

Occlusion Handling in Generic Object Detection: A Review

Kaziwa Saleh

Doctoral School of Applied Informatics
and Applied Mathematics

Óbuda University

Budapest, Hungary

kaziwa.saleh@stud.uni-obuda.hu

Sándor Szénási

John von Neumann Faculty of
Informatics

Óbuda University

Budapest, Hungary

szenasi.sandor@nik.uni-obuda.hu

Zoltán Vámossy

John von Neumann Faculty of
Informatics

Óbuda University

Budapest, Hungary

vamossy.zoltan@nik.uni-obuda.hu

Abstract— The significant power of deep learning networks has led to enormous development in object detection. Over the last few years, object detector frameworks have achieved tremendous success in both accuracy and efficiency. However, their ability is far from that of human beings due to several factors, occlusion being one of them. Since occlusion can happen in various locations, scale, and ratio, it is very difficult to handle. In this paper, we address the challenges in occlusion handling in generic object detection in both outdoor and indoor scenes, then we refer to the recent works that have been carried out to overcome these challenges. Finally, we discuss some possible future directions of research.

Keywords—Object detection, Indoor scene, Outdoor scene, Generative Adversarial Network, Amodal Perception, Instance segmentation, Compositional models

I. INTRODUCTION

As human beings, we are extremely fast and accurate in detecting and recognizing objects in surrounding environments even under various conditions where the object is partially visible. Our minds are capable of compensating for invisible parts and link the visible areas to identify the object [1]. Computers are still far away from accomplishing this task. However, since the development of deep neural networks and availability of large amount data, there has been a remarkable progress in the field of computer vision in general and object detection in particular. Object detection consists of two sub-tasks: classification and localization of an object.

Object detectors are categorized into two types: one-stage detectors, and two-stage detectors. The latter uses region proposal network to produce regions of interest (ROI) and apply a deep neural network to classify each proposal into class categories. The first type, however, considers the object detection as a regression problem, hence it uses a unified framework to learn the class probabilities and coordinates of bounding boxes. This makes one-stage detectors faster compared to its counterpart. The most effective state-of-art detectors are Faster RCNN [2], SSD [3] and YOLO [4].

However, object detection is a challenging task due to many factors such as clutter, imaging conditions, large number of object categories and instances, and occlusion [5].

Occlusion happens either when the object is hidden by the same type of object, which is called intra-class occlusion, or the object is occluded by a fixed element or an object of another type, this is called inter-class occlusion.

With partial occlusion, the deep neural network based classifiers are less robust compared to humans [6], and it worsens the performance of detectors [7]. Therefore, occlusion handling has been studied extensively such as in pedestrian detection [8] [9] [10] [11], object tracking [12] [13] [14], face detection [15] [16], stereo images [17], car detection [18] [19], semantic part detection [20] [21], etc. However due

to a huge number of variations in object category and instances, occlusion handling in generic object detection from a single still image is much harder.

Despite the existence of many datasets for object detection in outdoor scenes, the major problem in dealing with occlusion is the lack of availability of annotated occluded data. Other problems are the detection of occlusion existence, recovering the occluded region(s) of the object, and detecting the occluded object.

Occlusion has also been studied in indoor scenes, but there are several problems which makes it challenging. First, the rigid nature of furniture limits viewing the object fully from different angles. Second, there is not a large-scale dataset of real indoor scenes for occlusion. Finally, the size of objects is relatively smaller compared to those found in outdoors, which means when they are occluded, the visible region may not have sufficient information to be recognized. Thus it becomes harder to regenerate the occluded object.

There are several review and survey papers about object detection available in the literature. Among the most recent ones is the work of Liu et al. in [5] which surveys the available datasets and methods for generic object detection based on deep learning. Also Chen et al. [22] present the challenges and solutions for small object detection, and Zhao et al. [23] provide a review and an analysis for recent deep learning methods for object detection.

Nevertheless, to our best knowledge there is not a recent review about occlusion handling in the literature. Due to the significance of handling occlusion in generic object detection, we provide a review of the recent works that have been done in this field. Then we propose some directions for future research.

This review has been done with respect to object detection in still images. As such, occlusion handling in other applications are out of the scope of this work.

II. APPLICATIONS OF OBJECT DETECTION

Two significant applications of object detection in outdoor scenes are autonomous driving [24] [25] and object tracking. In autonomous vehicles, the car needs to have the ability to detect variety of objects on the road such as other vehicles, pedestrians, traffic and road signs, obstacles, etc. [26].

On the other hand, the most apparent application of detecting objects in indoor environments is the ‘fetch-and-delivery’ task which is the main function of service robots [27]. The robot not only needs to identify the objects around it in order to find the object it is looking for, but it also has to semantically label the region of space it is located in and categorize it, which is another application called scene understanding and categorization [28].

III. DATASETS

Since the development of deep neural networks, datasets have been the crucial element in the advancement of the field of computer vision. The most widely used datasets for generic object detection are PASCAL VOC [29], MS COCO [30], ImageNet [31], and Open Images [32]. Liu et al. [5] have thoroughly discussed these datasets. To avoid repeating what is already explained in [5], we will focus on the datasets for images in indoor environments.

