Specialized Indoor and Outdoor Scene-specific Object Detection Models

Mahtab Jamali[a], Paul Davidsson[a], Reza Khoshkangini[a], Martin Georg Ljungqvist[b], and Radu-Casian Mihailescu[a]

[a]Internet of Things and People Research Center, Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden
[b]Axis Communications AB, Lund, Sweden

ABSTRACT

Object detection is a critical task in computer vision with applications across various domains, ranging from autonomous driving to surveillance systems. Despite extensive research on improving the performance of object detection systems, identifying all objects in different places remains a challenge. The traditional object detection approaches focus primarily on extracting and analyzing visual features without considering the contextual information about the places of objects. However, entities in many real-world scenarios closely relate to their surrounding environment, providing crucial contextual cues for accurate detection. This study investigates the importance and impact of places of images (indoor and outdoor) on object detection accuracy. To this purpose, we propose an approach that first categorizes images into two distinct categories: indoor and outdoor. We then train and evaluate three object detection models (indoor, outdoor, and general models) based on YOLOv5 and 19 classes of the PASCAL VOC dataset that consider places. The experimental evaluations show that the specialized indoor and outdoor models have higher mAP (mean Average Precision) to detect objects in specific environments compared to the general model that detects objects found both indoors and outdoors. Indeed, the network can detect objects more accurately in similar places with common characteristics due to semantic relationships between objects and their surroundings, and the network's misdetection is diminished. All the results were analyzed statistically with t-tests.

Keywords: object detection, YOLOv5, indoor object detection, outdoor object detection, scene classification.

1. INTRODUCTION

Object detection is a vital process of numerous artificial intelligence applications, such as autonomous driving, robotics, image retrieval, healthcare, and intelligent video surveillance,[1,2] to name a few. Over the past decades, Convolutional Neural Networks (CNN)[3] have gone a long way and CNN-based models built for object detection brought revolutionary changes in such applications. The models can generally be divided into three distinct categories; The first category is *Two-stage* object detection models such as R-CNN,[4] Fast R-CNN,[5] Faster R-CNN,[6] RFCN,[7] and Mask R-CNN.[8] These models generate the region of interest and then classify and predict objects using a regression model. Although two-stage models are typically more accurate than single-stage models, they are slow for real-time applications due to having two stages and complex calculations. The second type is *One-stage* models such as YOLO,[9] SSD,[10] RetinaNet,[11] EfficientDet,[12] CornerNet,[13] CenterNet[14] that perform coordinate regression (predicting object locations) and classification (identifying object classes) simultaneously. End-to-end processing makes these models quicker than two-stage models and more applicable to real-time scenarios. The third type of object detection models employs *Transformers*,[15] such as DETR,[16] which directly calculates results using encoder-decoder structures. While deep learning and CNNs have significantly advanced object detection practices, there are still challenges that need to be solved in this field. Occlusion, light environment, poor-quality images, scale variation, complex backgrounds, and cluttered scenes are some of these challenges.[17,18]

Objects in the real world do not exist in isolation, and their associations and places enrich their meanings. Using contextual information around objects or scene-level context is one of the methods that has been used in many studies[19–22] for tackling the above challenges to improve the performance of the models. For example, Liu

et al. in[19] proposed a method that extracts the whole image feature as a scene and concatenates it with features of images and descriptors. In inside-outside net,[20] contextual information around objects and the region of interest are employed to improve the network's performance to detect objects. Contextual-YOLOv3,[23] TYOLOv5,[24] Context R-CNN,[25] and Fusion Faster R-CNN[26] are other models that leverage contextual information at the object or scene level. While these models have achieved satisfactory developments, they have yet to focus on scene-specific object detection, meaning that they do not detect objects in a manner tailored to specific places. Indeed, they are trained as general models applicable to all different places, and there is still a need to investigate the performance of specialized models for specific places.

