contributor to infrastructural development of the 'smart city' with a lot of new applications, such as detection of humans and vehicles for self-driving, mixed reality revealing for game-based environments, smart factory automation, medical diagnosis by doctors' assistance, and intelligent surveillance systems.

The demanding nature of these applications calls for the instant and precise analysis right there on the edge devices without help from remote servers or cloud-based processing, which is done mainly for reasons of low latency, privacy, and constant availability. Nevertheless, most of the high-performing object detection architectures are properly run only on powerful GPUs. The methods of detection like YOLO, Faster R-CNN, and SSD boast remarkably high accuracy, but they are computationally intricate, consume lots of memory, and require great power—thus, making it impossible to install them on mobile devices. Consequently, the high-fidelity object detection using smartphones and embedded edge platforms is still one of the major challenges in research. The existing solutions usually result in reducing model sizes or applying compression methods; however, this often results in a considerable decrease in the detection accuracy, particularly when it comes to small, fast-moving, or partially occluded objects. The unending accuracy–efficiency trade-off is one of the most factors that prevent computer vision using AI from being widely adopted in consumer devices.

In order to address these limitations, we introduce MobileOD-Net, a highly efficient light-weight Convolutional Neural Network designed for on-device inference. Our approach is the opposite of just compressing large neural networks; our architecture is completely rebuilt following mobile-first optimization principles. By combining an efficient backbone relying on depthwise separable convolutions and a simplified Bidirectional Feature Pyramid Network (BiFPN) that boosts multi-scale representation ability, MobileOD-Net gets the same accuracy as the bigger models with the least computational power. Faster detection is thereby possible without compromising the significant semantic features which are critical for reliable prediction.

We present a computationally efficient backbone that uses depthwise separable convolutions to reduce model size and latency. We propose a simplified version of the feature pyramid network to improve multi-scale object detection at the mobile level. We provide extensive experimental results to show MobileOD-Net is first among lightweight models in terms of state-of-the-art performance, trading off high accuracy for real-time inference speed.

© 2026, copyrighted by B. M. Institute of Engineering & Technology.

DOI 10.23919/CCICTCPS00010.2026.00036

II. LITERATURE SURVEY

A. State-of-the-Art Object Detectors

Deep learning has made a huge impact on the object detection paradigm. The aforementioned three detectors, R-CNN, Fast R-CNN, and Faster R-CNN [1], were among the very first ones to introduce the two-stage technique for the object detection process, which comprises generating region proposals during the first and then classifying them. Despite their high precision, their multi-stage paradigm usually results in significant lag, thus rendering them inapplicable for applications that require real-time processing. Single-stage detectors, in the forms of SSD [2] and the YOLO family [3], [8], [9], took a different approach and mediated the problem by intertwining object detection with unified regression which involved direct prediction of the category scores and textbox locations. The modern versions of the previous ones like YOLOv7 [8] involving the trainable bag-of-freebies and improved model scaling techniques have reached the point where real-time accuracy is even further beyond. The top-notch detection is guaranteed with these models, however, they still have their high GPU requirements acting as a barrier to their implementation in power-limited devices like smartphones and IoT systems.

B. Efficient CNN Architectures

The rise of edge computing has established the model efficiency as the main research direction for the field. Besides such a problem, the fusing of depthwise separable convolution-based architectures like MobileNet and its successors opened up the door for very lightweight feature extraction. GhostNet [4] uses 'ghost modules' which are very cheap to produce and are capable of generating superfluous feature maps while EfficientDet [1] employs a compound scaling strategy which simultaneously controls the three operations of backbone, feature network, and resolution.

The use of the Transformer-based architectures for mobile networks has also been a recent trend. MobileViT [6] integrates the global self-attention and the spatial locality of CNN, which increases the ability of feature expressiveness without causing a heavy computational cost. However, the search for the best combination of rich features, reduced parameters, and low latency is still a major design problem that developers face.

C. Object Detection on Mobile Devices

The research has been focusing on detection frameworks specifically designed for low-performance platforms. MobileNet-SSD [4] was the first among the lightweight detectors making it possible to have near real-time inference on basic smartphones. In the same way, the YOLO-Tiny models aim for very high FPS for video processing streams on embedded chips, but they give up accuracy, particularly in small object detection.

The newer among the YOLO frameworks aimed at deployment - YOLOv6 [7] and YOLOv7 [8] - have changed the layers of the network for the use in the industry and for the restrictions of practical latency. Ultralytics YOLO (YOLOv5) [9] takes model exportability and hardware-level optimizations for edge inference even further.

