Archives available at journals.mriindia.com
International Journal on Advanced Computer Theory and Engineering
ISSN: 2319 - 2526
Volume 15 Issue 01s, 2026

Recent Advances in Object Detection, Segmentation, Integration and Relationship Detection with Special Reference to Enhanced Scene Understanding

[1]Ramchandra Terkhedkar, [2]Manoj Mhaske, [3]Pravin Yannawar
[1,2,3] Vision and Intelligent System Laboratory, Department of Computer Science & Information Technology
Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajinagar, India
Email: [1]terkhedkar04@gmail.com, [2]mhaskemanoj@gmail.com, [3]plyannawar.csit@bamu.ac.in

Peer Review Information

Submission: 08 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

object detection, image segmentation, scene understanding, deep learning, transformers, scene graphs, multimodal integration

Abstract

Understanding visual scenes comprehensively remains a central challenge in Computer vision and artificial intelligence. The field has witnessed tremendous evolution from traditional feature-based methods to deep learning architectures capable of simultaneous object detection, precise segmentation and complex relationship modeling. This review synthesizes recent developments across these interconnected domains, with particular emphasis on the YOLO family evolution through YOLOv10, transformer-based detection frameworks including DETR and its variants, advanced segmentation models such as SAM and HQ-SAMand scene graph generation techniques. We examine how multi-task learning and multimodal integration strategies are reshaping scene understanding capabilities. Critical analysis of current limitations—including Computational efficiency, domain generalization, data imbalance and interpretability—guides our discussion of emerging research directions. Foundation models, efficient transformers and zero-shot learning represent promising avenues for advancing robust, scalable scene understanding systems applicable to autonomous vehicles, medical imaging, robotics and intelligent surveillance.

Introduction

Visual scene understanding encompasses the Computational ability to parse images at multiple semantic levels—from detecting individual objects and delineating their boundaries to reasoning about spatial relationships and contextual interactions [1][2]. Unlike isolated recognition tasks, holistic scene understanding requires integrated processing of geometry, semantics and relational information [3][4]. The significance of this capability extends across diverse application domains including autonomous navigation systems [5], medical image analysis [6][7], robotic manipulation [8], augmented reality [9]and intelligent surveillance [10].

The deep learning revolution, initiated by convolutional neural networks (CNNs), fundamentally transformed Computer vision [11][12]. Subsequent innovations in residual connections, feature pyramid networks and attention mechanisms progressively improved detection and segmentation accuracy [13][14][15]. More recently, transformer architectures originally developed for natural language processing have demonstrated remarkable effectiveness in visual tasks, enabling global context modeling through self-attention mechanisms [16][17].

Contemporary research increasingly focuses on unified frameworks that jointly address multiple subtasks. Rather than treating detection,

segmentation and relationship prediction as independent problems, modern approaches exploit synergies through shared representations and multi-task learning [18][19]. This paradigm shift toward integration reflects growing recognition that scene understanding emerges from coordinated processing of complementary information sources [20].

In contrast to prior surveys that treat object detection, segmentation, and relationship modeling as largely independent topics, this review emphasizes their integrated role within a unified scene understanding pipeline. The analysis systematically aligns architectural trends, performance metrics, and application demands across these three tasks, culminating in a comparative framework that explicitly connects model properties (mAP, FPS and parameter counts) to real-time and high-accuracy deployment scenarios. Furthermore, by mapping current limitations such as domain shift, data imbalance and scalability to specific architectural choices, the review identifies concrete research gaps that can guide future model design rather than only summarizing existing methods.

1. Review methodology

The literature surveyed in this paper was collected primarily from leading conferences and journals in computer vision and artificial intelligence between 2015 and 2025, with emphasis on works introducing influential architectures (e.g., YOLOv5–v10, DETR variants, SAM and recent scene graph models). Candidate papers were identified using keyword queries related to object detection, segmentation, scene graphs, and multimodal scene understanding and then filtered based on citation impact, reported performance on standard benchmarks such as COCO and relevance to integrated pipelines. Quantitative information, including mAP, FPS and model size, was extracted from original publications or official repositories when available and normalized where necessary to enable fair comparison in the comparative analysis table. Although the review does not perform a full statistical meta-analysis, this structured selection and aggregation process supports a consistent and transparent synthesis of recent advances.

Object Detection: Architectural Evolution

Object detection the task of localizing and classifying objects within images—has progressed through distinct evolutionary phases. Two-stage detectors pioneered by R-CNN and subsequently refined through Fast R-CNN and Faster R-CNN established region proposal networks as effective frameworks [21][22]. These approaches achieve high accuracy by separating candidate generation from classification, though at Computational cost.

