WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Real-Time Object Detection in Video Sequences: Techniques, Challenges, and Future Directions

[1]Khalid M.A. Abufayyadh

1Student, 2Assistnat Professor
Masters of Computer Applications, School of CS & IT, Jain (Deemed-To-Be-University), Bangalore, India,
[1]k.abufayyad1@gmail.com

A B S T R A C T

Real-time object detection in video sequences is pivotal area of research in computer vision. The applications span autonomous driving. Surveillance and healthcare also benefit. This paper provides comprehensive review of current state-of-the-art techniques used for detecting and tracking objects in video data. It highlights both classical methods and recent advances driven by deep learning. We delve into unique challenges. These challenges are posed by time-domain data exemplified by motion blur and occlusions. We also explore necessity for low-latency processing.

Key methodologies are discussed. They include convolutional neural networks (CNNs). Recurrent neural networks (RNNs) are another focus. Transformer-based models are also analyzed. Each offers distinct advantages for handling temporal dependencies. They also manage dynamic environments efficiently. Furthermore, the paper examines practical applications. It illustrates how techniques are employed across various domains. Such implementation enhances both performance and reliability.

Finally we outline future research directions. Emphasizing need for improved robustness and efficiency. And generalization in real-time object detection systems. Through this review. We aim to provide detailed understanding of the field's current landscape. Simulate further advancements. In this crucial area of computer vision

Keywords: Real-time object detection, Video sequences, Computer vision, Deep learning, Convolutional neural networks (CNNs), Transformer-based models, Temporal coherence, Autonomous driving, Healthcare applications, Time-domain data, Dynamic environments

INTRODUCTION

Real-time object detection in video sequences is cornerstone of modern computer vision research. It provides critical functionality across wide range of domains. From autonomous vehicles. Navigating complex urban environments. To surveillance systems monitoring public spaces. The ability to detect and track objects in real-time is essential. This enables intelligent decision-making and interaction with dynamic environments. Furthermore in medical imaging healthcare. Real-time object detection plays vital role in assisting surgeons during procedures. It's crucial for monitoring patient vital signs and for facilitating analysis of medical imaging data.

The complexity of real-time object detection in video sequences arises from continuous and dynamic nature of video data. Unlike static images. Objects are typically isolated and motionless. Video sequences capture objects in motion. Often undergoing complex transformations such as changes in size and shape. Also orientation over time. As a result. Real-time object detection algorithms must be capable of analyzing and interpreting temporal relationships. These can be between objects across multiple frames. Allowing accurate identification. And tracking as objects move through scene.

This allows detection of object motion and movement patterns. Background subtraction techniques identify moving objects by subtracting static background model from each frame. It highlights regions of change corresponding to potential objects of interest. Feature-based tracking methods identify and track specific visual features. Features such as corners or edges across frames allow tracking of objects with distinctive visual characteristics.

While traditional methods have proven effective in certain scenarios they often struggle with complex scenes occlusions. Variations in lighting conditions can be challenging. In recent years, advent of deep learning has revolutionized real-time object detection. This has enabled development of highly accurate efficient detection systems.

Convolutional Neural Networks (CNNs) have emerged as cornerstone of deep learning-based object detection. They leverage hierarchical feature representations to detect objects in images. Also in video frames. CNN-based object detection systems include Faster R-CNN YOLO (You Only Look

Once) and SSD (Single Shot MultiBox Detector). These have achieved remarkable success in real-time applications. They offer a balance between accuracy and efficiency.

In addition to CNNs Recurrent Neural Networks (RNNs) have been employed to capture temporal dependencies in video sequences allowing for tracking of objects over time. Long Short-Term Memory (LSTM) networks have demonstrated effectiveness in modeling long-range dependencies. They also capture complex temporal relationships. By integrating CNNs with LSTM networks researchers have developed powerful architectures. These are for real-time video object detection. Enabling accurate tracking. They also assist in prediction of object trajectories.

More recently transformer-based models gained popularity for the ability to capture long-range dependencies. They also parallelize computation across sequences. Models such as DETR (Detection Transformer) and Video Swin Transformer demonstrated impressive performance. They tackle real-time object detection tasks. Leveraging attention mechanisms, this facilitates processing of entire sequences in parallel. State-of-the-art results are achieved

Despite advancements made in real-time object detection several challenges remain. Motion blur. Occlusions variations in lighting conditions degrade detection performance. Requiring robust algorithms capable of handling these challenges. Additionally achieving low-latency processing is essential for real-time applications. Necessitating efficient algorithms and hardware optimizations to meet stringent timing requirements. Furthermore addressing issues such as data imbalance and domain shift is critical for ensuring generalization. And reliability of real-time object detection systems across diverse environments and scenarios.

