

WYA: A Novel Spatial Scene Classification Framework Based on Surrounding Object Detection

Ruoling Wu

Spatial Intelligent Research Lab, College of Information Science and Technology, Beijing University of Chemical Technology
Beijing, China

Danhua Guo*

gdh@buct.edu.cn

Spatial Intelligent Research Lab, College of Information Science and Technology, Beijing University of Chemical Technology
Beijing, China

ABSTRACT

Spatial scene classification has long been a prominent area of research in the field of geographic information science. In the past, traditional approaches heavily relied on retrieval methods based on image features. However, given the rapid advancements in deep learning and artificial intelligence, the efficient classification of complex spatial scenes has become increasingly crucial. This paper presents a novel framework named WYA (Where You At) that combines surrounding object detection with knowledge graph to automate the process of spatial scene classification. Initially, the input images undergo processing using object detection techniques to identify key entities within the scenes. Subsequently, a knowledge graph, which encompasses various spatial scenes, entities, and their relationships, is utilized to identify spatial scene categories. To validate the effectiveness of the framework, experiments were conducted using eight spatial scene categories as an example. The results demonstrated a high level of consistency with actual spatial types, thus affirming the efficacy of the framework and highlighting its potential application value in the domain of spatial scene classification.

CCS CONCEPTS

- Computing methodologies → Scene understanding

KEYWORDS

Spatial Scene Classification, Object Detection, Knowledge Graph

ACM Reference Format:

Ruoling Wu and Danhuai Guo. 2023. WYA: A Novel Spatial Scene Classification Framework Based on Surrounding Object Detection. In *Proceedings of The 8th ACM SIGSPATIAL International Workshop on Security Response using GIS 2023 (EM-GIS '23)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3615884.3629428>

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

EM-GIS '23, November 13, 2023, Hamburg, Germany

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0346-1/23/11...\$15.00
<https://doi.org/10.1145/3615884.3629428>

1 INTRODUCTION

In recent years, deep learning has demonstrated exceptional proficiency in image processing and recognition. Object detection, a focal point within computer vision, has made substantial advancements. This technology excels in efficiently analyzing intricate scenes, facilitating the detection of multiple objects. While prevailing object detection tasks, such as pedestrian detection, traffic signal recognition, and remote sensing target identification [29], predominantly focus on foreground objects, the background elements within images often harbor vital information essential for a comprehensive scene understanding. Consequently, the task of background element detection in images, coupled with the determination of their spatial context, presents a significant and multifaceted challenge.

Background element detection in images plays a pivotal role in assisting analysts in identifying salient objects, their positions, and distinctive features within the background context. This process contributes to the construction of a more precise representation of the spatial scene. For instance, within the domain of criminal investigations, background element detection aids law enforcement agencies in pinpointing the whereabouts and movements of suspects. By discerning critical background elements such as buildings, roadways, and landmarks, law enforcement can more effectively ascertain the suspect's location. This technology not only facilitates a more profound comprehension of the crime scene but also furnishes vital evidentiary support, streamlining the process of criminal investigations and judicial proceedings. Consequently, it holds the potential to enhance the efficiency and accuracy of criminal investigations, thereby making a substantial contribution to social security and the administration of justice.

In light of the aforementioned research context and the associated imperatives, this paper introduces a framework to identify and categorize the classes of spatial scenes. Leveraging object detection models based on deep learning, this framework enables the recognition of background elements within images. Furthermore, it encompasses the creation of a knowledge graph specifically tailored to spatial scenes, mapping spatial scenes to their corresponding key objects. Through the process of entity recognition and search within the knowledge graph, the framework ultimately ascertains the spatial scene category of the image.

2 RELATED WORKS

2.1 Spatial Scene Recognition and Classification

The recognition and classification of spatial scenes can be categorized into feature-based image retrieval methods and image retrieval methods based on deep learning[25].