One of the available datasets is introduced by Ehsani et al. in [33] called DYCE dataset which contains synthetic occluded objects. The images are taken in indoor scenes. There are 11 synthetic scenes, which contain 5 living rooms and 6 kitchen. Every scene has 60 objects and the number of visible (at least 10 visible pixels) objects per image is 17.5.

Furthermore, authors in [34] present TUT indoor dataset which contains 2213 frames containing 4595 object instances from 7 classes. The size of each frame is 1280×720 . But the number of instances are not equal through the classes. There are 1684 maximum instances from a single class while the minimum is 81. The dataset has different backgrounds, light conditions, occlusion, and high inter-class variations. An object detector (Faster RCNN using ResNet-101 [35]) is trained to generate proposal annotations for half of the dataset. While the other smaller half is manually annotated by a human. Thus a faster bounding box annotation is presented in the study.

To efficiently train a model on occlusions in real indoor scenes we need a large scale fully annotated dataset with ground truth labels for occluded objects in the scene. Unfortunately, there is no such dataset yet.

For outdoor scenes, Qi et al. [36] created KINS dataset from KITTI [37] for amodal instance segmentation. It contains amodal instance segmentation masks and corresponding occlusion order. Three expert annotators marked every image. Ambiguities were resolved by crowdsourcing to ensure that the occluded regions were labeled consistently.

Apart from the previously mentioned datasets, there are others such as ScanObjectNN (contains indoor scenes created with CAD data) [38], BigBird [39], NYC v2 [40], Places [41], and datasets for 3D indoor reconstruction and Simultaneous localization and mapping (SLAM) mentioned in [42]. But those are out of the scope of this review.

IV. OCCLUSION HANDLING IN OUTDOOR SCENES

In this section we address the challenges in occlusion handling in outdoor environments and the solutions for each of them.

A. Data Collection

Because of large variety of object categories and instances, collecting and labeling a dataset with possible occlusions of each instance in every category seems impossible. Consequently, many studies rely on synthetic datasets or automatically generated examples:

1) Generative Adversarial Networks (GAN): Since their inventions by Goodfellow et al. [43], Generative Adversarial Networks (GAN) have been studied widely to train a generative model. The framework has two adversarial networks which are trained simultaneously, namely a generator and a discriminator. Whereas the generator is trained to learn mapping samples from a random latent space to data, the discriminator is trained to differentiate between real and generated (fake) samples. The goal of the generator is to fool the discriminator through making samples that appear similar to the real data as possible. To ensure that the generated data comes from a particular class or data, Conditional Generative Adversarial Networks (cGAN) can be used [44].

Wang et al. [45] argue that potential occlusions and deformations cannot be covered even by large scale datasets. Thus they propose a strategy that uses two adversarial networks (shown in “Fig. 1”) to generate examples from COCO dataset that would be hard for a Fast-RCNN to classify. The first adversarial network is Adversarial Spatial Dropout Network (ASDN) which learns how to occlude an object, and the second network is Adversarial Spatial Transformer Network (ASTN) which learns how to rotate object parts to create deformations. By training these two networks simultaneously against the Fast-RCNN, the latter learns to deal with occlusions and deformations.

2) Amodal perception: is the ability of inferring the object’s physical structure when parts of it are occluded. Recent studies such as [46] [47] [36] have used it for instance segmentation. Amodal segmentation is significant for occlusion handling because through comparing amodal and modal segmentation masks the existence, degree, contour, and parts of occlusion can be deduced [46].

Nonetheless, the difficulty of data preparation for amodal instance segmentation makes it a challenging task. Therefore, Li and Malik [46] create amodal training data by adding synthetic occlusion to the modal mask. First, an image patch that has at least one foreground object instance is randomly cropped. Afterwards, the patch is overlaid by random object instances that are extracted from other images. The random position and scale of the overlaid object ensures a mild occlusion with the bottom patch. Then by using the original modal segmentation mask, the pixels congruous to the mask in each patch are labeled as positive (belongs to the object),

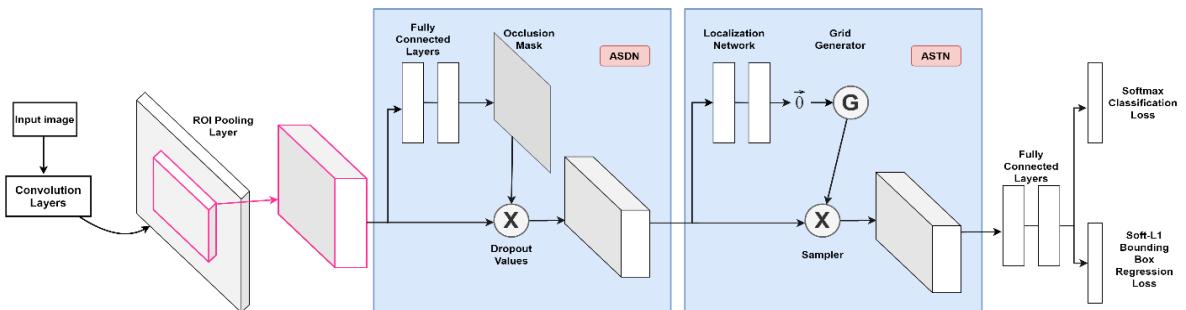


Fig. 1. Proposed architecture for combining ASDN and ASTN in [45]

negative (background), and unknown (belongs to other objects). Eventually, the original modal mask contains parts of the occluded object in the composite image that were originally visible. This becomes the true amodal segmentation mask for the composite patch.