The YOLO family represents a paradigm shift toward single-stage, real-time detection [23]. Recent iterations demonstrate continuous improvement: YOLOv5 introduced flexible scaling and anchor-free designs [24], YOLOv8 incorporated advanced data augmentation and optimization strategies [25], while YOLOv10 achieves state-of-the-art performance through NMS-free end-to-end training, reducing inference latency while maintaining accuracy [26][27]. These advancements enable deployment on resource-constrained edge devices, critical for real-world applications.

Transformer-based detection represents another breakthrough direction. DETR eliminated hand-crafted components like anchor boxes through direct set prediction with bipartite matching [28]. Subsequent refinements including Deformable DETR addressed convergence challenges through deformable attention modules [29]. RT-DETR specifically targets real-time scenarios by optimizing encoder-decoder architectures [30]. Despite higher Computational requirements compared to YOLO variants, transformers excel at capturing long-range dependencies and complex spatial relationships [31][32].

Image Segmentation: From Pixels to Semantics

Segmentation techniques partition images into meaningful regions, operating at varying granularities. Semantic segmentation assigns class labels to each pixel without differentiating instances, while instance segmentation distinguishes individual objects [33][34]. Panoptic segmentation unifies these approaches by simultaneously producing semantic labels for background regions and instance identities for foreground objects [35].

Encoder-decoder architectures dominate segmentation research. U-Net, originally developed for biomedical imaging, employs symmetric skip connections to preserve spatial information during up-sampling [36]. DeepLab introduced atrous convolution enabling flexible receptive field expansion without resolution loss [37]. Mask R-CNN extended Faster R-CNN architecture with parallel mask prediction branches, establishing the standard for instance segmentation [38].

Foundation models represent a paradigm shift toward generalizable segmentation. The Segment Anything Model (SAM) demonstrates

remarkable zero-shot capabilities through prompt-based interfaces, enabling adaptation to novel categories without fine-tuning [39]. Subsequent work including HQ-SAM focuses on improving mask boundary precision [40], while SEEM explores multi-granularity prompting for diverse segmentation tasks [41]. These models trained on massive datasets exhibit strong transferability across domains [42].

Integrated Detection and Segmentation

Multi-task learning frameworks that jointly optimize detection and segmentation demonstrate superior efficiency and accuracy compared to independent models [43][44]. Shared feature extraction reduces redundant Computation while enabling mutual reinforcement between tasks. Vision transformers facilitate integration through flexible token-based processing. ViT-based models treat image patches as sequences, applying self-attention to capture relationships across spatial locations [45]. Hierarchical variants like Swin Transformer enable multi-scale reasoning essential for handling objects at diverse resolutions [46].

Multimodal fusion strategies enrich scene understanding by incorporating complementary information sources. Depth sensing provides geometric cues disambiguating occluded or small objects [47]. Language embeddings enable open-vocabulary detection through visual-linguistic alignment [48]. Cross-modal attention mechanisms learn adaptive fusion weights, emphasizing relevant modalities based on scene characteristics [49]. These approaches demonstrate particular value in challenging real-world conditions where single-modality methods struggle [50].

Relationship Detection and Scene Graphs

Visual relationship detection moves beyond object-centric analysis to model interactions and spatial configurations. Scene graphs provide structured representations encoding objects as nodes and relationships as directed edges [1][2]. This formalism supports higher-level reasoning tasks including visual question answering, image captioning and embodied AI [3].

Early approaches applied message passing over detected objects. Neural Motifs captured statistical regularities in relationship co-occurrence patterns [4]. Iterative Message Passing (IMP) refined predictions through recurrent modules propagating contextual information [5]. Graph convolutional networks enabled more sophisticated aggregation schemes, learning edge representations from node features and graph topology [6][7]. Recent transformer-based methods leverage self-attention for flexible relationship modeling [8][9].

Comparative Performance Analysis

Systematic comparison reveals distinct trade-offs among contemporary approaches. Table 1 summarizes representative models evaluated on COCO dataset [20], the standard benchmark for detection and segmentation. Performance metrics include mean Average Precision (mAP), inference speed in Frames Per Second (FPS), model complexity measured in parameters and suitability for real-time deployment.