Traditional feature-based retrieval methods utilize a Bag-of-Visual-Words (BoVW) approach. Initially, they employ local feature descriptors such as SIFT[16] to represent local features. Subsequently, the distance between image features and each visual word in a visual dictionary is computed, generating a histogram of visual words. This process facilitates image retrieval and scene recognition.

While traditional feature-based methods adequately serve scene recognition needs, the emergence of deep learning has brought about groundbreaking improvements in scene classification and recognition, revolutionizing the entire field of computer vision. Convolutional Neural Networks (CNN) like AlexNet[10], VGG[22], and ResNet[8] have played pivotal roles in image classification. The availability of large-scale image datasets such as ImageNet[4], COCO[14], and Places[28] has further enriched this domain. Scene recognition is predominantly concerned with the geographical characteristics of spatial scenes, and the application of CNN for feature extraction has significantly advanced this research. CNN autonomously learn and capture features at different levels, encompassing edges, textures, shapes, and high-level semantic features, which are indispensable for spatial scene recognition. They can handle input images of various scales and adaptively capture both large-scale global features and small-scale local features, thereby enhancing the understanding of complex scenes. Furthermore, the utilization of data augmentation expands the training dataset, improving the model's generalization capacity. This aspect proves crucial for handling scenes under varying temporal, weather, and lighting conditions.

2.2 Object Detection

Due to the complexity of spatial scenes, encompassing both foreground and background elements, achieving scene classification directly presents certain challenges. Therefore, it is more effective to initially employ object detection methods to extract entities of interest within the scene images. Subsequently, by classifying these extracted entities, we can achieve improved scene classification. The evolution of object detection can be divided into two main periods: the era of traditional object detection algorithms and the algorithms based on deep learning[26].

Traditional object detection methods generally involve three key steps: candidate target region selection, feature extraction, and classification[27]. Representative traditional object detection algorithms include the Viola-Jones (VJ) detector[24] proposed by P. Viola and M. Jones in 2001, Histogram of Oriented Gradients (HOG) detector[3] introduced by N. Dalal and B. Triggs in 2005, and the Deformable Part Model (DPM)[5] presented by P. Felzenszwalb in 2010.

The advent of CNN significantly enriched feature extraction, surpassing manual feature engineering. This advancement led to

the rapid development of object detection based on deep learning. Depending on the design structure of detectors, it can be categorized into two main types: "Two-stage" and "One-stage". In 2014, R. Girshick and colleagues introduced R-CNN, pioneering the "Two-stage" object detection approach. This approach operates in a "candidate region + prediction" mode and offers excellent detection accuracy. Subsequently, researchers proposed Fast R-CNN[6], Faster R-CNN[21], Mask R-CNN[7], and Feature Pyramid Network (FPN)[12], among others. In response to the efficiency issues of "Two-stage" object detection, in 2016, R. Joseph and colleagues introduced the YOLO (You Only Look Once) model[18], which marked the inception of One-stage object detectors. The YOLO series models employ a single convolutional neural network to input the entire image. They divide the image into multiple regions and predict bounding boxes and class confidences for each region. Although slightly less accurate compared to "Two-stage" models, they excel in speed and lightweight design, offering promising practical applications. The YOLO series models have continuously evolved and improved[1, 19, 20]. In January 2023, YOLOv8, the latest iteration, was introduced, optimized for its backbone network modules and loss functions, demonstrating excellent performance. Additionally, in recent years, several other One-stage models have emerged, including the Single Shot MultiBox Detector (SSD)[15], RetinaNet[13], and EfficientDet[23], among others. These models have demonstrated impressive performance in object detection tasks, providing significant impetus to the advancement of this field.

