Multi-modal Variational Faster R-CNN for Improved Visual Object Detection in Manufacturing

Panagiotis Mouzenidis Antonios Louros Dimitrios Konstantinidis
Kosmas Dimitropoulos Petros Daras

Centre for Research and Technology Hellas (CERTH)
57001, 6th km Charilaou-Thermi, Thessaloniki, Greece
{mouzenidis, alouros, dikonsta, dimitrop, daras}@iti.gr

Theofilos Mastos
KLEEMANN HELLAS SA
61100, Industrial Area of Kilkis, Greece
t.mastos@kleemannlifts.com

Abstract

Visual object detection is a critical task for a variety of industrial applications, such as robot navigation, quality control and product assembling. Modern industrial environments require AI-based object detection methods that can achieve high accuracy, robustness and generalization. To this end, we propose a novel object detection approach that can process and fuse information from RGB-D images for the accurate detection of industrial objects. The proposed approach utilizes a novel Variational Faster R-CNN algorithm that aims to improve the robustness and generalization ability of the original Faster R-CNN algorithm by employing a VAE encoder-decoder network and a very powerful attention layer. Experimental results on two object detection datasets, namely the well-known RGB-D Washington dataset and the QCONPASS dataset of industrial objects that is first presented in this paper, verify the significant performance improvement achieved when the proposed approach is employed.

1. Introduction

With the advent of Industry 4.0 and the technological breakthroughs in ICT technologies, there is a growing demand for automation in industrial environments by using robotics and relevant applications to assist workers in their tasks and facilitate production [9]. One such automatic procedure is object detection that aims at identifying the location and class of an industrial object from images captured by specialized sensors and processed by machine and deep learning algorithms. Object detection in industrial environments is crucial for a huge variety of robotic applications involving assembling, sorting, robot navigation, fault detection, etc.

However, object detection is a very challenging task as it involves the detection of a large number of industrial objects that may vary slightly in appearance and/or size. In addition, object detection requires significant computational time, adaptability to different objects of the same category, and durability against lighting conditions, dust, and fast-changing environments [18]. To this end, it is imperative to develop automatic object detection algorithms that can achieve high accuracy, robustness and generalization. Most research works explore the processing of RGB images using highly specialized object detection algorithms [1, 2, 4, 22], but there is limited research on the use of additional data modalities and/or generative models to further enhance the accuracy and the generalization ability of object detection algorithms. Such limitations are further enhanced by the lack of large publicly available datasets of industrial objects.

To overcome the aforementioned challenges, this work introduces a new dataset of RGB-D images of industrial objects and proposes a novel deep learning approach for the detection of such objects based on the processing of RGB and depth images. The proposed approach utilizes a novel enhanced variant of the well-known Faster R-CNN algorithm [20], called Variational Faster R-CNN for the processing of RGB and depth images separately prior to the fusion (i.e., concatenation) of the RGB and depth information and the further processing using another Variational Faster R-CNN network architecture. The novel Variational Faster R-CNN algorithm extracts image features using a ResNet network, forwards these features to a Variational Auto-Encoder (VAE) network architecture and an Efficient Channel Attention (ECA) layer and finally feeds them to Region Proposal Networks (RPN) and classification modules for object classification and localization. The aim of the Variational Faster R-CNN algorithm is to improve the generalization ability and accuracy of the original Faster R-CNN algorithm through the projection of the image features to a highly discriminative latent space and a powerful attention mechanism. The main contributions of this work are summarized below:

- We propose the novel Variational Faster R-CNN that extracts features from RGB and Depth images, feeds these features to a VAE framework with an ECA attention layer and then processes these features using a RPN and a classification module for robust object detection results.
- We introduce a new dataset of industrial objects, named QCONPASS, captured using a RGB-D sensor.

2. Related Work

Two-stage detectors, such as Faster R-CNN [20] and Mask R-CNN [12], are very popular algorithms for object detection and localization due to their high accuracy and speed. In the first stage, such detectors employ deep convolutional networks and region proposal networks to propose candidate image regions, where objects of interest may exist, while in the second stage they classify these regions to object classes. Especially, in industrial settings, where an autonomous robot is required to navigate and assist workers in their tasks, the lightweight Faster R-CNN algorithm is highly employed. Several works employed Faster-RCNN to train robotic arms in an industrial environment to identify and manipulate objects using RGB images [1, 2], while Saeed et al. [22] used the Faster R-CNN algorithm for fault detection in industrial images. In a different scenario, Sun et al. [23] employed Faster R-CNN to safely navigate an autonomous robot in a warehouse by identify shelf-legs and tags in an image.