Looking ahead future research in real-time object detection is poised to address these challenges. It will push boundaries of performance and efficiency. Promising directions include exploration of self-supervised learning techniques. It will leverage unlabeled video data. Multi-modal fusion approaches will integrate information from diverse sensor modalities. Incremental learning strategies will adapt models. This helps them to evolve with environments continuously. Through continued innovation and collaboration real-time object detection systems will play central role in enabling intelligent decision-making. They will facilitate interaction with dynamic environments. This will drive progress across wide range of domains and applications.

Background and Importance

Real-time object detection in video sequences is of paramount importance across wide range of applications including autonomous driving security surveillance and healthcare. In autonomous driving vehicles must continuously detect and track other vehicles. Pedestrians. Obstacles. To navigate safely. Delays or inaccuracies in detection can lead to catastrophic outcomes. This highlights critical role of real-time object detection systems. In ensuring safety of passengers pedestrians alike.

In healthcare settings, real-time object detection plays vital role in patient monitoring surgical assistance and medical imaging analysis. For instance in patient monitoring applications, real-time detection of vital signs. Patient movements helps healthcare providers identify potential medical emergencies. This enables timely interventions. In surgical assistance applications real-time detection of surgical instruments. Anatomical structures aids surgeons. This assistance allows them to perform complex procedures with precision

The evolution of object detection techniques has been marked by significant advancements. This particularly true with advent of deep learning. Traditional methods like optical flow. Background subtraction laid groundwork for early developments. However these methods often struggled to cope with complexities of real-world scenarios. Occlusions varying lighting conditions and non-rigid object deformations posed significant challenges.

Deep learning has revolutionized field of object detection. It does so by providing powerful tools for automatically learning discriminative features from raw data. Convolutional Neural Networks (CNNs) have emerged as cornerstone of many state-of-the-art object detection systems. They offer superior performance in terms of accuracy. And robustness.

Two-stage detectors like Faster R-CNN. Single-stage detectors like YOLO (You Only Look Once) SSD (Single Shot MultiBox Detector). All have achieved remarkable success. These models perform exceptionally well in real-time applications. They provide balance. Between accuracy and efficiency.

Despite significant advancements enabled by deep learning real-time object detection still faces several challenges. These challenges include motion blur. Occlusions. Variations in lighting conditions. Data imbalance. Motion blur can obscure object details. It makes detection difficult. Especially in fast-moving scenes. Occlusions occur when objects overlap. Or are blocked by other objects. Leading to partial or complete loss of visibility.

Addressing these challenges and advancing state-of-art in real-time object detection requires innovative research. Development efforts. Future directions in this field may include exploration of novel deep learning architectures. Development of robust and efficient algorithms. Handling motion blur and occlusions. Integration of multi-modal sensor data. To improve detection accuracy and reliability. Through continued innovation and collaboration. Real-time object detection systems will continue to play critical role. Enabling intelligent decision-making. Interaction with dynamic environments across various domains and applications

2.REVIEW LITERATURE

Study	Methodology/Approach	Challenges Addressed	Key Findings/Contributions
Redmon et al. (2016) [1]	YOLO (You Only Look Once)	Real-time processing, object detection in varying scales, accuracy vs. speed trade-off	Introduced YOLO, achieving real-time object detection with high accuracy on a single pass through the network.
Ren et al. (2015) [2]	Faster R-CNN (Region-based Convolutional Neural Networks)	Accuracy and efficiency in object detection, multi-scale object detection	Proposed a two-stage object detection framework, achieving state-of-the-art results in accuracy and speed.
Feichtenhofer et al. (2016) [3]	Two-Stream Convolutional Networks	Temporal modeling for action recognition in videos, integration of spatial and temporal features	Introduced a two-stream architecture for video-based action recognition, incorporating both spatial and temporal information.
Carion et al. (2020) [4]	DETR (Detection Transformer)	Long-range dependency modeling, object detection in complex scenes, parallel processing of sequences	Proposed a transformer-based model for object detection, demonstrating strong performance in complex scenes.
Simonyan and Zisserman (2014) [5]	Two-Stream Convolutional Networks	Action recognition in videos, spatial and temporal feature extraction	Presented a two-stream architecture for action recognition, achieving state-of-the-art performance on benchmark datasets.
He et al. (2016) [6]	Deep Residual Learning	Addressing vanishing gradient problem, training deeper neural networks	Introduced residual learning, enabling the training of very deep neural networks with improved performance.
Vaswani et al. (2017) [7]	Attention is All You Need (Transformer)	Sequence modeling, parallel processing of sequences, long-range dependency modeling	Proposed a transformer-based model for sequence transduction tasks, achieving state-of-the-art performance on various tasks.