2.3 Knowledge Graph

Knowledge Graph (KG) is a pivotal concept in the fields of artificial intelligence and natural language processing, which is a method of organizing and representing knowledge in the form of a graph. The concept of knowledge graphs can be traced back to the 1960s with "Semantic Web", which used interconnected nodes and edges to represent knowledge. Nodes signify objects or concepts while edges denote relationships between nodes, forming a precursor to today's knowledge graphs. In 2012, Google introduced the concept of the "Knowledge Graph", initially applied to enhance search engine results. Its objective is to provide structured information to users, allowing them to address their queries without navigating to other websites or manually aggregating information. In the contemporary information age, knowledge graphs are not only utilized for improving search engines but also find widespread application in natural language processing, recommendation systems, intelligent assistants, healthcare, AiFinance, public safety and government services, and various other domains. They aid machines in understanding context, conducting inference, and providing intelligent decision support, thus propelling advancements in the field of artificial intelligence. With the proliferation of open data and the development of open-source tools, creating knowledge graphs has become more accessible. Large-scale knowledge graph projects such as Wikidata and DBpedia have made substantial structured data openly available for researchers and developers.

A knowledge graph is a graph data structure used for organizing and representing knowledge, comprising entities and their relationships. Entities are the core elements in a knowledge graph and

typically represent real-world entities, concepts, individuals, or objects. They can encompass concrete entities like people, places, and books, as well as abstract concepts such as emotions, theories, and events. Attributes are related properties or features associated with entities, aiding in providing detailed descriptions and additional information about the entities. Relationships denote connections or associations between entities, describing the relational aspects between different entities. Relationships usually have names and directions. For example, "contain" is a relationship that can connect a bookstore entity and a book entity. Relationships can be unidirectional or bidirectional and may possess attributes to describe properties such as weight or strength.

The structured nature of knowledge graphs makes them an effective tool for managing vast knowledge repositories, fostering developments in semantic search, natural language processing, intelligent recommendations, and decision support systems. In the task of this article, a knowledge graph can store relationships between spatial scenes and entities within those scenes, enabling matching and associations. This allows for the determination of spatial scene types based on detected entities, carrying practical significance.

3 METHOD

To achieve spatial scene classification, we propose a comprehensive framework based on object detection and knowledge graphs. This framework consists of two main components: (1) Spatial Entity Recognition based on Object Detection and (2) Spatial Scene Classification based on Knowledge Graphs. The framework diagram is shown below.

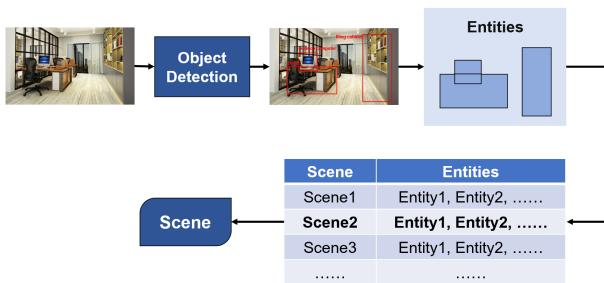


Figure 1: A Spatial Scene Classification Framework Based on Target Detection and Knowledge Graphs

3.1 Spatial Entity Detection

In the spatial scene classification framework, the primary step involves the utilization of object detection techniques to process input images or videos. You Only Look Once(YOLO) model, known for its lightweight nature and ease of training, is widely adopted today. For this specific task, the YOLO model proves to be highly competent, and hence, it is selected. This stage aims to recognize and locate various objects or entities appearing within the scene in the input images, determining both their positions and categories.

3.2 Spatial Scene Classification

(1) Knowledge Graph Construction

Currently, there exist numerous spatial scene datasets, with some datasets predominantly focusing on indoor scenes, such as MIT67[17] and InteriorNet[11], while outdoor scene datasets are more oriented towards the field of Autonomous Driving, exemplified by Cityscapes[2]. Moreover, several datasets encompass both indoor and outdoor scenes, including SUN397[9] and Places[28], among others. These public datasets effectively cater to the requirements of tasks such as spatial scene classification and various other applications. Guided by the aforementioned public datasets, in this phase, a knowledge graph of spatial scenes and the spatial entities within them has been constructed. Given the vast diversity of spatial scenes and entities, only a set of common spatial scenes have been enumerated, and their most crucial elements have been selected for correspondence. The specific relationships are outlined in **Table 1**.