Scene awareness has been studied using scene classification for choosing a change detection method suitable for the scene type.[27] However, they did not perform object detection. Scene-specific object detection has been done where a CNN was adapted to detect objects with context,[28] as well as in an unsupervised manner[29] without using context or CNNs. Other works[30–32] have also investigated scene-specific object detection, but their networks use overall scene contextual information to detect objects without being specialized, which means that these networks are not trained separately for specific places. Studies have also been performed on scene specific pedestrian detection.[28,33]

Motivated by the above, we propose specialized scene-specific object detection models to detect objects based on their places. Since objects are situated in various places, for instance, birds are often in the sky (outdoors), and keyboards are often near computers (indoors), specialized scene-specific models might perform well in object detection. The extraction of contextual information by a network that simultaneously detects objects in different places may reduce the network's efficacy. A network that processes diverse information from objects in various types of places may not be as precise as a model that focuses on objects in a specific place or places with similar characteristics. Thus in this study, two main phases construct the proposed approach: The first phase classifies images based on their places into indoor and outdoor categories, and the second phase applies specialized object detection to each category separately.

This work is the first study to investigate scene-specific object detection with CNNs specialized on indoor and outdoor places and classes, and it precisely compares the accuracy of the specialized models' classes with the general model classes. Earlier studies concentrated more on general models by considering extracted information from all data, or models that leverage scene contextual information without training specialized networks for specific places; however, building and applying specific models could lead to higher accuracy and less effort.

In this paper, we first classified the images into two categories, indoor and outdoor, and then two specialized models for the indoor and outdoor groups and one general model for different types of places were trained. We also demonstrate how this specialized approach enhanced the prediction performance compared with the general model.

The remainder of the paper is organized as follows; The proposed approach is described in Section 2. Section 3 covers the datasets used in this study, experimental evaluation, and results. Section 4 gives a discussion and summary of the work.

2. METHOD

2.1 Scene-specific object detection

Figure 1 shows the high-level structure of the proposed approach. It contains two main phases; In the first phase, images are divided into indoor and outdoor groups based on their places. To do this, wideresnet18[34] is used as the classification network's backbone. Since in object detection datasets, images are often not annotated with places, the pre-trained weights of the places365 dataset[35] are used to classify images into indoor and outdoor categories.

Some objects are more visible than others in specific places. For example, it is highly uncommon to encounter an elephant or a truck in an indoor place. Since this study aims to design specialized models for specific places, we created an agent that precludes non-commonly observed objects from entering the network training procedure. Accordingly, after classifying the images into indoor and outdoor categories, based on the number of instances of each class in each indoor or outdoor category, the classes of the dataset are also divided into two main categories.

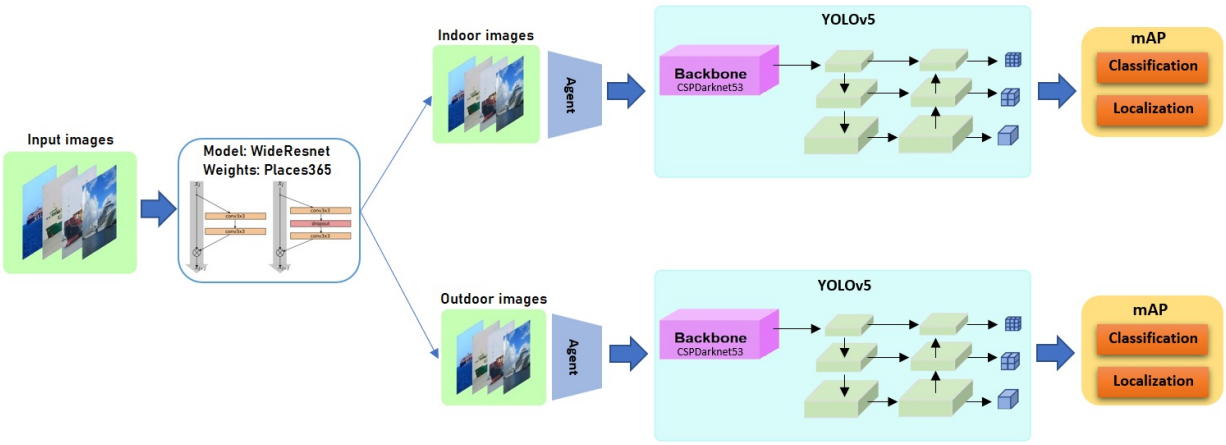

Figure 1. The conceptual view of the proposed approach.