With the high availability and the low cost of RGB-D sensors and to further improve the accuracy and robustness of object detection (RGB-based methods are usually sensitive to illumination changes), several works employed depth information as an additional modality, with complementary information to the RGB data [10]. Depth information can be really useful in industrial environments, in which the objects that needs to be detected vary significantly in size. Multi-modal information has been successfully incorporated in a Faster R-CNN network as well. Mocanu et al. in [17] proposed the processing of RGB and depth modalities using VGG networks in two streams and the early fusion of these streams prior to their introduction in the RPN network of a Faster R-CNN architecture. On the other hand, Zhu et al. in [25] proposed custom CNNs as backbone networks for the Faster R-CNN algorithm in order to process RGB and depth images. The authors then used the depth information to predict object boundaries, while they performed late fusion of the RGB and depth modalities to accurately classify the detected objects.

The huge variety of objects and/or industrial parts that need to be detected in a typical industrial environment, as well as the lack of large industrial object datasets makes the need for generalized object detection algorithms imperative. Generative models, such as Variational Autoencoders (VAEs) [5, 14] and Generative Adversarial Networks (GANs) [11], demonstrate tremendous learning capacity and generalization capabilities and are widely employed in several computer vision tasks. In object detection and classification, Eslami et al. [8] proposed a recurrent neural network and a VAE framework to identify several objects in a scene by attending to a single object at a time. Crawford et al [3] improved the previous approach by replacing the recurrent with a convolutional neural network and proposed a novel algorithm that can achieve higher accuracy and generalize better in scenes containing several objects.

Although advances in CNNs and VAEs have led to improved accuracy and generalization in several computer vision tasks, robustness problems still pertain. To remedy these problems, attention layers have been proposed in the literature, empowering deep networks to attend to certain aspects of the input data, reducing their overall sensitivity to noise. Vaswani et al. [24] proposed the modelling of the interdependencies between spatial or temporal features using a Transformer network. On the other hand, Hu et al. [13] proposed a network architecture that performs channel-wise attention and explicitly models interdependencies between channel features. In a similar fashion, Wang et al. [19] designed a layer called Efficient Channel Attention (ECA), aiming to be lightweight and model cross-channel interactions, and they demonstrated its ability to significantly boost the performance of deep networks.

Leveraging the aforementioned advances in deep learning, this work proposes a novel object detection approach that can be applied in industrial environments. The proposed approach is based on the processing of RGB-D images using the novel Variational Faster R-CNN and the subsequent fusion of the RGB and depth information for the accurate localization of industrial objects. The incorporation of a VAE encoder-decoder network and an ECA attention layer to the proposed Variational Faster R-CNN algorithm leads to improved accuracy and robustness as verified by the experimental results.

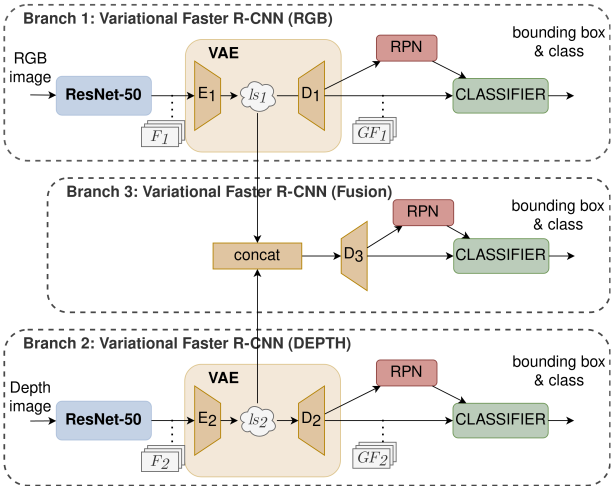


Figure 1: Network architecture of the proposed method.