(2) Knowledge Graph Querying

Utilizing the results obtained from object detection in the preceding phase as query parameters, it is possible to interact with a carefully constructed knowledge graph. This knowledge graph encompasses a multitude of spatial scenes, entities, and the intricate network of relationships among them. By systematically comparing the entities discerned during the object detection phase with those residing within the knowledge graph, the specific spatial scene category associated with the given image can be inferred. To illustrate, if the detected entities encompass elements such as desks, filing cabinets, and a printer, and the knowledge graph delineates interconnections among these entities, it's certain that the current scene corresponds to an office.

4 EXPERIMENT

4.1 Dataset

The size of the dataset and the clarity of its images significantly influence a model's fitting capacity and prediction generalization. Therefore, the selection of the dataset is of paramount importance. The ImageNet dataset is one of the most commonly utilized datasets for image classification, detection, and localization. Its diversity and extensiveness make it a valuable resource for research in image understanding and scene recognition, aligning well with the tasks of the proposed framework. This dataset comprises a large number of images with annotated class and location information. Therefore, during the object detection training phase, entities relevant to the task were selected from the ImageNet dataset for experiment. The chosen dataset consisted of images from 37 different categories, with a relatively balanced distribution of images per category, totaling 22,019 images and encompassing 25,630 target entities.

4.2 Settings

The implementation of YOLOv5 model is on Ubuntu 18.04, leveraging GPU acceleration, specifically utilizing the GeForce RTX 3080 graphics card equipped with GA102 GPU. The model training parameters were configured as follows: a batch size of 16, a learning rate of 0.01, and a total of 300 epochs for iterations.

Table 1: Correspondence between Spatial Scenes and Entities

Scenes	Entities
office	desk, desktop computer, notebook computer, filing cabinet, folding chair, printer
kitchen	frying pan, pitcher, plate rack, refrigerator
restaurant	dining table, menu, plate, hot dog, cabbage
city	check, traffic sign, traffic light, bookstore, shoe store
coast	seashore, shoal, yawl, swimsuit
mountain	alp, volcano, cliff, valley
forest	hay, cardoon
countryside	barn, boathouse, thatched roof, ox, hen, duck, goose

The construction of the spatial scene knowledge graph is accomplished using the high-performance NoSQL graph database Neo4j, allowing for visual representation. Connectivity to the Neo4j dataset is established and queries are executed on the database using Python.

4.3 Procedure

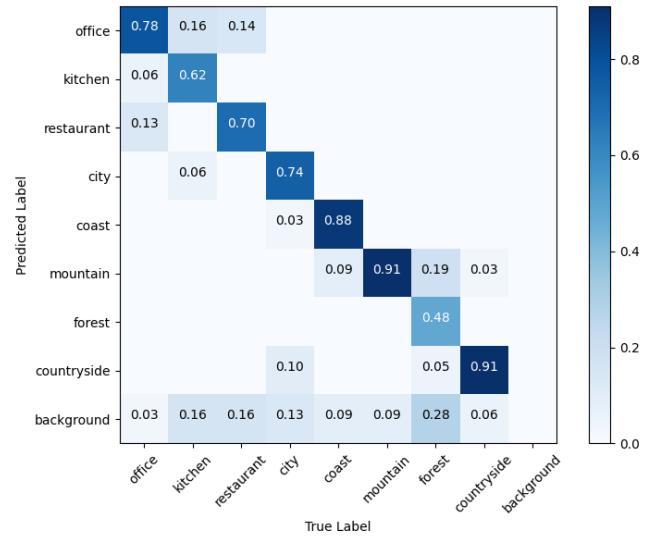
To validate the effectiveness of our proposed framework, experiments are conducted on eight representative spatial scenes. These scenes are characterized by distinct target entities and encompass three indoor scenes: office, kitchen, and restaurant, as well as five outdoor scenes: city, forest, coast, countryside, and mountain. A total of 240 test images were collected from the internet, spanning across these various scenes, with an average of 30 images per category. Subsequently, the YOLOv5 model is employed for object detection, resulting in the generation of position coordinates and category labels for the detected objects in each image.