Emphasizing intricate aspects systemically unravels multifaceted features existing within the subject under scrutiny. Systematic dissection of each facet elucidates inherent complexities. Simultaneously unearths underlying principles. Inevitably advances comprehension. Moreover in-depth exploration is necessary enhances a scholarly understanding. It initiates a profound discourse permeability.

The intricacies operational frameworks necessitate rigorous examination. Focus should remain on interrelations among components. Their collective interaction reveals nuanced dynamics that define systemic behavior. Enhanced comprehension emerges. Sequential analysis each segment allows delineation of functional interdependencies. Holistic perspective fosters augmentation of theoretical foundations. Enabling. Practical applications extend toward empirical validations.

Methodological rigor essential. Employment of advanced analytical tools is imperative. Quantitative models and qualitative assessments substantiate findings. Integration of data-driven insights augments relevance. Facilitated. Further, incorporation diverse perspectives infuses study with multidimensional depth. Interdisciplinary collaboration becomes cornerstone. Also enriches contextual understanding.

Transformative potential resides in this endeavor. Contributions extend beyond academic precincts. Encompassing sociocultural and technological spheres. Promoting progressive paradigms foundational initiative catalyze innovation myriad domains. Resulting in paradigm shifts elicited by knowledge proliferation.

Study: Redmon et al. (2016)

Methodology/Approach: YOLO (You Only Look Once)

YOLO is real-time object detection system that processes images in single pass through a convolutional neural network. It divides input image into a grid. Predicts bounding boxes. Class probabilities for each grid cell.

Challenges Addressed: Real-time processing. Object detection in varying scales. Accuracy vs. speed trade-off

YOLO addresses challenge of real-time processing by optimizing network architecture for speed. While maintaining high accuracy. In object detection across different scales.

Key Findings/Contributions: Introduced YOLO. Achieving real-time object detection with high accuracy on single pass through network.

YOLO significantly reduces computational complexity. Compared to traditional object detection methods. Making it suitable for real-time applications. Such as autonomous driving. And surveillance.

Study: Ren et al. (2015)

Methodology/Approach: Faster R-CNN (Region-based Convolutional Neural Networks)

Faster R-CNN is two-stage object detection framework. It first generates region proposals using Region Proposal Network (RPN) and then classifies these proposals using region-based CNN.

Challenges Addressed: Accuracy and efficiency in object detection. Multi-scale object detection

Faster R-CNN achieves high accuracy and efficiency. Decoupling region proposal generation from object classification. This allows for multi-scale object detection without sacrificing speed.

Key Findings/Contributions: Proposed two-stage object detection framework achieving state-of-the-art results in accuracy, speed.

Faster R-CNN significantly improves upon previous object detection methods. Introducing unified framework for region proposal generation and object classification. Leading to superior performance on benchmark datasets.

Study: Feichtenhofer et al. (2016)

Methodology/Approach: Two-Stream Convolutional Networks

Two-Stream Convolutional Networks consist of two separate CNNs. One processes spatial information (RGB frames). Other processes temporal information (optical flow or motion vectors). These are fused to extract spatial and temporal features.

Challenges Addressed: Temporal modeling for action recognition in videos integration of spatial and temporal features

Two-Stream Convolutional Networks address challenge of action recognition in videos. Capturing both spatial and temporal information enables more robust and accurate recognition of complex actions.

Key Findings/Contributions: Introduced two-stream architecture for video-based action recognition, incorporating both spatial and temporal information.

Two-Stream Convolutional Networks significantly improve action recognition performance. Leveraging both spatial and temporal cues in video data achieves state-of-the-art results on action recognition benchmarks.

Study: Carion et al. (2020)

Methodology/Approach: DETR (Detection Transformer)

DETR is transformer-based object detection model that casts object detection as set prediction problem. It directly predicts object bounding boxes and class labels from sequence of image features.

Challenges Addressed: Long-range dependency modeling object detection in complex scenes. Parallel processing of sequences

DETR addresses challenge of object detection in complex scenes. It captures long-range dependencies. Computation is parallelized across sequences. This enables efficient and accurate object detection.

Key Findings/Contributions: Proposed transformer-based model for object detection. Demonstrated strong performance in complex scenes.

DETR introduces novel approach to object detection. It demonstrates that transformers can handle long-range dependencies in visual data and achieve competitive performance with traditional CNN-based methods.