3. Proposed Method

3.1. Motivation

The motivation behind the proposed object detection method lies in the effort to overcome limitations of single data modalities, as well as design networks with high discrimination and generalization abilities for accurate and robust industrial object localization results. The network architecture of the proposed method consists of three distinct branches that are trained separately. Initially, Branches 1 and 2 employ the novel Variational Faster R-CNN algorithm in order to create meaningful feature representations of the RGB and depth information, respectively, through the encoding of the RGB and depth images as points in descriptive latent spaces. Subsequently, points from the latent spaces are concatenated and fed to the RPN network and the classifier of the Variational Faster R-CNN in Branch 3 in order to derive the final positions (i.e., bounding boxes) and classes of the objects depicted in the images. The proposed method is illustrated in Fig. 1, while the novel Variational Faster R-CNN algorithm is described in detail below.

3.2. Variational Faster R-CNN

The proposed Variational Faster R-CNN comprises an enhanced variant of the original Faster R-CNN, aiming to improve its accuracy, robustness and generalization ability through the use of a VAE framework and an attention mechanism. More specifically, the proposed Variational Faster R-CNN algorithm initially extracts descriptive features from input images using a backbone CNN network. Lee et al. in [16] performed a comparison of various CNN networks that can be employed as backbone networks for the Faster R-CNN algorithm and showed that ResNet-50 outperforms the other networks in terms of accuracy. The high accuracy and computational speed, as well as the low memory footprint of the ResNet-50 network are the main reasons behind the adoption of the specific network for the proposed Variational Faster R-CNN since an algorithm that is employed in industrial robotic applications should meet such criteria.

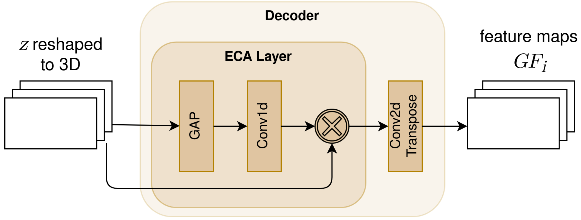


Figure 2: The architecture of the proposed VAE decoder.

Next, the extracted image features are fed to a VAE framework that is responsible for creating a mapping between the feature space and a new highly discriminative latent space. The purpose of this step is to improve the generalization ability of the proposed Variational Faster R-CNN algorithm by projecting the input data to a new space that can better model their underlying attributes and correlations, allowing the algorithm to generalize better on unseen data. Firstly, the encoder E_i gets as input the 3D feature maps F_i, reshapes them to vectors and processes them to generate two fixed-size vectors μ_i and σ_i that describe the mean and standard deviation of the distribution of the features in the latent space ls_i, respectively. These vectors are then used to sample a new vector z from the Gaussian distribution $N(\mu_i, \Sigma_i)$, where $\Sigma_i = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$ and d is the dimensionality of the latent space. The loss function that is used to optimize the weights of the encoder is Kullback-Leibler [14]:

$$L_{KL} = \frac{1}{2}\sum_{j=1}^{d}(\sigma^2(j) + \mu^2(j) - ln(\sigma^2(j)) - 1)$$

Afterwards, the latent space vector z is reshaped to a 3D tensor and fed to the VAE decoder. The decoder, illustrated in Fig. 2, consists of an ECA attention layer and a transpose convolution layer. The ECA layer (consisting of a global average pooling (GAP) and a convolution layer) is responsible for performing cross-channel attention by weighing feature maps accordingly based on their relevance for the task and diminishing the effect of noisy features, while the transpose convolution layer processes the feature maps to become more meaningful. The output of the decoder is the generated feature maps GF_i.

The feature maps GF_i generated by the VAE decoder are then fed to two specialized networks that are utilized by the original Faster R-CNN algorithm as well. The first network is called Region Proposal Network (RPN) and it consists of three convolutional layers that act as spatial sliding window, box-regression (reg) and box-classification (cls), respectively. For each sliding window, the RPN network provides object proposals (Regions of Interest (ROIs)), in the form of bounding boxes specified by their 2D coordinates and probabilities that the detected boxes depict an object or the background. A critical parameter in the RPN network is the designation of the anchors. The anchors represent a set of sizing and scaling parameters that may vary based on the size of input images and are used by the reg layer to define its object proposals per sliding window. Consequently, the improper setup of the anchors may lead to slow convergence of the RPN network or even to its failure. The loss that optimizes the RPN network is described by the following equation:

$$L_{RPN} = \frac{1}{N_{cls}} \sum_{i=1}^{n} L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_{i=1}^{n} p_i L_{reg}(t_i, t_i^*),$$

where L_{cls} is the box-classification layer loss and L_{reg} is the box-regression layer loss defined as binary cross-entropy and L1 loss, respectively. Moreover, p_i^* represents the ground-truth of the ROI class (0 or 1 depending on background or object), t_i^* represents the ground-truth of the bounding box, p_i is the predicted probability of the anchor class and t_i is the coordinates of the selected anchor. Finally, N_{cls}, N_{reg}, and λ are normalization and balancing parameters.