As a result, objects that are rarely seen in indoor or outdoor places are removed from that group and are not included in the training process. Figure 2. depicts the distinct number of classes in each group.

The second phase is training object detection models. We chose the YOLOv5 object detection model as the foundation of our approach due to its accuracy and speed of inference in real-time. YOLOv5 has several configurations with different sizes, parameters, speeds, and mAPs. In this research, we used YOLOv5l, which has approximately 46.5 million parameters. YOLOv5l is slower than smaller configurations, such as YOLOv5n or YOLOv5s, but it has a higher mAP. Some essential advantages of YOLOv5 compared to previous versions are smaller volumes, higher speed, higher precision, and implementation in the PyTorch open-source ML framework.[36]

2.2 Overview of YOLOv5 and WideResNet

The YOLOv5[37] has been designed to provide real-time and accurate object detection in one-stage operation. It belongs to the YOLO series of one-stage object detectors[9,38–42] . The last version of this family is YOLOv8, but an official paper has yet to be published for it. We chose the YOLOv5 model for this work due to its outstanding performance, which makes it an ideal candidate for achieving our research objectives.

YOLOv5 has three main parts: backbone, neck, and head. The backbone of this model is CSP Darknet53,[43] which stands for Cross Stage Partial Darknet53 and is a modified version of Darknet53.[39] In CSPDarknet53, cross-stage partial connections are introduced in addition to Darknet53. These connections allow features to be passed forward and backward between different stages in the network. Apart from CSP, this backbone has spatial pyramid pooling (SPP)[44] to extract feature maps of various sizes from the input images. YOLOv5 uses PANet[45] as a neck in its architecture. Several upsampling and downsampling modules are included in PANet to enable detection at multiple scales and resolutions. A custom head within this model also predicts bounding boxes and class probabilities of objects in the scenes. YOLOv5 has different architectural scales with different parameters, including YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x.

WideResNet is a development of the original ResNet architecture. A key concept behind WideResNet is to increase the number of channels in the convolutional layers to learn complex models. In this architecture, "L" is the number of convolutional layers in each residual block, and "K" is used to control the width of the network. Increasing the width factor makes the network wider and leads to more channels. WideResNet has several advantages compared to the original ResNet, such as reducing overfitting, and flexibility in network design.

3. EXPERIMENTS

3.1 Dataset

The PASCAL VOC 2012 dataset,[46] is a common benchmark for object detection in computer vision. This dataset contains 20 object classes, such as car, bottle, cat, bird, and household items in different positions and

light environments from various real-world scenarios. We used it to train the three models presented in this paper. The images and classes of this dataset were divided into two categories: indoor, and outdoor, which are further explained in the 3.2 section. In addition, the pre-trained weights of the places365 dataset[35] were used to classify the images of the PASCAL dataset based on their places into indoor and outdoor categories. This dataset contains approximately 1.8 million images classified into 365 scene classes, representing various indoor and outdoor places.

3.2 Experiment setup

Using the WideResNet network and pre-trained weights of the Places365 dataset, the images of the PASCAL dataset were initially classified as indoors or outdoors. Following the specialization of the models, and due to the fact that certain objects are observed more frequently than others in any specific place, the designed agent was employed to eliminate specific classes. The number of instances of each class was checked in two indoor and outdoor categories. The classes with a small number of instances in each indoor or outdoor category were removed and placed solely in the group with the greatest number of instances. 8 classes of this dataset were allocated to the indoor model, and 11 classes were assigned to the outdoor model. Person class was excluded from this experiment because its instances were almost equal in the indoor and outdoor categories. This class was removed so that the networks could be trained in a specialized manner without common classes, leading to a fair comparison of the three models. Therefore, 19 classes of the PASCAL dataset were used in this paper. The general model was trained on all images and classes of the indoor and outdoor categories without categorizing or considering scene-specific characteristics. Figure 2 depicts the classes and their instances of the three models analyzed in this paper (the indoor, outdoor, and general models).