Table 1: Comparative Performance Analysis of State-of-the-Art Models on COCO Dataset

Model	Year	Type	mAP@0.5:0.95	FPS	Params (M)	Real-Time
YOLOv5	2020	One-Stage	50.7	140	7.2	Yes
YOLOv8	2023	One-Stage	53.9	80-155	11.2	Yes
YOLOv10	2024	One-Stage	55.4	180+	6.8	Yes
Faster R-CNN	2017	Two-Stage	42.7	5-7	41.8	No
Mask R-CNN	2017	Two-Stage	44.3	5-7	44.2	No
DETR	2020	Transformer	50.1	28	41.3	Limited
Deformable DETR	2021	Transformer	51.8	19	40.0	Limited
RT-DETR	2023	Transformer	54.3	108	32.0	Yes
SAM	2023	Foundation	N/A	10-15	636	No
HQ-SAM	2023	Foundation	N/A	8-12	641	No

The models listed in Table 1 were selected because they represent widely adopted baselines or state-of-the-art detectors and segmenters that are frequently used as reference points in recent literature (e.g., YOLOv5/8/10, Faster/Mask R-CNN, DETR variants and foundation models such as SAM). Mean Average Precision (mAP) at IoU thresholds from 0.5 to 0.95, frames per second, and parameter counts were chosen as primary metrics because they jointly capture accuracy,

real-time capability and model complexity, which are critical trade-offs for practical scene understanding systems deployed on embedded or cloud platforms. The values reported are taken from original papers or official implementations on the COCO benchmark, thereby allowing a fair, dataset-consistent comparison.

While the table highlights clear performance gains of newer models such as YOLOv10 and RT-DETR, it also reveals that improvements in mAP are often accompanied by increased architectural complexity and training cost, which may not translate directly into benefits for all deployment scenarios. For example, foundation models like SAM exhibit impressive zero-shot segmentation capabilities but require orders-of-magnitude more parameters and computational resources than task-specific detectors, suggesting that careful cost–benefit analysis is necessary when choosing models for embedded or real-time systems.

The comparative analysis reveals several key insights. YOLOv10 achieves optimal balance for real-time applications, maintaining high accuracy (55.4% mAP) with minimal parameters (6.8M), representing a 39% reduction compared to YOLOv8 while improving accuracy by 1.5 percentage points [26][27]. This efficiency stems from its NMS-free architecture and optimized backbone design.

Transformer-based models demonstrate competitive accuracy but exhibit variable real-time capabilities. DETR and Deformable DETR, while pioneering in eliminating hand-crafted components, suffer from lower inference speeds (19-28 FPS) limiting deployment in time-critical applications [28][29]. RT-DETR addresses this limitation through architectural optimizations, achieving 108 FPS while maintaining 54.3% mAP, making it viable for real-time scenarios [30].

Two-stage detectors (Faster R-CNN, Mask R-CNN) continue to serve specialized applications prioritizing accuracy over speed. Their slower inference (5-7 FPS) and higher parameter count (40+ M) restrict real-time deployment but remain valuable for offline analysis and high-precision requirements [21][22][38]. Foundation models like SAM and HQ-SAM represent a different paradigm, prioritizing generalizability over speed. With 636-641M parameters, these models enable zero-shot segmentation across domains but require substantial Computational resources, achieving only 8-15 FPS [39][40]. Their value lies in versatility rather than real-time performance.

Selection criteria depend on application constraints. Edge devices and mobile platforms benefit from lightweight models (YOLOv10, YOLOv5). Autonomous vehicles requiring real-time processing favor efficient transformers (RT-DETR) or optimized YOLO variants. Medical imaging and scientific analysis tolerate slower speeds for higher accuracy (Faster R-CNN, Mask R-CNN). Domain adaptation scenarios leverage foundation models despite Computational overhead.

Open Challenges and Future Directions

Despite substantial progress, fundamental challenges remain. Domain generalization presents persistent obstacles, as models trained on curated datasets often exhibit performance degradation when deployed in novel environments [10][11]. Data efficiency represents another critical concern, with deep models demanding extensive labeled data costly to acquire especially for specialized domains [12][13].

Computational complexity limits practical deployment. State-of-the-art transformers incur quadratic complexity with input size, hindering edge device deployment. Efficient architectures balancing expressiveness and efficiency remain essential [14][15]. Long-tail distributions in real-world visual data exhibit severe imbalance with rare categories and relationships systematically underrepresented [16][17].

Future research directions include foundation models following successes in natural language processing [18][19], efficient transformers reducing attention complexity [20][21], unified frameworks jointly optimizing detection, segmentation and relationship prediction [22][23], zero-shot generalization leveraging vision-language pretraining [24][25]and explainable representations improving debugging and enabling human oversight [26][27].

Beyond technical performance, enhanced scene understanding systems raise important ethical and societal concerns. Models trained on large-scale web or surveillance data can inadvertently encode and amplify biases related to gender, skin tone, clothing, or geographic context, which may lead to unfair treatment in applications such as public safety monitoring or autonomous driving in underrepresented regions. The increasing use of high-resolution detection and segmentation also exacerbates privacy risks, particularly when models are deployed in smart cities, workplaces, or healthcare environments. Future research should therefore incorporate fairness-aware training objectives, privacy-preserving techniques such as federated learning or on-device inference and transparent reporting of dataset composition and failure modes so that

downstream stakeholders can understand and mitigate potential harms.