Study: Simonyan and Zisserman (2014)

Methodology/Approach: Two-Stream Convolutional Networks

Similar to Feichtenhofer et al. (2016) Simonyan and Zisserman propose two-stream architecture for action recognition in videos. They use separate streams for spatial and temporal information processing.

Challenges Addressed: Action recognition in videos. Spatial and temporal feature extraction

Study addresses challenge of action recognition. By extracting both spatial and temporal features from video data, it improves the discriminative power and robustness of action recognition models.

Key Findings/Contributions: Presented two-stream architecture for action recognition. Achieving state-of-the-art performance on benchmark datasets.

Proposed two-stream architecture significantly advances state-of-the-art in action recognition. It demonstrates effectiveness of integrating spatial and temporal information for video analysis tasks.

Study: He et al. (2016)

Methodology/Approach: Deep Residual Learning

Deep Residual Learning introduces residual learning where residual connections added to neural network architectures mitigate the vanishing gradient problem. Enable training of very deep networks.

Challenges Addressed: Addressing vanishing gradient problem. Training deeper neural networks

The study addresses the challenge of training deep neural networks by introducing residual connections. These allow for training of networks with hundreds of layers without suffering from vanishing gradients.

Key Findings/Contributions: Introduced residual learning. Enabled the training of very deep neural networks with improved performance.

Deep Residual Learning revolutionizes the field of deep learning. It enables training of extremely deep neural networks. This leads to significant improvements in performance. These improvements span various tasks, including object detection and image classification.

Study: Vaswani et al. (2017)

Methodology/Approach: Attention is All You Need (Transformer)

Attention is All You Need introduces transformer architecture for sequence transduction tasks. It relies solely on self-attention mechanisms to capture dependencies between input/output sequences.

Challenges Addressed: Sequence modeling parallel processing of sequences, long-range dependency modeling

The study addresses challenge of sequence modeling by proposing a transformer-based model. This can efficiently capture long-range dependencies. It can parallelize computation across sequences.

Key Findings/Contributions: Proposed transformer-based model for sequence transduction tasks achieving state-of-the-art performance on various tasks.

Attention is All You Need demonstrates effectiveness of transformer architectures for sequence modeling tasks. Surpassing previous approaches.

3.TOOLS & ALGORITHMS

Tools:

TensorFlow: Open-source deep learning framework developed by Google. Widely used for building and training deep neural network models. Including those for object detection.

PyTorch: Another popular deep learning framework. Developed by Facebook. Known for its flexibility. Ease of use in building and training neural network models.

OpenCV: Open-source computer vision library providing various tools and functions for image and video processing. Including object detection algorithms.

CUDA: Parallel computing platform. Application programming interface model developed by NVIDIA. Used to leverage power of NVIDIA GPUs for accelerating deep learning computations.

Caffe: Deep learning framework developed by Berkeley AI Research (BAIR). Particularly suited for image classification. Segmentation tasks. Also used for object detection.

Methods:

Convolutional Neural Networks (CNNs): Deep learning architectures specifically designed for processing visual data. Consisting multiple layers of convolutional and pooling operations. These operations are used for feature extraction.

Recurrent Neural Networks (RNNs): Neural network architectures capable of processing sequential data by maintaining internal memory states. These are commonly used for capturing temporal dependencies. Particularly in video sequences.

Transformer Networks: Attention-based neural network architectures originally designed for sequence-to-sequence tasks in natural language processing. These networks have been recently adapted for processing video sequences. They capture long-range dependencies effectively.

Two-Stream Networks: Architectures that process both spatial and temporal information separately. Typically consisting two parallel streams of CNNs. One for spatial features. The other for temporal features.

Region-based Methods: Object detection approaches that first generate region proposals candidate object bounding boxes These methods classify and refine these proposals using deep learning models.

Algorithms:

YOLO (You Only Look Once): Real-time object detection algorithm. It predicts object bounding boxes and class probabilities directly from image pixels. Achieves this feat in single pass through convolutional neural network.

Faster R-CNN (Region-based Convolutional Neural Networks): Two-stage object detection algorithm. First generates region proposals using Region Proposal Network (RPN) Then classifies and refines these proposals. This is done using region-based CNN.

SSD (Single Shot MultiBox Detector): Single-stage object detection algorithm. Directly predicts object bounding boxes and class probabilities from feature maps at multiple scales. It achieves real-time performance.

DETR (Detection Transformer): Transformer-based object detection algorithm. Casts object detection as set prediction problem directly predicting object bounding boxes and class labels. From sequence of image features.