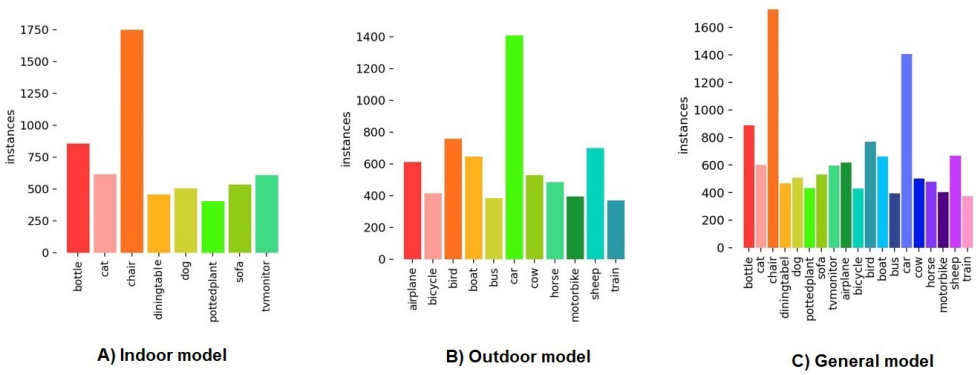

Figure 2. Object statistics in the indoor, outdoor, and general models.

In the following step, all three models were trained to determine the impact of indoor and outdoor places on object detection mAPs. Each model was trained five times with 300 epochs using the YOLOv5l model. Image size and batch size of the networks are 416×416 and 16, respectively. Finally, the average values were calculated for each model.

3.3 Results

The results of training the indoor, outdoor, and general models are shown in Table 1. The outdoor model has higher recall, precision, and mAP than the indoor and general models. Higher values of the outdoor model reflect the positive effect of incorporating information of similar places into the training process. Even though the outdoor model has fewer training images than the general model, the mAP is higher. The results of the outdoor model indicate that objects can be detected with greater precision when they exist in specific places with similar characteristics. It could be due to similar spatial relationships and arrangements.

The overall mAP of the indoor model is lower than the outdoor and general models. One of the factors for the lower mAP can be attributed to the smaller sizes and fewer pixels of indoor objects in the PASCAL dataset. The term "small objects" refers to objects that fill areas less than and equal to 32×32 pixels.[47] Due to limited

Table 1. Comparison of Recall, Precision, and mAP between the three models.

Model	Classes	Recall	Precision	mAP
Indoor model	8 indoor classes	61.9	79.0	50.5
Outdoor model	11 outdoor classes	**70.5**	**89.2**	**61.4**
General model	19 indoor+outdoor classes	68.0	83.9	57.6

Spatial Information, it is typically more difficult for a network to detect small objects. Small objects occupy fewer pixels within an image, limiting the network's ability to derive meaningful features from the image.[48] Another reason can be the diverse orientations and heavy occlusion of objects in indoor places.[49] As a result, it is more difficult for a network to detect indoor objects.

In order to gain a deeper understanding of the performance of the models, we compared the mAPs of all classes across three models, as shown in Table 2. Based on the results, 5 out of 8 classes in the indoor model and 8 out of 11 classes in the outdoor model have higher mAPs than mAPs of the same classes in the general model. In the general model, only 6 out of 19 classes have better performance than the specialized indoor and outdoor models. It is noteworthy that when the specialized model outperforms the general model in certain classes, the differences between the mAPs of the same classes in both models are substantially larger than when the situation is reversed. Although the overall mAP of the general model is higher than the indoor model (57.6 vs. 50.5), it was determined that the indoor model has performed better for most classes than the same classes in the general model. The superior performance of the specialized models in the majority of the classes demonstrates the positive effect of training objects in similar places.