The generated composite image and the amodal mask are used to train a CNN to attain the original mask. When tested on images with real occlusion, the model can effectively predict amodal masks even if it is trained on synthetic data.

B. Occlusion Detection

One of key problems in occlusion handling is to determine whether the object in question is occluded by other objects or not. It is hard to tell if the observed appearance of the object in the image is the object's real shape or it is a result of occlusion.

Qi et al. in [36] propose Multi-Level Coding (MLC) network with an occlusion classification branch to improve the amodal perception ability of inferring the occluded parts. MLC has two branches: extraction and combination. In the first branch the abstract global features of the objects are extracted. To produce the masks for visible and invisible parts, the combination branch fuses global and specific local features. Meanwhile, the occlusion classification predicts the existence of occlusion which enhances the network's amodal perception. Experimental results on KINS dataset show that amodal and inmodal instance segmentation can be enhanced through using the proposed model.

C. Generating the Occluded Regions

A major challenge in handling occlusion is determining how the invisible parts of the object can be recovered. Currently there are three solutions:

1) Amodal instance segmentation: In addition to previously mentioned works in amodal instance segmentation, Follmann et al. [48] propose an end-to-end trainable amodal instance segmentation model dubbed Occlusion R-CNN (ORCNN). The model is an extension of Mask R-CNN with amodal mask head and an occlusion mask head for predicting amodal, inmodal, and occlusion masks for object instances concurrently in a single forward pass (as seen in "Fig. 2"). The authors also introduce a new D2S amodal dataset and COCOA cls. The first one is based on D2S [49] and the latter is originated from COCOA dataset from [47]. Data augmentation is used to include moderate to heavy occlusion occluded objects in D2S amodal dataset. The results verify that the model obtains competitive results on D2S amodal without any amodally annotated data, it even performs better than its baselines on COCOA cls dataset.

2) Partial completion: a self-supervised framework developed by Zhan et al. [50] for partial completion of occluded objects for scene de-occlusion. The framework depends on two principles of partial completion concept. First, in case of having an object occluded by several other objects, the partial completion can be performed progressively with one object involved at a time. Second, by intentionally pruning an occluded object and training a network to recreate the unpruned object, the network can learn to partially complete the occluded object. The authors attain partial completion through two networks: Partial Completion Network-mask (PCNet-M) and Partial Completion Network-content (PCNet-C). While the first network is used to generate the occlusion mask of the occluded object corresponding to the occluding object, the second network furnishes the mask with RGB content. The framework is tested on KINS [36] dataset and COCOA [47], and the results illustrate a similar performance to fully supervised baselines although the framework is trained without ground truth occlusion ordering and amodal masks.

3) Context Encoders: Pathak et al. [51] suggest a convolutional neural network (CNN) that is able to generate the missing patches of an image based on context. The proposed model has an encoder and a decoder. While the encoder produces a compact latent feature representation from the image's context, the decoder can generate the missing part of the image from the produced representation. Due to its unsupervised nature, the model has to learn the semantic of the image and generate a reasonable hypothesis for the absent parts. Therefore, the model is trained to reduce a reconstruction loss which catches the structure of the absent part based on the context, and an adversarial loss which selects a specific mode from the distribution. The results show that the model can inpaint semantic parts of an image better than its baselines. However, the performance of the model deteriorates with high-level textures regions.

D. Occluded Object Detection

The accuracy of state-of-art object detectors and deep neural network based classifiers degrade with partial occlusion [52] [7]. As such, many studies focus on improving the ability of current methods to recognize and localize occluded objects. We divide these into two categories:

1) Classification: According to Fawzi and Frossard [53] Deep Convolution Neural Networks (DCNN) are not robust under partial occlusion.