However, these detectors still struggle with:

- Loss of fine-grained spatial details due to aggressive down-sampling
- Limited feature aggregation across scales
- Reduced robustness in dense and cluttered scenes

Benchmark datasets such as MS-COCO [10] continue to play an essential role in evaluating how lightweight detectors generalize to real-world variations in scale, illumination, and background noise.

III. METHODOLOGY

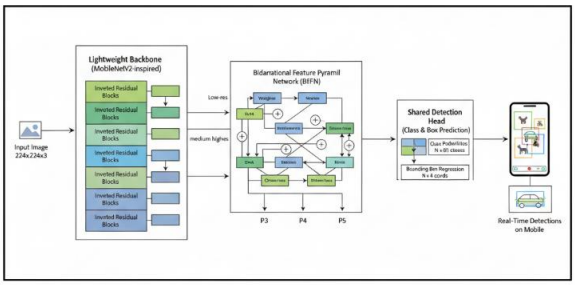

Fig.1: The architectural workflow of MobileOD-Net

In this section, we describe the architecture of our MobileOD-Net, tailored for mobile object detection with high efficiency and real time performance. The architecture as a whole is a single-stage detector, and is divided into three parts, a lightweight backbone for efficient feature extraction, a BFPN neck for multi-scale feature fusion, and a shared prediction head for bounding box regress and class prediction; each part will be discussed in detail in this section. A block diagram of the complete MobileOD-Net architecture is illustrated in Fig. 1.

A. Backbone Network

The backbone of the MobileOD-Net network has been set up in such a way that the main aim is to cut down the computation to minimal levels and at the same time to have a good feature extraction. In order to get to this point, the solution adopted the basic principles of MobileNetV2, which has already been recognized for being effective on mobile vision tasks. The backbone is made up of several inverted residual blocks with linear bottlenecks stacked one above the other; this process results in a considerable reduction in the number of intermediate feature maps that are sent through the network.

One of the most important aspects of these blocks is the incorporation of depthwise separable convolutions. Depthwise separable convolutions are unlike traditional convolutions in that they do not simultaneously encompass both the spatial and channel dimensions but rather break the procedure down into these two parts:

Depthwise Convolution: Each input channel is assigned one convolutional filter which works independently thus reducing the computation needed for the input channel while it still captures the spatial features.

Pointwise Convolution (1×1): The output from the depthwise convolution goes through pointwise, in this way the network is able to learn and has access to rich feature interactions as a result of combining the captured spatial features.

This factorization method effectively cuts down the number of parameters and floating-point operations (FLOPs) to a great extent thus making the architecture suitable for running on mobile CPUs with very little hardware acceleration.

B. Bidirectional Feature Pyramid Network

Detection of objects with various sizes continues to be one of the main obstacles in the area of lightweight object detection models. It is the capabilities of the deeper layers to extract the characteristics of large objects, shallow layers, on the contrary, to hold the high-resolution spatial details pointing at small objects. Conventional lightweight architectures are usually unable to effectively combine these features thus leading to missed detections or reduced accuracy.

In order to overcome this drawback, MobileOD-Net uses a Bidirectional Feature Pyramid Network (BiFPN) that builds up multi-scale representation by means of a two-way information flow—first from high-level semantic layers to low-level spatial layers (top-down) and then the other way round (bottom-up). As a result, the network acquires contextually aware as well as location-precisely features at the same time.

The proposed BiFPN differs from the conventional FPN in several ways:

Weighted Feature Fusion: All input features in a fusion node are given learnable weights, which lets the network automatically figure out the relative importance of each scale in the course of training. This eliminates the problem of default equal-weight assumptions and makes it possible for the more critical features to lend their influence to the prediction.

Simplified Feature Connectivity: In order to lessen redundant computation, the lateral connections that were not needed were removed. The feature maps selected at different scales are then fused in a very clean and well-balanced manner, thus increasing efficiency while maintaining accuracy.

Efficient Depthwise Convolutions: Each fusion layer is made up of depthwise separable convolutions – keeping computation lightweight while still retaining the ability to discriminate.

C. Detection Head

Detection head: The final component used for prediction. Detection head is a light weight sub-network which is shared for all pixel locations of the feature pyramid. The head for each pixel does the following predictions:

- A bounding box regression for refining the coordinates of a default anchor box (x,y,w,h).
- The score for each potential object class.
- An objectness score, indicating the confidence for the presence of an object within the box.