Table 2. Comparison of mAP for all 19 classes between the three models.

Class	Indoor model	Outdoor model	General model
Bottle	**44.6**		35.3
Cat	**76.8**		69.7
Chair	**44.0**		41.0
Diningtable	52.5		**52.8**
Dog	64.9		**66.8**
Pottedplant	**27.1**		22.8
Sofa	**57.0**		52.7
TV-Monitor	56.9		**59.8**
Airplane		**68.4**	63.6
Bicycle		**63.1**	62.1
Bird		**50.4**	45.5
Boat		**41.4**	39.1
Bus		75.8	**80.9**
Car		**62.2**	60.9
Cow		**57.6**	54.7
Horse		**68.5**	64.6
Motorbike		64.8	**65.5**
Sheep		48.4	**50.9**
Train		**77.3**	64.7

3.4 Statistical analysis of the results

To evaluate the results of the models statistically, we compared them using charts and t-tests. In Figure 3, mean mAPs of 8 classes of the indoor model and mean mAPs of the same classes from the general model were compared. On the other hand, mean mAPs of 11 classes of the outdoor model were compared to their corresponding classes in the general model. The charts show that the specialized models have higher means than the general model, indicating their better performance in detecting objects seen in specific places.

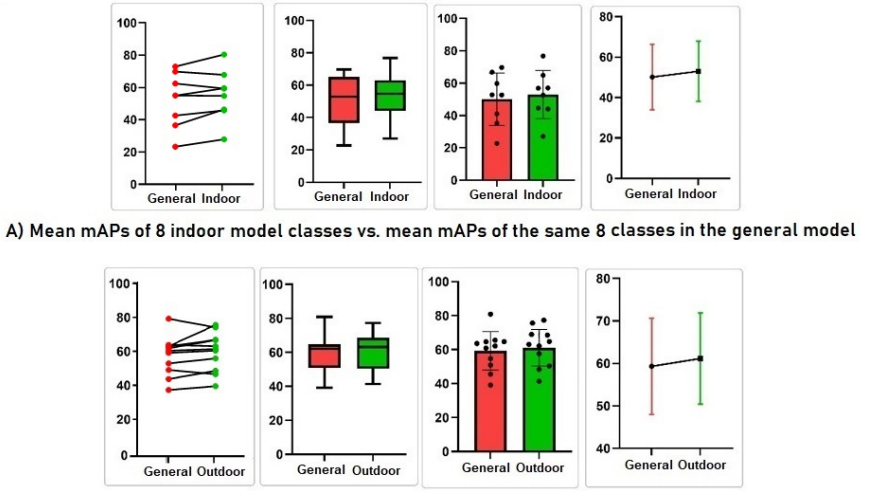

Figure 3. timeline of object detectors

The t-test is used in statistical analysis to compare the means of two groups to determine whether they have a statistically significant difference. It quantifies the extent of the difference within each group based on the t-value calculated from the t-distribution. It considers both groups' sample size, means, and standard deviations. When the t-value exceeds the critical value and the p-value is below a predetermined significance level, there is a statistically significant difference between the groups, and equality of the means is rejected.

For comparing the overall mAPs of the models, we used independent t-tests, whereas paired t-test was used for comparing the mAPs of the classes in three models. In this comparison, the confidence interval is 95%, and the significance level is 0.10.

Table 3. The independent t-test of the overall means of the models

Group	T-value	P-values
General and Outdoor	-6.02729	0.000314
General and Indoor	4.83919	0.001289

Based on Table 3, since the p-values are lower than the significance level, the differences between the means of the two groups are big enough to be statistically significant. As a result, the outdoor model has statistically outperformed the general model. The general model has superior results than the indoor model, but when we analyzed the results of each class separately, we discovered that the indoor model performed better than the general model in the majority of classes involving objects found in indoor places. To check this improvement statistically, we examined the means of all classes in the three models using the paired t-test. In this test, the classes of specialized models were evaluated with their corresponding classes in the general model. The results in Table 4 and Figure 4 indicate that both specialized models perform better than the general model in detecting objects in specific places.