DeVries and Taylor [54] argue that CNNs are prone to overfitting which leads to poor generalization in case of

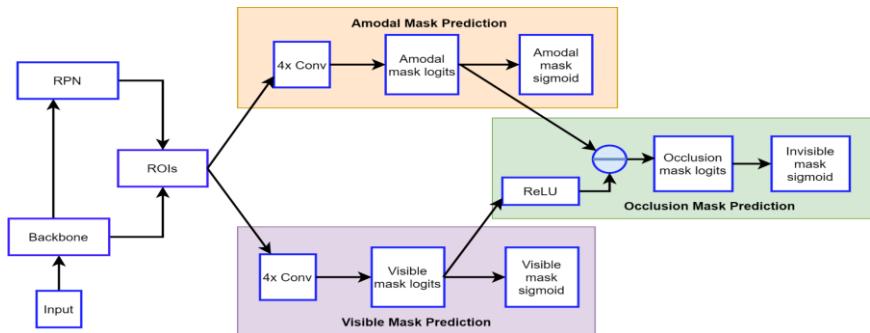


Fig. 2. Architecture of ORCNN [48]

occlusion. Thus they propose a simple regularization technique for CNN called Cutout, where they augment the training data with partially occluded images. For each input image, a pixel is randomly chosen as a center point for a fixed-size zero mask to drop out neighboring section of the image. However, increasing the training data leads to the increase in training time and cost [7].

On the other hand, Xiao et al. [55] propose TDAPNet to address two problems in DCNN: overfitting, and contamination of occlusion during feature extraction. They propose TDAPNet, a deep network which has three parts: prototype learning, partial matching, and a top-down attention mechanism. The first two parts help to solve the first addressed issue, whereas the third part deals with the second issue. After extracting the features by DCNN, prototype learning is applied. Then, partial matching compares features and prototypes by only comparing the visible parts thus removing irrelevant feature vectors. Finally, the top-down attention regulation produces more pure features around the occluder by filtering out irregular activations caused by occlusion. Through removing occluded features in the lower layers, the model becomes more robust against occlusion. But based on the experiments done by Kortylewski et al. [7] on the images with real occlusion, the model is not as reliable as it is with artificial occlusion.

Moreover, many recent studies have focused on using compositional models to handle occlusions [52] [56] [57]. Through compositional models, an object can be represented in terms of its parts and their spatial structure [52]. Compositional models have two benefits: 1) we can create an occlusion model which ignores regions of the model that are occluded. 2) The model can provide explanation about its results such as: the location of object's detected individual parts, and the occluded regions of the object [52].

Kortylewski et al. in [52] combine the discriminative power of DCNN and the ability of compositional to generalize well to partially occluded objects. In the proposed model, a standard DCNN is used for image classification during training. The extracted features are grouped into dictionaries. The elements within the dictionaries are similar to object part detectors and learn the spatial allocation of parts of each class. During testing, the DCNN tries to classify the input image through a feedforward pass. It is probable that the image is partially occluded if the network cannot certainly predict its class. Then the extracted features are used to detect the parts of the compositional model. The results demonstrate that the model can recognize partially occluded 3D objects better than DCNN even when it has been trained on occlusion-free data. However, the model is not as discriminative as DCNN at recognizing non-occluded objects [57].

2) *Detection*: despite the success of state-of-art detectors in detecting non-occluded objects, detection of occluded objects remains an open problem. Nevertheless, several recent works have implemented compositional models to detect partially occluded objects.

Kortylewski et al. in [57] introduce a unified model of DCNN and compositional models called Compositional Convolutional Neural Network (CompositionalNet). In their

proposed architecture, the authors apply a differentiable generative compositional layer instead of the fully-connected head in DCNN. The network can robustly classify the partially occluded objects and localize the occluders due to the generative characteristic of the compositional layer. The experiments show that, although CompositionalNet is trained with class labels only, it can localize occluders correctly. It also outperforms a standard DCNN and other related techniques in classifying partially occluded objects despite not being trained on occluded objects.

However, according to Wang et al. [56] CompositionalNets have no explicit separate representation of context from the object, subsequently the context has a negative effect on detection in severe occlusion. And there is no mechanism for robust bounding box prediction for partially occluded objects in CompositionalNets. To manage the effect of context on detecting the occluded objects, the authors suggest using bounding box annotations for segmenting the context. Meanwhile, the part-based voting technique in CompositionalNets is extended to consider the vote for the two opposite corners of the bounding box beside the object center. The results demonstrate that the proposed context-aware CompositionalNet can robustly detect and estimate bounding boxes even under severe occlusion.

Apart from the previously mentioned efforts, contextual information [58] [59] [60], and deformable convolution [61] [62] [63] can be used to reduce occlusion.

V. OCCLUSION HANDLING IN INDOOR SCENES

Due to complex and diverse layout of indoor scene structure, clutter, occlusion, and illumination, object detection is more challenging compared to that of outdoor scenes [42]. In this section we present the challenges of object occlusion handling in indoor environments, and existent methods in the literature to solve them:

A. Scene Structure

The layout and design of indoor elements and furniture, prevents obtaining full view of objects or even partial view in some cases. For example, looking for a hidden box behind other objects inside a cupboard. Because of the static and rigid nature of the cupboard, it would be impossible to view the object from different viewpoints. To overcome this issue, there are two solutions:

1) *Turntable*: is used to avoid occlusion by placing the objects on a round table and allow the camera that is installed on a robot to go around the table to capture the objects in different viewpoints [64]. The different views are used to provide the next-best view (NBV) of the object to model the invisible parts.