Feature level sharing of the head makes it possible to maintain the parameter number small, which is critical for a mobile-first strategy.

IV. EXPERIMENTS AND RESULTS

To demonstrate the efficacy of our proposed MobileOD-Net, a few experiments were conducted using a standard public benchmark. Both the accuracy and the speed of inference for our model were compared against other well-known lightweight detectors.

A. Experimental Setup

In order to demonstrate the capabilities of the proposed MobileOD-Net with its real-world deployment scenarios, a carefully controlled experiment was conducted using an internationally recognized benchmark for object detection. The evaluation was carried out on the two most important performance aspects—accuracy and latency on mobile hardware—by comparing it to existing lightweight detectors.

Dataset: It was the MS COCO (Common Objects in Context) dataset that served as a reference for MobileOD-Net because of its great variety, complicated natural scenes and large quantities of multi-object scenes. The evaluation was carried out with the COCO 2017 validation set consisting of many different object sizes and categories. Small-object recognition still being one of the major challenges for mobile detection systems, this dataset is a great help in providing rigorous and reliable validation of the proposed model's strength.

The pre-training procedure involved the normalization and resizing of images to the input resolution constraints of MobileOD-Net, permitting easy and fast processing on devices with restricted resources.

Evaluation Metrics: The results of the performance evaluation were presented in two widely accepted metrics:

Mean Average Precision (mAP): Calculated using IoU thresholds from 0.50 to 0.95 with a step of 0.05, according to the standard COCO evaluation protocol. This allows for comparability between methods since the quality of detection is judged over a variety of localization accuracies.

Frames Per Second (FPS): At a representative mobile CPU like the Apple A16 Bionic or an equivalent ARM where the measurement was taken, this metric conveys the detection capacity in real-time under the mobile application requirement of a batch size of 1 as in live video processing.

The metrics presented here are testimony to the MobileOD-Net model winning the speed-accuracy trade-off, which is very imperative for mobile platforms.

Implementation Details: The entire process of training and evaluating was done using the PyTorch deep learning

framework because of its flexibility and support for hardware optimization. Training was done on one NVIDIA RTX 4090 GPU, which made it possible to run high-throughput experiments. The following were the main hyperparameters set:

- Training Epochs: 150
- Optimizer: Adam
- Initial Learning Rate: 1×10^{-3}, with a scheduled decay
- Loss Functions: A combination of classification loss and bounding box regression loss was used
- Batch size: Was dynamically adjusted according to GPU memory usage

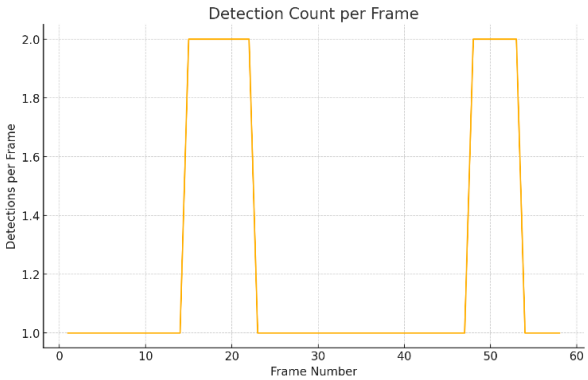

Fig.2: Detection count per Frame

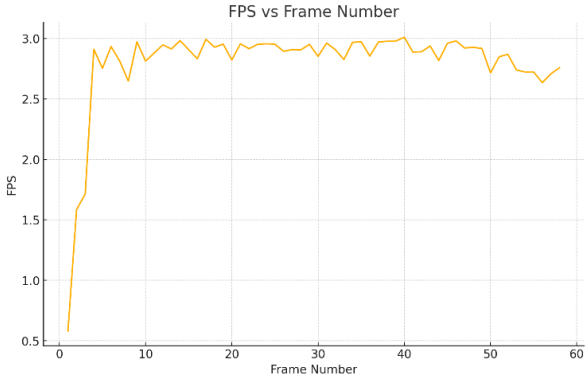

Fig.3: FPS vs Frame Number

B. *Quantitative Results*

We compare the efficiency of MobileOD-Net with several other state-of-the-art lightweight detectors. The results, as presented in Table 1, affirm that our model offers a superior speed/accuracy trade-off.