The second and final specialized network is the classifier that receives as input the generated feature maps GF_i and the ROIs generated by the RPN network and performs ROI pooling to refine the predicted bounding boxes prior to the prediction of the class for each final bounding box. The loss function utilized for the classifier network is defined as:

$$L_{classifier} = \sum_{i=1}^{n} L_{cls}(p_i, p_i^*) + \lambda \sum_{i=1}^{n} L_{reg}(t_i, t_i^*),$$

where L_{cls} denotes the cross-entropy loss and all other variables denote the same quantities as those presented in the equation of the L_{RPN} loss.

4. QCONPASS dataset

Motivated by the scarcity of publicly available datasets to effectively train object detection algorithms for industrial applications, we introduce in this work the QCONPASS dataset that consists of RGB-D images depicting industrial objects.

4.1. Data Collection

The QCONPASS dataset consists of synchronised and registered RGB and Depth images captured using an Intel RealSense D435 sensor placed on an autonomous Kobuki robot [21]. The robot navigates in an elevator manufacturing and assembly factory collecting images of elevator components that need to be recognised during assembly and packaging processes. The elevator components are placed firmly on a platform and moved around by workers so that the sensor can capture images of the components from different distances and viewing angles, thus increasing the variability of depicted objects and improving the robustness of an object detection algorithm trained on the QCONPASS dataset. The objects are placed at a distance of around 1.5-2m and the sensor could capture a maximum distance of 4m. Due to the aforementioned capturing procedure, each image of the QCONPASS dataset depicts a single elevator component. In addition, there are images, in which the depicted objects are slightly occluded due to the presence of workers that manipulate the platform.

4.2. Data Processing

During the preprocessing stage, the images of the dataset are initially filtered using two criteria: i) The object is viewed at an angle lower than 60 degrees and ii) At least 50% of the size of the object is visible on the image. Afterwards, a manual annotation procedure is performed using the VIA software [6], during which a human annotator view the object and annotate it with a class name and a bounding box. In total, 2051 RGB and depth images of size 848x400 pixels depicting 13 different elevator components (classes) are obtained, meaning that there are around 150 images per class. These images are randomly split in a training and test set consisting of 1661 and 390 images, respectively. All different elevator components present in the QCONPASS dataset are shown in Fig. 3, in which we can observe that most classes have similar shapes and sizes but they differ in texture.

5. Experimental Results

5.1. Datasets and Metrics

For the experimental evaluation of the proposed method, we employ the RGB-D Washington dataset [15], as well as the new QCONPASS dataset. The RGB-D Washington dataset contains around 210k RGB-D images, taken from different viewpoints. The images are subsampled every 5th video frame, resulting in around 42k RGB-D images of 300 everyday objects that are organized into 51 categories. For our experiments, we randomly sample one object per category to be employed for testing (i.e., 51 objects), while the remaining objects are used for training (i.e., 249 objects). The RGB-D dataset is chosen due to its large size that makes it suitable for training a deep network and the fact that it provices RGB-D images of objects. The RGB-D Washington dataset is very challenging as there is large vari-