2) *Interactive manipulation*: authors of [65] argue that finding an occluded object within a cluttered scene cannot be done by using a single image. Therefore using a robotic arm with wrist-mounted RGB and depth camera, active perception and interactive perception are applied to find the object of interest. While active perception is where the camera is moved to capture the object from several viewpoints, interactive perception provides better understanding of the scene from interactions. Then a reinforcement learning based control algorithm and a color detector are used to find target object of a particular color.

Although the agent is trained with a detached gripper during simulation, some end-effector poses are not kinematically feasible in real world.

Also Dogar et al [66] state that the problem of searching for an object requires both perception and manipulation to move the objects that might hide the targeted object. They propose two search algorithms, the greedy search algorithm that considers visibility and accessibility relations between objects; and the connected components algorithm which utilizes the expected time to find the target as an optimization criterion.

Interactive manipulation of an object by a robot arm in front of a static camera is used by Krainin et al. [67] to generate a 3D model of an object.

Although the mentioned techniques might be necessary and required in certain cases where applicable, they require a controlled setup of the environment which is unrealistic and not always possible in real world.

B. Training Data

There is not a large-scale database available for object occlusion for indoor scenes.

Georgakis et al [27] argue that creating annotated dataset that covers all the possibilities of indoor environments such as differences in viewpoint, lighting conditions, occlusion and clutter, would be laborious and time consuming. Meanwhile the trained models poorly generalize across various environments and backgrounds. Therefore, they propose an approach to generate a new dataset from two other datasets: GMU-Kitchens and Washington RGB-D Scenes v2. The synthesized images are generated by superimposing object instances at several positions, scales, pose, and location.

On the other hand, Dwibedi et al [68] suggest an easy cut and paste method to synthesize training data with minimal effort. Their key intuition is based on state-of-the-art detectors such as Faster RCNN that mostly work on local region-based features rather than global-based features. The method automatically cuts object instances and pastes them on random backgrounds. However, to avoid subtle pixel artifacts the training algorithms is forced to ignore the artifacts and concentrate on the object appearance instead.

C. Recover the Occluded Regions

Occlusion can happen in different scale, location, and levels. Thus it is hard to train a model to regenerate the invisible regions of an object. The following methods are used to segment and regenerate the occluded area of objects:

1) *GAN*: Ehsani et al. [33] use GAN to address the problem of generating the occluded region of objects. Their

proposed model called SeGAN segments the invisible part first, then produces its appearance by painting. Based on an input image and a segmentation of the visible region of the object, the model produces an RGB image where the invisible region of the object is rebuilt. To do this, the model consists of two parts: the segmentation part which is a CNN that outputs a mask for the object by using the information from visible regions, and a painting part which uses a cGAN to produce the occluded parts of the object. The loss function of the SeGAN model is a combination of the losses for segmentation and painting. “Fig. 3”, shows the architecture of SeGAN model. The authors report better results of SeGAN compared to its baselines on their photo-realistic DYCE dataset.

2) *Semantic segmentation*: Purkait et al. [69] propose that instead of assigning a label per pixel in an image, a set of semantic labels can be achieved indicating visibility or invisibility per each pixel. Their work uses a synthesized dataset that was augmented from SUNCG dataset to predict the semantic category of visible regions and occluded regions. They use a U-Net architecture [70] that has an encoder and a decoder with ReLUs, but in the final layer a sigmoid activation function is used with a group-wise softmax. There are skip connections from the encoder to the decoder.

The results of the network with cross-entropy loss is compared with the network that utilizes a grouped semantic loss, the latter produces better results in predicting semantic labels for the occluded pixels. However since the training set is synthetic, the implementation of the proposed method is bounded to environments where such a synthetic models are found. In real world scenes, certain estimation of occluded parts has to be considered.

3) *Instance segmentation*: By extending the Mask RCNN architecture, Wada et al. [71] propose a model for occlusion segmentation which they call it “relook Mask RCNN”. To train their model, the authors create a dataset of synthesized images of a pile of object created from the instances of the objects. Instead of considering instance occlusion segmentation as a single class (visible only) problem, the authors look at it as a multi-class (visible, occluded). They also consider the relationship between the masks to infer the occlusion status of the instances correctly. To learn the relationship, the instance masks predicted from the Mask RCNN is converted to a density map to predict the instance mask in the second stage (relook stage). As an output, the model predicts three masks: visible, occluded, and other (not belonging to the object).