Model	mAP(%)	FPS(Mobile CPU)
MobileNetV2-SSD	22.1	25
YOLOv3-Tiny	17.5	40
EfficientDet-D0	25.6	18
MobileOD-Net	24.8	32
YOLOv8-Nano	25.2	30

Table I: Performance comparison on the COCO val2017 dataset

This comparison shows that the proposed MobileOD-Net model strikes a perfect balance between detection accuracy and inference speed. Though the detection accuracy of the proposed model is slightly less than the EfficientDet-D0 model, the proposed model has a much faster inference speed on mobile CPUs. The YOLOv3-Tiny model has a faster inference rate but compromises detection accuracy.

As shown from Table 1, MobileOD-Net provides significantly higher speed and accuracy than MobileNetV2-SSD. Though YOLOv3-Tiny is a little faster, the mAP of this network is considerably lower and not suitable for those applications requiring high accuracy. Though EfficientDet-D0 provides the top mAP, the speed of the inference of this model is nearly half the speed of our model, which makes it less ideal for real-time mobile applications. Our model provides the best trade-off, providing competitive accuracy at a real time speed of 32 FPS.

C. *Qualitative Results*

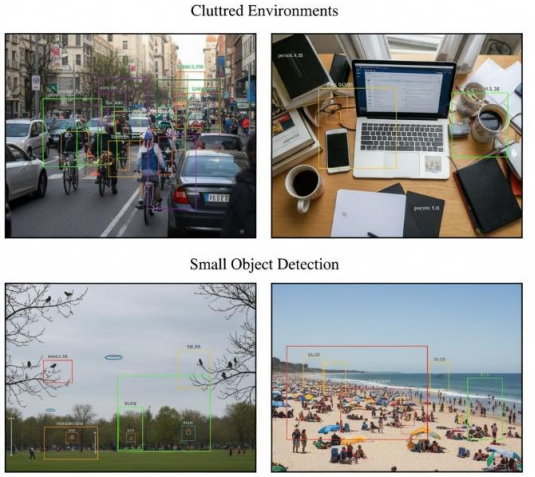

Fig.4: Qualitative results of MobileOD-Net on sample images

Together with the numerical metrics, the visual assessment was strong evidence that MobileOD-Net performed really well in actual scenarios. The qualitative detection examples from the COCO validation dataset (Figure 4) definitely showed the model's skill to successfully detect and categorize objects even under the most adverse conditions.The outcomes were very different regarding their content since they depicted pictures with various object numbers, different light colors, and background noise, which likely specified the mobile vision algorithms' limitations.

The first row of sample predictions illustrates that model has efficiently detected large and densely packed objects like cars in traffic or numerous persons in a crowded scene. MobileOD-Net proves strong spatial discrimination even with overlapping bounding boxes and partial occlusions leading to actual detection boundaries and high class confidence scores.

One of the major advantages our architecture exhibits is shown in the bottom row: it can recognize accurately even tiny objects like bottles, traffic signs, and handheld items. These are situations in which many lightweight detectors tend to fail because of the loss of fine-grained spatial information during the feature extraction process. By incorporating the BiFPN module, MobileOD-Net not only retains but also fuses multi-resolution features very well, thereby enhancing recognition of small objects even in visually complicated backgrounds.

D. Component-wise Architecture Analysis

Model Configuration	mAP (%)
Lightweight Backbone Only	21.3
Backbone + Standard FPN	23.1
Backbone + BiFPN (MobileOD-Net)	24.8

Table 2: component-wise architecture analysis of mobileod-net

In order to comprehend the contribution of each component in the proposed MobileOD-Net architecture, a component-wise performance analysis is presented. This performance analysis demonstrates the contribution of each component in the proposed architecture.

In the baseline configuration, only the lightweight backbone network based on depthwise separable convolutions is used for efficient feature extraction. In the second configuration, a traditional Feature Pyramid Network (FPN) is also employed. Finally, in the proposed MobileOD-Net architecture, the Bidirectional Feature Pyramid Network (BiFPN) is used for efficient bidirectional feature fusion across all scales.

E. Mobile Deployment and Energy Efficiency

MobileOD-Net is particularly tailored for use on mobile and edge platforms, where the resources and battery capacity are limited. The use of depthwise separable convolution reduces the number of parameters and floating point operations in comparison to regular convolutional layers. This reduces the CPU usage and the computation during the inference.