Table 4. The paired t-test of 8 and 11 classes in the indoor and outdoor models, respectively, with their corresponding classes in the general model.

Group	T-value	P-values
General and Outdoor	1.828713	0.09738
General and Indoor	1.899446	0.09929

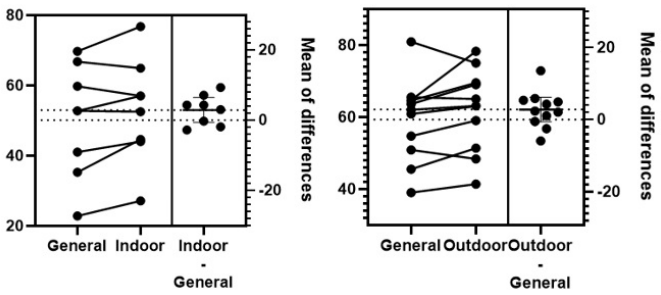

Figure 4. Comparing the mAPs of classes in the indoor and outdoor models with their corresponding classes in the general model (general vs. indoor and general vs. outdoor)

We tested the specialized models on test images. Some of the results are shown in Figure 5. The specialized outdoor model has detected the airplane and boats with higher mAPs. On the other hand, the general model has a higher detection rate for the incorrect class (detection of humans as airplanes with high mAPs).

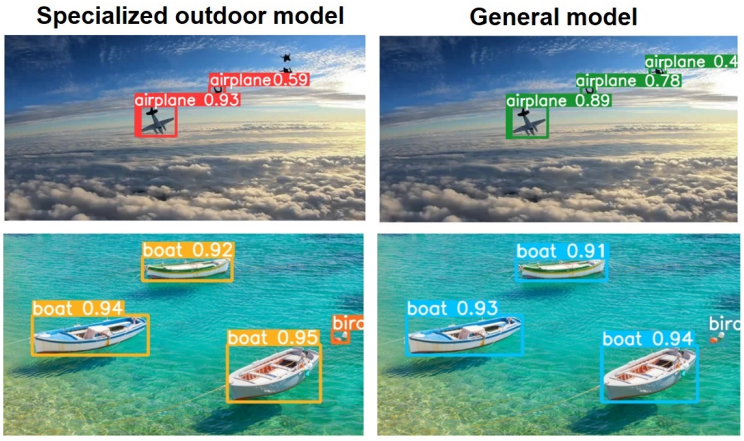

Figure 5. Testing the specialized outdoor model and the general model on test images

4. CONCLUSIONS AND FUTURE WORKS

This work studied the task of detecting objects based on their places. For this purpose, three models were investigated (the indoor model, the outdoor model, and the general model). In specialized indoor and outdoor models, the WideResnet network and the pre-trained weights of the places365 dataset were used to classify the images into indoor and outdoor categories. The YOLOv5l model and the PASCAL dataset were utilized for all three networks. The experimental results indicate that the outdoor model has a higher overall mAP than the general model. Although the overall mAP of the indoor model is lower than the general model, both specialized indoor and outdoor models perform better in detecting objects in specific places. As a result, when objects are located in similar places (for example, they are all indoors), the network can be trained and detects objects more accurately. We have observed that the numbers of images and classes of the general model have no positive effect on the model's performance, which even led to an increase in the network's misdetection in recognizing

objects. The statistical t-test also showed and confirmed a significant difference between the results obtained from specialized and general models.