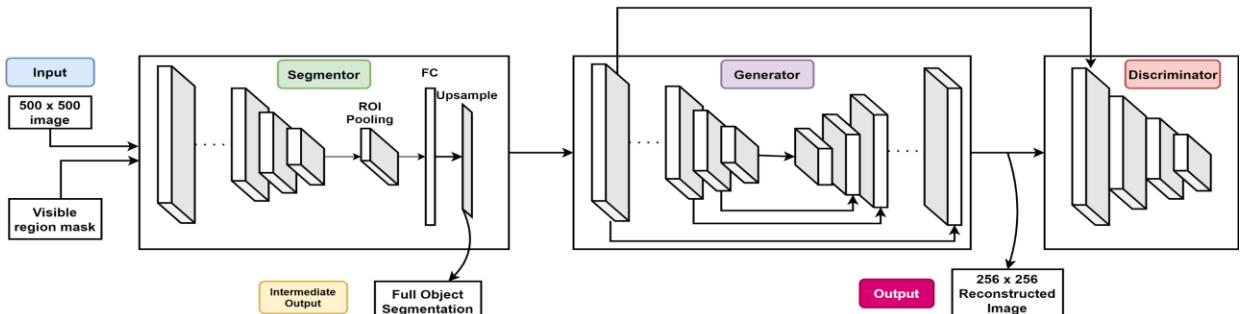


Fig. 3. Architecture of SeGAN as proposed in [33]

The proposed system is tested on real picking task from a stack of objects and the reported results illustrate the effectiveness of the system. However the system requires a dataset with all possibilities of occlusion state of the objects and their corresponding labels and masks, the effort for achieving this increases exponentially with the increase of the number of objects.

4) Contextual information: Huetting et al. [72] introduce SEETHROUGH model that utilizes the fact that there is a considerable regularity of objects co-occurrence in indoor scenes. By considering that co-occurrence information as explicit priors, the object identity, location and orientation can be predicted even with heavy inter- or intra-class occlusions. The authors train a neural network on real indoor annotated images to extract 2D keypoints. The extracted key points are fed into a 3D candidate object generation stage. Then object co-occurrence statistics extracted from a large 3D scene database is used to solve a selection problem between 3D object proposals. The process is repeated to use the location of the already discovered objects for incrementally detecting nearby candidates with low keypoint response. The results show that SeeThrough can more accurately detect chairs in scenes even with medium or heavy occlusion than its two baselines, Faster-3D RCNN and SeeingChairs.

VI. DISCUSSION AND FUTURE DIRECTIONS

Based on the reported results of GAN, amodal segmentation, and compositional models, we see a great potential of these techniques in occlusion detection and handling. GAN have been effectively used to enhance features when there are not enough of them in small object detection [73], and to regenerate occluded faces [74].

However, there remains a set of key problems in occlusion handling that could serve as a future direction of research in the field.

First, there is no large-scale dataset that could be used for object occlusion in indoor environments. The available ones currently are mostly synthetic datasets, which means when a model is trained on such a dataset it might not generalize well for a real world scene. In addition, the available datasets for outdoor scenes lack sufficient annotations for occluded regions.

Second, in order to apply de-occlusion on an object, it is vital that we first determine that the object is occluded or not. Although amodal segmentation can be used to resolve this issue, this solution may be ineffective if the annotation is incorrect or insufficient.

Finally, increase the ability of current models to regenerate the occluded objects under different degrees of occlusion using the real world training data instead of synthetic data.

VII. CONCLUSION

Object detection has come a long way since the invention of region based networks. Not only they are more accurate than before but they also achieve real-time results. Although there is abundant studies in generic and specific object category detection, occlusion handling in generic object detection remains comparatively unexplored. Many prominent models fail to detect objects when they are

occluded. In this review, we have presented the most recent works that endeavor to resolve occlusions and in many cases regenerate the object. We have also addressed the future directions for research if we want to overcome challenges of occlusion handling in generic object detection.

ACKNOWLEDGEMENT

The authors would like to thank the GPGPU Programming Research Group of Óbuda University for its valuable support.

REFERENCES

- [1] R. A. Rensink and J. T. Enns, "Early completion of occluded objects," *Vision Research*, vol. 38, no. 15-16, pp. 2489-2505, 1998.
- [2] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in neural information processing systems*, pp. 91-99, 2015.
- [3] W. Liu et al., "SSD: Single shot multibox detector," in *European conference on computer vision*, 2016.
- [4] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779-788, 2016.
- [5] L. Liu et al., "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261-318, 2020.
- [6] H. Zhu, P. Tang, J. Park, S. Park and A. Yuille, "Robustness of object recognition under extreme occlusion in humans and computational models," *arXiv preprint arXiv:1905.04598*, 2019.
- [7] A. Kortylewski, Q. Liu, A. Wang, Y. Sun and A. Yuille, "Compositional Convolutional Neural Networks: A Robust and Interpretable Model for Object Recognition Under Occlusion," *Int J Comput Vis*, 2020.
- [8] S. Zhang, L. Wen, X. Bian, Z. Lei and S. Z. Li, "Occlusion-aware R-CNN: detecting pedestrians in a crowd," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [9] Y. Tian, P. Luo, X. Wang and X. Tang, "Deep Learning Strong Parts for Pedestrian Detection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [10] S. Gilroy, E. Jones and M. Glavin, "Overcoming Occlusion in the Automotive Environment-A Review," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1-13, 2019.
- [11] C. Ning, L. Menglu, Y. Hao, S. Xueping and L. Yunhong, "Survey of pedestrian detection with occlusion," *Complex Intell. Syst.*, 2020.
- [12] Y. Liu, X. Jing, J. Nie, H. Gao, J. Liu and G. Jiang, "Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in RGB-D videos," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 664-677, 2018.
- [13] S. Zhao, S. Zhang and L. Zhang, "Towards occlusion handling: object tracking with background estimation," *IEEE Transactions on Cybernetics*, vol. 48, no. 7, pp. 2086-2100, 2017.
- [14] J. Ong, B. -T. Vo, B. -N. Vo, D. Y. Kim and S. Nordholm, "A Bayesian Filter for Multi-view 3D Multi-object Tracking with Occlusion Handling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [15] D. Zeng, R. Veldhuis and L. Spreeuwiers, "A survey of face recognition techniques under occlusion," *arXiv:2006.11366v1*, 2020.
- [16] S. Yang, P. Luo, C. C. Loy and X. Tang, "Faceness-Net: Face Detection through Deep Facial Part Responses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1845-1859, 2018.
- [17] A. Li and Z. Yuan, "SymmNet: a symmetric convolutional neural network for occlusion detection," *arXiv preprint arXiv:1807.00959*, 2018.