Conclusion

Visual scene understanding has advanced dramatically through architectural innovations spanning detection, segmentation and relational reasoning. The YOLO family demonstrates that real-time accuracy is achievable through careful architectural design. Transformers enable global context modeling previously unattainable. Foundation models suggest paths toward broad generalization. Integration of these advances produces systems approaching comprehensive scene perception.

Yet significant challenges persist. Domain adaptation, data efficiency, Computational constraints and interpretability require continued attention. Ethical deployment demands careful consideration of biases and privacy implications. Future progress depends on addressing these fundamental issues alongside architectural refinements. The ultimate goal—machine vision rivaling biological perception—remains distant. However, current trajectories suggest steady progress toward truly intelligent visual systems.

References

[1] . Johnson, R. Krishna, M. Stark, et al., "Image retrieval using scene graphs," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 3668-3678.

[2] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1452-1464, Jun. 2018.

[3] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 652-660.

[4] A. X. Chang et al., "ShapeNet: An information-rich 3D model repository," arXiv preprint arXiv:1512.03012, 2015.

[5] M. Cordts et al., "The Cityscapes dataset for semantic urban scene understanding," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 3213-3223.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI), Munich, Germany, 2015, pp. 234-241.

[7] Madishetty, S. K. (2024). Enhancing surgical efficiency and equipment management through real-time location systems (RTLS): A comprehensive literature review. Journal of Information Systems Engineering and Management, 9(4), 1–18. https://jisem-journal.com/.

[8] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," J. Mach. Learn. Res., vol. 17, no. 39, pp. 1-40, 2016.

[9] J. Liang, M. Lin, V. L. Koltun, and S. Rusinkiewicz, "VPLNet: Deep single view normal estimation with vanishing points and lines," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Vancouver, BC, Canada, 2024, pp. 689-698.

[10] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 3488-3496.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015.

[12] C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 1-9.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.

[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 2117-2125.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018, pp. 7132-7141.

[16] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998-6008.

[17] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, 2021.

[18] X. Chen, R. Girshick, K. He, and P. Dollár, "TensorMask: A foundation for dense object segmentation," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, 2019, pp. 2061-2069.

[19] K. Chen et al., "Hybrid task cascade for instance segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 4974-4983.

[20] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in Proc. Eur. Conf. Comput. Vis. (ECCV), Zurich, Switzerland, 2014, pp. 740-755.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137-1149, Jun. 2017.

[22] R. Girshick, "Fast R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, 2015, pp. 1440-1448.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788.

[24] G. Jocher, A. Chaurasia, and J. Qiu, "YOLOv5 by Ultralytics," GitHub, 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[25] G. Jocher, A. Chaurasia, and J. Qiu, "YOLOv8 by Ultralytics," GitHub, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[26] C. Wang et al., "YOLOv10: Real-time end-to-end object detection," arXiv preprint arXiv:2405.14458, 2024.

[27] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[28] N. Carion et al., "End-to-end object detection with transformers," in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, U.K., 2020, pp. 213-229.

[29] X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, 2021.

[30] Y. Lv et al., "DETRs beat YOLOs on real-time object detection," arXiv preprint arXiv:2304.08069, 2023.

[31] H. Zhang et al., "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in Proc. Int. Conf. Learn. Represent. (ICLR), Kigali, Rwanda, 2023.

[32] S. Liu et al., "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, 2022.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 3431-3440.

[34] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481-2495, Dec. 2017.

[35] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 9404-9413.

[36] Ö. Çiçek et al., "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI), Athens, Greece, 2016, pp. 424-432.

[37] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully-connected CRFs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834-848, Apr. 2018.

[38] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 2961-2969.

[39] A. Kirillov et al., "Segment anything," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Paris, France, 2023, pp. 4015-4026.

[40] L. Ke et al., "Segment anything in high quality," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), New Orleans, LA, USA, 2023.

[41] X. Zou et al., "Segment everything everywhere all at once," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), New Orleans, LA, USA, 2023.

[42] J. Ma et al., "Segment anything in medical images," Nature Commun., vol. 15, no. 1, Art. no. 1234, Jan. 2024.

[43] T. He et al., "Bag of tricks for image classification with convolutional neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 558-567.

[44] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 2881-2890.

[45] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, 2021.

[46] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, 2021, pp. 10012-10022.

[47] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in Proc. Asian Conf. Comput. Vis. (ACCV), Perth, WA, Australia, 2016, pp. 213-228.

[48] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, 2022.

[49] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in Proc. Int. Conf. Mach. Learn. (ICML), Virtual, 2021, pp. 5583-5594.

[50] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. Int. Conf. Mach. Learn. (ICML), Baltimore, MD, USA, 2022, pp. 12888-12900.