Considering the objective, and the experimental evaluation of the specific models, we could state that; *the specialized scene-specific indoor and outdoor models outperform the general model, and training the networks with images from similar places enables them to detect objects more accurately based on their appearances, identical characteristics, and similar scene contextual information.*

Despite the detailed examination of the results acquired in this paper, some limitations still need to be investigated. We used one dataset and one object detection model to train the networks. To verify and extend the results of the specialized models, conducting additional experiments on other datasets and object detection models would be advantageous. Another interesting future work would be to design and examine more specialized models for object detection that are not limited to indoor and outdoor places. For example, designing a model that detects objects based on their exact location, such as a library, an office, a park, etc., and takes advantage of more precise contextual information about different places.

REFERENCES

[1] Kaur, J. and Singh, W., "Tools, techniques, datasets and application areas for object detection in an image: a review," *Multimedia Tools and Applications* **81**(27), 38297–38351 (2022).

[2] Kaur, R. and Singh, S., "A comprehensive review of object detection with deep learning," *Digital Signal Processing* , 103812 (2022).

[3] Chauhan, R., Ghanshala, K. K., and Joshi, R., "Convolutional neural network (cnn) for image detection and recognition," in [*2018 first international conference on secure cyber computing and communication (ICSCCC)*], 278–282, IEEE (2018).

[4] Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 580–587 (2014).

[5] Girshick, R., "Fast r-cnn," in [*Proceedings of the IEEE International Conference on Computer Vision*], 1440–1448 (2015).

[6] Ren, S., He, K., Girshick, R., and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems* **28** (2015).

[7] Dai, J., Li, Y., He, K., and Sun, J., "R-fcn: Object detection via region-based fully convolutional networks," *Advances in Neural Information Processing Systems* **29** (2016).

[8] He, K., Gkioxari, G., Dollár, P., and Girshick, R., "Mask r-cnn," in [*Proceedings of the IEEE International Conference on Computer Vision*], 2961–2969 (2017).

[9] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., "You only look once: Unified, real-time object detection," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 779–788 (2016).

[10] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C., "Ssd: Single shot multibox detector," in [*Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*], 21–37, Springer (2016).

[11] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection," in [*Proceedings of the IEEE International Conference on Computer Vision*], 2980–2988 (2017).

[12] Tan, M., Pang, R., and Le, Q. V., "Efficientdet: Scalable and efficient object detection," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 10781–10790 (2020).

[13] Law, H. and Deng, J., "Cornernet: Detecting objects as paired keypoints," in [*Proceedings of the European Conference on Computer Vision (ECCV)*], 734–750 (2018).

[14] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q., "Centernet: Keypoint triplets for object detection," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 6569–6578 (2019).

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need," *Advances in Neural Information Processing Systems* **30** (2017).

[16] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S., "End-to-end object detection with transformers," in [*Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*], 213–229, Springer (2020).

[17] Saleh, K., Szénási, S., and Vámossy, Z., "Occlusion handling in generic object detection: A review," in [*2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*], 000477–000484, IEEE (2021).

[18] Mukherjee, R., Bessa, M., Melo-Pinto, P., and Chalmers, A., "Object detection under challenging lighting conditions using high dynamic range imagery," *IEEE Access* **9**, 77771–77783 (2021).

[19] Liu, Y., Wang, R., Shan, S., and Chen, X., "Structure inference net: Object detection using scene-level context and instance-level relationships," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 6985–6994 (2018).

[20] Bell, S., Zitnick, C. L., Bala, K., and Girshick, R., "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 2874–2883 (2016).

[21] Shrivastava, A. and Gupta, A., "Contextual priming and feedback for faster r-cnn," in [*Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*], 330–348, Springer (2016).

[22] Zeng, X., Ouyang, W., Yang, B., Yan, J., and Wang, X., "Gated bi-directional cnn for object detection," in [*Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*], 354–369, Springer (2016).

[23] Luo, H.-W., Zhang, C.-S., Pan, F.-C., and Ju, X.-M., "Contextual-yolov3: Implement better small object detection based deep learning," in [*2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*], 134–141, IEEE (2019).

[24] Corsel, C. W., van Lier, M., Kampmeijer, L., Boehrer, N., and Bakker, E. M., "Exploiting temporal context for tiny object detection," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 79–89 (2023).