- [18] N. D. Reddy, M. Vo and S. G. Narasimhan, "Occlusion-Net: 2D/3D Occluded Keypoint Localization Using Graph Networks," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
- [19] S. Yan and Q. Liu, "Inferring occluded features for fast object detection," *Signal Processing*, vol. 110, pp. 188-198, 2015.
- [20] J. Wang, C. Xie, Z. Zhang, J. Zhu, L. Xie and A. L. Yuille, "Detecting semantic parts on partially occluded objects," arXiv preprint arXiv:1707.07819, 2017.
- [21] Z. Zhang, C. Xie, J. Wang, L. Xie and A. L. Yuille, "Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1372-1380, 2018.
- [22] G. Chen et al., "A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020.
- [23] Z. Zhao, P. Zheng, S. Xu and X. Wu, "Object Detection With Deep Learning: A Review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212-3232, 2019.
- [24] D. Drexler, Á. Takács, T. Nagy D. and T. Haidegger, "Handover Process of Autonomous Vehicles – technology and application challenges," *Acta Polytechnica Hungarica*, vol. 16, no. 9, pp. 235-255, 2019.
- [25] G. Gosztolya, T. Grósz, L. Tóth, A. Markó and T. G. Csapó, "Applying DNN Adaptation to Reduce the Session Dependency of Ultrasound Tongue Imaging-based Silent Speech Interfaces," *Acta Polytechnica Hungarica*, vol. 17, no. 7, pp. 109-124, 2020.
- [26] J. Lu, S. Tang, J. Wang, H. Zhu and Y. Wang, "A Review on Object Detection Based on Deep Convolutional Neural Networks for Autonomous Driving," in 2019 Chinese Control And Decision Conference (CCDC), 2019.
- [27] G. Georgakis, A. Mousavian, A. C. Berg and J. Kosecka, "Synthesizing Training Data for Object Detection in Indoor Scenes," arXiv:1702.07836v2.
- [28] Young et al., "Making sense of indoor spaces using semantic web mining and situated robot perception," in European Semantic Web Conference, Springer, 2017, pp. 299-313.
- [29] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303-338, 2010.
- [30] T.Y. Lin et al, "Microsoft COCO: Common Objects in Context," in European conference on computer vision, 2014.
- [31] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017.
- [32] A. Kuznestsova et al, "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale," arXiv preprint arXiv:1811.00982, 2018.
- [33] K. Ehsani, R. Mottaghi and A. Farhadi, "SeGAN: Segmenting and Generating the Invisible," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6144-6153.
- [34] B. Adhikari, J. Peltomaki, J. Puura and H. Huttunen, "Faster Bounding Box Annotation for Object Detection in Indoor Scenes," in 2018 7th European Workshop on Visual Information Processing (EUVIP), Tempere, 2018.
- [35] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778, 2016.
- [36] L. Qi, L. Jiang, S. Liu, X. Shen and J. Jia, "Amodal Instance Segmentation with KINS Dataset," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.
- [37] A. Geiger, P. Lenz and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in CVPR, 2012.
- [38] M. Uy, Q. Pham, B. Hua, D. Nguyen and S. Yeung, "Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1588-1579.
- [39] A. Singh, J. Sha, K. S. Narayan, T. Achim and P. Abbeel, "Bigbird: A large-scale 3d database of object instances," in 2014 IEEE international conference on robotics and automation (ICRA), Hong Kong, IEEE, 2014, pp. 509-516.
- [40] N. Silberman, D. Hoiem, P. Kohli and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in ECCV, 2012.
- [41] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba, "Places: A 10 million Image Database for Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [42] Z. Kang, J. Yang, Z. Yang and S. Cheng, "Environments, A Review of Techniques for 3D Reconstruction of Indoor," *International Journal of Geo-Information*, vol. 9, no. 330, pp. 1-31, 2020.
- [43] I. Goodfellow et al., "Generative Adversarial Nets," in Advances in neural information processing systems, 2014, pp. 2672-2680.
- [44] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.
- [45] X. Wang, A. Shrivastava and A. Gupta, "A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017.
- [46] K. Li and J. Malik, "Amodal Instance Segmentation," in European Conference on Computer Vision, 2016.
- [47] Y. Zhu, Y. Tian, D. Metaxas and P. Dollár, "Semantic Amodal Segmentation," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017.
- [48] P. Follmann, R. König, P. Härtinger, M. Klostermann and T. Böttger, "Learning to See the Invisible: End-to-End Trainable Amodal Instance Segmentation," in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 2019.
- [49] P. Follmann, T. Böttger, P. Härtinger, R. König and M. Ulrich, "MVTec D2S: Densely Segmented Supermarket Dataset," *Computer Vision - ECCV 2018*, pp. 581-597, 2018.
- [50] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin and C. C. Loy, "Self-Supervised Scene De-occlusion," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020.
- [51] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2536-2544, 2016.
- [52] A. Kortylewski, Q. Liu, H. Wang, Z. Zhang and A. Yuille, "Combining Compositional Models and Deep Networks For Robust Object Classification under Occlusion," in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 2020.
- [53] A. Fawzi and P. Frossard, "Measuring the effect of nuisance variables on classifiers," in British Machine Vision Conference (BMVC), 2016.
- [54] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," arXiv preprint arXiv:1708.04552, 2017.
- [55] M. Xiao, A. Kortylewski, R. Wu, S. Qiao, W. Shen and A. Yuille, "TDAPNet: Prototype Network with Recurrent Top-Down Attention for Robust Object Classification under Partial Occlusion," arXiv preprint arXiv:1909.03879, 2019.
- [56] A. Wang, Y. Sun, A. Kortylewski and A. Yuille, "Robust Object Detection Under Occlusion With Context-Aware CompositionalNets," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020.
- [57] A. Kortylewski, Q. J. He, Liu and A. L. Yuille, "Compositional Convolutional Neural Networks: A Deep Architecture With Innate Robustness to Partial Occlusion," in 2020 IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020.