Following the object detection phase, knowledge graph queries are performed based on the detection results to establish connections between the detected entities and those in the knowledge graph. The detected entities are corresponded to the knowledge graph sequentially, and the queried spatial scenes are counted. If the highest number exists, it is judged to be the scene. In cases where multiple scene categories have the same highest count, the confidence scores of the detected target entities are considered jointly to determine the predicted scene. As a result, based on the outcomes of knowledge graph queries, the prediction scene category for each image can be determined.

5 RESULTS

The confusion matrix, as a matrix-based tool, serves as a valuable instrument for assessing the performance of classification models. It provides a more comprehensive evaluation of performance, reducing the potential for misguidance that may arise from relying only on accuracy or precision. In this experiment, the classification results are presented in the form of a confusion matrix, as depicted in the figure below.

Upon examination of the experimental results, it can be observed that for the selected eight scenes, the majority of the images are correctly categorized into their respective classes, demonstrating the effectiveness and generality of the framework across different scenes. However, a minority of images are misclassified due to the failure to detect entities during the object detection phase,

**Figure 2: Spatial Scene Classification Results**

consequently impacting the scene classification. Furthermore, there is a degree of similarity among certain scenes. For instance, both offices and restaurants may have tables, valleys could be present in both forests and mountains, and so on. These similarities contribute to minor errors in both detection and classification tasks. In an overarching assessment, the framework proposed in this study exhibited commendable performance in the realm of spatial scene classification. The specific implementation results are depicted in the **Table 2**.

6 CONCLUSION

This paper introduces a spatial scene classification framework based on object detection and knowledge graphs, aiming to achieve understanding and classification of scenes within images. Through testing across eight scenes, the effectiveness of the framework has been verified. By amalgamating object detection techniques from the computer vision domain with semantic representations from knowledge graphs, this framework have successfully bridged the gap between visual perception and structured knowledge, bringing new possibilities for spatial scene classification tasks.

Table 2: Spatial Scene Classification Framework Classification Effects

Scene Images	Detected Entities	Knowledge Graph Queries	Classification Results
		<pre> graph TD office[office] -- have --> filingCabinet[filing cabinet] office[office] -- have --> desk[desk] filingCabinet[filing cabinet] -- have --> foldingChair[folding chair] </pre>	office
		<pre> graph TD office[office] -- have --> desk[desk] office[office] -- have --> kitchen[kitchen] kitchen[kitchen] -- have --> refrigerator[refrigerator] </pre>	kitchen (after comparing confidence)
		<pre> graph TD restaurant[restaurant] -- have --> diningTable[dining table] </pre>	restaurant
		<pre> graph TD city[city] -- have --> check[check] check[check] -- have --> trafficLight[traffic light] </pre>	city
		<pre> graph TD coast[coast] -- have --> yawl[yawl] coast[coast] -- have --> seashore[seashore] </pre>	coast
		<pre> graph TD cliff[cliff] -- have --> mountain[mountain] mountain[mountain] -- have --> alp[alp] </pre>	mountain
		<pre> graph TD forest[forest] -- have --> hay[hay] </pre>	forest
		<pre> graph TD countryside[countryside] -- have --> thatchedRoof[thatched roof] countryside[countryside] -- have --> boathouse[boathouse] </pre>	countryside

On the practical front, this framework holds immense potential, particularly in the realms of criminal justice and public safety. Its application for identifying crime scenes has promising prospects. It can assist in case investigations, facilitate the analysis of crime scenes, and aid in evidence collection, ultimately contributing to the reconstruction of crime scenes. This provides robust technological support for law enforcement and enhances the efficiency of the judicial system, promoting public safety and the overall societal sense of security.