[25] Beery, S., Wu, G., Rathod, V., Votel, R., and Huang, J., "Context r-cnn: Long term temporal context for per-camera object detection," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 13075–13085 (2020).

[26] Fang, P. and Shi, Y., "Small object detection using context information fusion in faster r-cnn," in [*2018 IEEE 4th International Conference on Computer and Communications (ICCC)*], 1537–1540, IEEE (2018).

[27] Chan, Y.-T., "Deep learning-based scene-awareness approach for intelligent change detection in videos," *Journal of Electronic Imaging* **28**(1), 013038 (2019).

[28] Li, X., Ye, M., Liu, Y., and Zhu, C., "Adaptive deep convolutional neural networks for scene-specific object detection," *IEEE Transactions on Circuits and Systems for Video Technology* **29**(9), 2538–2551 (2019).

[29] Luo, D., Lei, S., Guo, P., Gao, C., Chen, Y., Li, J., and Wei, L., "Learning scene-specific object detectors based on a generative-discriminative model with minimal supervision," *Pattern Recognition Letters* **159**, 108–115 (2022).

[30] Hattori, H., Naresh Boddeti, V., Kitani, K. M., and Kanade, T., "Learning scene-specific pedestrian detectors without real data," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 3819–3827 (2015).

[31] Yao, J., Fidler, S., and Urtasun, R., "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in [*2012 IEEE conference on computer vision and pattern recognition*], 702–709, IEEE (2012).

[32] Stalder, S., Grabner, H., and Gool, L., "Exploring context to learn scene specific object detectors," in [*Proc. PETS*], **3** (2009).

[33] Ardö, H., Ahrnbom, M., and Nilsson, M., "Height normalizing image transform for efficient scene specific pedestrian detection," in [*2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*], 1–11 (2022).

[34] Zagoruyko, S. and Komodakis, N., "Wide residual networks," *arXiv preprint arXiv:1605.07146* (2016).

[35] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A., "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(6), 1452–1464 (2017).

[36] Horvat, M. and Gledec, G., "A comparative study of yolov5 models performance for image localization and classification," in [*Central European Conference on Information and Intelligent Systems*], 349–356, Faculty of Organization and Informatics Varazdin (2022).

[37] et. al., G. J., "ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support," (Oct. 2021).

[38] Redmon, J. and Farhadi, A., "Yolo9000: better, faster, stronger," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 7263–7271 (2017).

[39] Redmon, J. and Farhadi, A., "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767* (2018).

[40] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M., "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934* (2020).

[41] Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., et al., "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976* (2022).

[42] Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M., "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 7464–7475 (2023).

[43] Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., and Yeh, I.-H., "Cspnet: A new backbone that can enhance learning capability of cnn," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*], 390–391 (2020).

[44] He, K., Zhang, X., Ren, S., and Sun, J., "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(9), 1904–1916 (2015).

[45] Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J., "Path aggregation network for instance segmentation," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 8759–8768 (2018).

[46] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A., "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision* **88**, 303–338 (2010).

[47] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L., "Microsoft coco: Common objects in context," in [*Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*], 740–755, Springer (2014).

[48] Tong, K., Wu, Y., and Zhou, F., "Recent advances in small object detection based on deep learning: A review," *Image and Vision Computing* **97**, 103910 (2020).

[49] Patel, T. A., Dabhi, V. K., and Prajapati, H. B., "Survey on scene classification techniques," in [*2020 6th international conference on advanced computing and communication systems (ICACCS)*], 452–458, IEEE (2020).

AUTHORS' BACKGROUND

Name	Title	Research field
Mahtab Jamali	PhD candidate	Machine learning
Paul Davidsson	Full professor	Internet of Things, AI
Martin Georg Ljungqvist	Researcher	Computer vision
Reza Khoshkangini	Senior lecturer	Machine learning
Radu-Casian Mihailescu	Senior lecturer	Machine learning