- [58] T. Gao, B. Packer and D. Koller, "A segmentation-aware object detection model with occlusion handling," in CVPR 2011, Providence, RI, 2011.
- [59] E. Hsiao and M. Hebert, "Occlusion reasoning for object detection under arbitrary viewpoint," in IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012.
- [60] J. Winn and J. Shotton, "The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 2006.
- [61] J. Dai et al., "Deformable Convolutional Networks," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017.
- [62] T. Mordan, N. Thome, G. Henaff and M. Cord, "End-to-End Learning of Latent Deformable Part-Based Representations for Object Detection," Int J Comput Vis 127, pp. 1659-1679, 2019.
- [63] W. Ouyang. and et al., "Deepid-net: Deformable deep convolutional neural networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.
- [64] T. Fäulhammer et al., "Autonomous Learning of Object Models on a Mobile Robot," IEEE Robotics and Automation Letters, vol. 2, no. 1, pp. 26-33, 2017.
- [65] T. Novkovic, R. Pautrat, F. Furrer, M. Breyer, R. Siegwart and J. Nieto, "Object Finding in Cluttered Scenes Using Interactive Perception," in 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 2020, pp. 8338-8344.
- [66] M. R. Dogar, M. Koval, A. Tallavajhula and S. Srinivasa, "Object search by manipulation," Autonomous Robots, vol. 36, no. 1-2, pp. 153-167, 2013.
- [67] M. Krainin, B. Curless and D. Fox, "Autonomous generation of complete 3D object models using next best view manipulation planning," in 2011 IEEE International Conference on Robotics and Automation, Shanghai, 2011.
- [68] D. Dwibedi, I. Misra and M. Hebert, "Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
- [69] P. Purkait, C. Zach and I. Reid, "Seeing Behind Things: Extending Semantic Segmentation to Occluded Regions," in 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 2019.
- [70] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, 2015.
- [71] K. Wada, S. Kitagawa, K. Okada and M. Inaba, "Instance Segmentation of Visible and Occluded Regions for Finding and Picking Target from a Pile of Objects," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, 2018.
- [72] M. Huetting, P. Reddy, V. Kim, N. Carr, E. Yumer and N. Mitra, "SEETHROUGH: Finding Chairs in Heavily Occluded Indoor Scene Images," arXiv preprint arXiv:1710.10473, 2017.
- [73] Y. Zhang, Y. Bai, M. Ding and B. Ghanem, "Multi-task Generative Adversarial Network for Detecting Small Objects in the Wild," International Journal of Computer Vision, pp. 1-19, 2020.
- [74] J. Dong, L. Zhang, H. Zhang and W. Liu, "Occlusion-Aware GAN for Face De-Occlusion in the Wild," in 2020 IEEE International Conference on Multimedia and Expo (ICME), London, United Kingdom, 2020.