Nevertheless, this study has limitations, with room for potential enhancement in two aspects: (1) Increasing the variety of scenes and scaling up the knowledge graph will definitely allow the framework to be more versatile and cover more practical scenes. (2) Integrating more advanced decision-making methods into the process from knowledge graph queries to final classification outcomes can enhance the framework's interpretability.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of China [NSFC41971366, 42371476] and Fundamental Research Funds for the Central Universities of China [buctrc202132].

REFERENCES

- [1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *ArXiv abs/2004.10934* (2020). <https://api.semanticscholar.org/CorpusID:216080778>
- [2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3213–3223. <https://doi.org/10.1109/CVPR.2016.350>
- [3] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 886–893 vol. 1. <https://doi.org/10.1109/CVPR.2005.177>
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [5] Pedro F. Felzenszwalb, Ross B. Girshick, and David McAllester. 2010. Cascade object detection with deformable part models. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2241–2248. <https://doi.org/10.1109/CVPR.2010.5539906>
- [6] Ross Girshick. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1440–1448. <https://doi.org/10.1109/ICCV.2015.169>
- [7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2980–2988. <https://doi.org/10.1109/ICCV.2017.322>
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.
- [9] Luis Herranz, Shuqiang Jiang, and Xiangyang Li. 2016. Scene Recognition with CNNs: Objects, Scales and Dataset Bias. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 571–579. <https://api.semanticscholar.org/CorpusID:15429030>
- [10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60 (2012), 84–90.
- [11] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. 2018. InteriorNet: Mega-scale Multi-sensor Photo-realistic Indoor Scenes Dataset. In *British Machine Vision Conference (BMVC)*.
- [12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 936–944. <https://doi.org/10.1109/CVPR.2017.106>
- [13] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP, 99 (2017), 2999–3007.
- [14] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.
- [15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). 21–37.
- [16] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110.
- [17] Ariadna Quattoni and Antonio Torralba. 2009. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 413–420. <https://doi.org/10.1109/CVPR.2009.5206537>
- [18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 779–788. <https://doi.org/10.1109/CVPR.2016.91>
- [19] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6517–6525. <https://doi.org/10.1109/CVPR.2017.690>
- [20] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *ArXiv abs/1804.02767* (2018). <https://api.semanticscholar.org/CorpusID:4714433>
- [21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
- [22] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.
- [23] Mingxing Tan, Ruoming Pang, and Quoc V. Le. 2020. EfficientDet: Scalable and Efficient Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10778–10787. <https://doi.org/10.1109/CVPR42600.2020.01079>
- [24] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Vol. 1. I–I. <https://doi.org/10.1109/CVPR.2001.990517>
- [25] Yajun Wang, Qizi Mu, Moyin Yu, Honghai Wang, and Liyuan Zhu. 2023. Outdoor Scene Recognition Based on Convolutional Neural Network. *Automation Application* 64, 201–207 (2023).
- [26] Degang Xu, Lu Wand, and Fan Li. 2021. Review of Typical Object Detection Algorithms for Deep Learning. *Computer Engineering and Applications* 10–25 (2021).
- [27] Shun Zhang, Yihong Gong, and Jinjun Wang. 2019. The Development of Deep Convolution Neural Network and Its Applications on Computer Vision. *Chinese Journal of Computers* 42, 453–482 (2019).
- [28] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1452–1464. <https://doi.org/10.1109/TPAMI.2017.2723009>
- [29] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2023. Object Detection in 20 Years: A Survey. *Proc. IEEE* 111, 3 (2023), 257–276.