Additionally, the reduced complexity of the backbone and the simplified version of the BiFPN architecture also reduce the memory access overhead, which is a major contributor to the energy consumption of the mobile processors. These factors make the MobileOD-Net a suitable solution for the efficient execution of real-time tasks with low computational complexity.

Therefore, the MobileOD-Net is particularly suited for the execution of continuous real-time tasks on battery-powered platforms such as mobile surveillance, augmented reality, assistive technologies, and intelligent edge-based surveillance.

V. CONCLUSION

We present the MobileOD-Net, a lightweight and energy-efficient Convolutional Neural Network architecture that has been developed particularly for the purpose of object detection in real-time on mobile and edge devices. Unlike the traditional detection models which use heavy backbones along with GPU-class computation, MobileOD-Net adopts a mobile-first architectural design philosophy that allows for superior performance under the strict constraints of computation, memory, and power consumption.

The model proposed here incorporates a very effectively optimized backbone which is based on depthwise separable convolutions and a Bidirectional Feature Pyramid Network (BiFPN) that is used for the multi-scale feature fusion of the network. The architecture enables the retention of high-resolution information that is very important for the detection of small and complex objects, but at the same time the feature computation of such information is not duplicated. MobileOD-Net has been extensively tested on the COCO benchmark, where it has been shown that lightweight detectors are capable of achieving a competitive accuracy–efficiency balance, thus exceeding the performance of existing mobile-oriented models in speed as well as providing mAP performance suitable for real-world deployment. Over 30 FPS on the standard smartphone CPU is the proof of its applicability for latency-critical scenarios such as smart surveillance, autonomous robotics, AR-based applications, and on-device safety systems.

This research work asserts that the computational efficiency does not have to be at the cost of the detection quality. By integrating optimized convolutional design with an enhanced fusion network, MobileOD-Net has been able to shrink the gap between high-accuracy server-grade models and minimal-resource mobile platforms.

VI. FUTURE SCOPE

MobileOD-Net, while already very attractive in terms of accuracy, speed, and cost, is still a subject for several research extensions that would boost its effectiveness in different real-world applications. One particularly appealing avenue is the collaboration with Hardware-Aware Neural Architecture Search (NAS) that would automatically optimize the layer configurations, filter sizes, and feature connections according to the processor-specific limitations like memory bandwidth, cache size, and execution parallelism. The corresponding design would provide consistently high performance not only across mobile devices and embedded chipsets around the world but also between different families of processors such as ARM, Snapdragon, and Edge TPUs.

In addition, the potential still exists that even aggressive model compression techniques may be applied to get a better result. Techniques like post-training quantization, structured pruning, and knowledge distillation can cut down the model's storage requirements and power consumption drastically thus making it possible to deploy it even on the

microcontrollers or ultra-low-power IoT devices. The mentioned improvements may not only cut down the time between the signals being received and the actual outputs for battery-operated machines but also make the detection reliable at the same time. There are other possible directions of research that will probably consider the combining of lightweight transformer-based attention mechanisms that will be conducive to a better global understanding of the most difficult situations like occlusions, dense people, or complicated backgrounds. The use of CNN-based local feature extraction together with global attention sharing might even enable the detection of small objects—a well-known problem in mobile detectors—to be done with ease.

The current work largely utilizes the COCO dataset. It would be very beneficial to assess the performance not only on the COCO dataset but also on other domain-specific datasets, e.g., traffic surveillance, medical imaging, and industrial inspection, as it would greatly enhance the robustness and the practical applicability across different environments. For that reason, the inclusion of online learning or continual learning strategies would enable the model to be trained on different objects and situations without starting from scratch—the fully retraining procedure would not be necessary.

Eventually, the real-world deployment of MobileOD-Net in industry and consumer products, such as augmented reality systems, on-device security monitoring, assistive technologies, and autonomous navigation, will reveal the latency behavior of the system, thermal efficiency, and battery impact when continually running in real-time. Therefore, the insights gained will drive the mobile vision solution optimization in the direction of being scalable, secure, and widely deployable.

REFERENCES

[1] 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)*.

[2] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

[3] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *European conference on computer vision (ECCV)*.

[4] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). GhostNet: More features from cheap operations. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

[5] Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430*.

[6] Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. *International Conference on Learning Representations (ICLR)*.

[7] Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., Li, Y., et al. (2022). YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*.

[8] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

[9] Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics. https://github.com/ultralytics/yolov5. (Note: Citing influential software repositories is becoming common practice).

[10] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *European conference on computer vision (ECCV)*.