Mask R-CNN: Extension of Faster R-CNN. Also predicts object masks. Adds function in addition to bounding boxes and class labels enabling instance segmentation. Avoid hyphenation at the end of a line. Symbols denoting vectors and matrices should be indicated in bold type. Scalar variable names should normally be expressed using italics. Weights and measures should be expressed in SI units. All non-standard abbreviations or symbols must be defined when first mentioned, or a glossary provided.

4.CONCLUSION & FUTURE

In conclusion real-time object detection in video sequences is crucial area of research with wide-ranging applications in fields such as autonomous driving. Security surveillance finds utility in healthcare. The evolution of object detection techniques propelled by advancements in deep learning has revolutionized the field. It enabled the development of highly accurate and efficient detection systems. From pioneering algorithms like YOLO and Faster R-CNN. Recent innovations such as DETR and transformer-based models have emerged. Researchers have continually pushed the boundaries of performance and efficiency in real-time object detection.

Despite these advancements challenges remain. These include handling motion blur and occlusions. Variations in lighting conditions and data imbalance also pose significant issues. Future research directions may focus on novel deep learning architectures. Efficient algorithms for handling complex scenes are critical. Integration of multi-modal sensor data to improve detection accuracy and reliability is also essential.

Overall real-time object detection in video sequences continues to be dynamic and rapidly evolving field. It is driven by quest to develop intelligent systems. These systems are capable of perceiving and interacting with environment in real time. Through continued innovation researchers and practitioners will continue push boundaries of what is possible. Unlocking new opportunities. Applications for real time object detection technologies will expand.

Future:

Improving Accuracy and Robustness: Future research efforts may focus on further improving accuracy robustness of real-time object detection systems. Particularly in challenging scenarios. Such as low-light conditions, adverse weather crowded environments. This could involve development of more sophisticated algorithms. For handling occlusions, partial visibility and object interactions.

Multi-Modal Fusion: Integrating information from multiple sensor modalities. Such as visual LiDAR, radar and thermal sensors promises improving accuracy and robustness of real-time object detection systems. Future research may explore methods for effectively fusing information from diverse sensor modalities. These efforts aim to enhance object detection performance in complex real-world scenarios.

Continual Learning and Adaptation: Real-world environments are dynamic. They constantly evolve. This necessitates real-time object detection systems that can adapt. Systems must learn from new data over time. Future research may focus on developing algorithms for continual learning. Adaptation can enable object detection systems adapt to changes in environment. Hence such systems will handle concept drift. They will maintain high performance periods.

Privacy-Preserving Object Detection: Concerns about privacy and data security continue to grow. There is need for real-time object detection algorithms that can operate while preserving privacy of individuals. They must also protect sensitive information. Future research may explore techniques for privacy-preserving object detection such as federated learning. Differential privacy and on-device processing ensure that personal data is protected. This protection is needed. It will still enable effective object detection.

Ethical Considerations and Bias Mitigation: Addressing ethical considerations and mitigating bias in real-time object detection systems is crucial for ensuring fairness. Transparency. Accountability are equally important. Future research may focus on developing methods for detecting and mitigating biases in training data. Ensuring that object detection systems are fair and equitable. This must be across different demographic groups and social contexts.

Application-Specific Solutions: As real-time object detection finds applications in diverse domains such as autonomous driving healthcare smart cities and robotics. Future research may focus on developing application-specific solutions. These solutions must be tailored to unique requirements and constraints of each domain. This could involve customizing algorithms. Sensor configurations. Deployment strategies to optimize performance. Address domain-specific challenges

References

1. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

2. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Advances in Neural Information Processing Systems (NeurIPS).

3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. In European Conference on Computer Vision (ECCV).

4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

5. Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional Two-Stream Network Fusion for Video Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems (NeurIPS).

7. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In European Conference on Computer Vision (ECCV).

8. Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. In Advances in Neural Information Processing Systems (NeurIPS).

9. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

10. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

11. Law, H., & Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV).

12. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., & Cottrell, G. (2019). Understanding convolution for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

13. Zhu, Y., Zhou, Q., Ye, Q., & Qiao, Y. (2018). Soft-nms--improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

14. Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

15. Lin, T. Y., Maire, M., Belongie, S., Bourdev, L. D., Girshick, R., Hays, J., ... & Ramanan, D. (2014). Microsoft coco: Common objects in context. In European conference on computer vision (ECCV).

16. Liu, W., Anguelov, D., Erhan, D., & Szegedy, C. (2016). Ssd: Single shot multibox detector. In European conference on computer vision (ECCV).

17. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

18. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., ... & Murphy, K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).