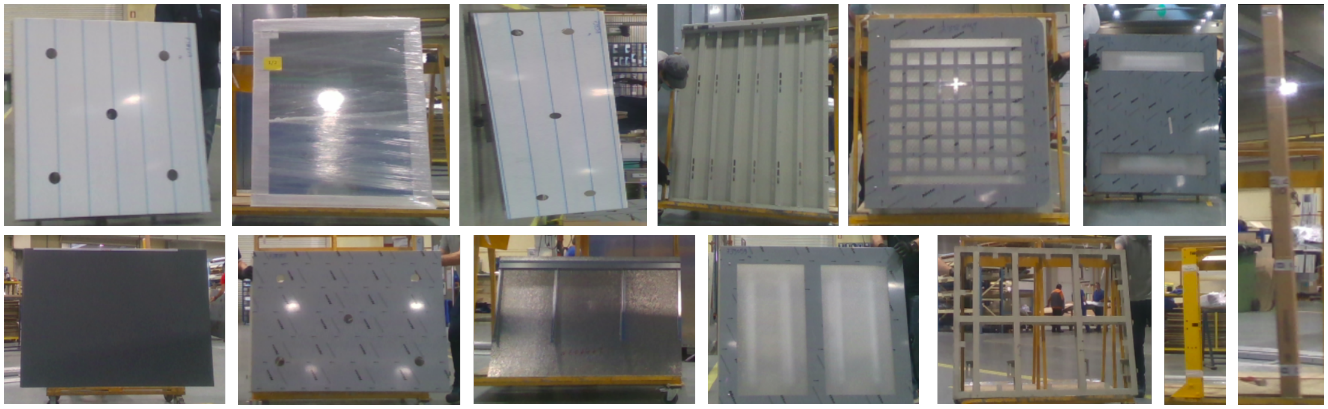


Figure 3: Different elevator components present in the QCONPASS dataset.

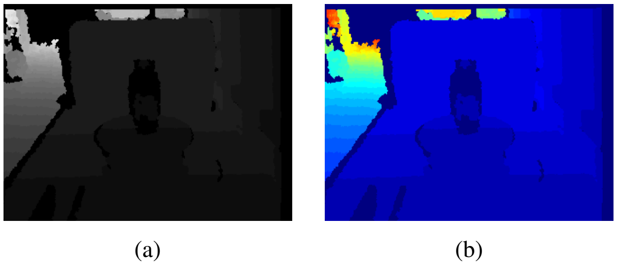
ance in the lighting conditions that affect the quality of the images. Additionally, the dataset contains several textureless objects, such as fruits and bowls, that can significantly deteriorate the accuracy of an object detection methodology.

To evaluate the performance of the proposed method and compare it with other techniques, we employ the well-known metrics of mean average precision (mAP), F1-score and Area Under Curve (AUC). To calculate AUC, we use the 11-point interpolated recall and precision curves. An object is considered detected if the ratio of the overlap between the predicted and the ground truth bounding box (i.e., Intersection-over-Union (IOU)) is above a threshold t_{IOU}. IOU is calculated using the formula below:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

5.2. Implementation details

In all Faster R-CNN implementations, we rescale the RGB-D images to the size of 320×240 and employ ResNet-50 as the backbone CNN network for fair comparison. More specifically, we get the output from the fourth convolutional layer of the ResNet-50 network, since the RPN network accepts 3D feature maps. The ResNet-50 network is already pretrained on ImageNet. We also encode the depth images from a single channel to 3 channels in colorjet format using the method proposed in [7], as it showed improved performance on the tested datasets. An example of the resulting depth images is illustrated in Fig. 4. Moreover, we set the dimensionality of the latent space for all branches to $d = 2240$. Finally, we set the number of object proposals generated by the RPN network to 16, we employed 5 scales with sizes of 8, 16, 32, 64 and 128, and 3 aspect ratios of 1:1, 1:2 and 2:1 for the anchors. The result is 15 anchors per sliding window (5 scales × 3 aspect ratios). We choose this configuration setup because the RGB-D Washington dataset contains objects with significant variations in size and image area coverage.



(a)

(b)

Figure 4: (a) Normalized depth image, (b) Colorjet depth image.

As far as the training is concerned, we initially train Branches 1 and 2 and then freeze their weights while training Branch 3. We also set the learning rate to the value of 10^{-5} and the threshold to decide whether a predicted bounding box is true positive or false alarm to $t_{IOU} = 0.5$.

5.3. Results

We evaluate our proposed method with different modalities in order to assess the importance of combining information from different modalities to achieve more accurate object detection results. Initially, we compare our proposed method against the original Faster R-CNN method on the RGB-D Washington dataset and the results are presented in Table 1.

From the results of Table 1, we can draw a few interesting conclusions. The RGB images carry slightly more important information than the depth images, which can be verified by the improved performance of both the original and the proposed Variational Faster R-CNN when RGB images are employed. In addition, utilizing multi-modal information is beneficial both to the original and the Variational Faster R-CNN, which verifies the importance of fusing dif-

Method	mAP	F1 Score	AUC
Faster R-CNN (RGB) [20]	50.93	59.23	0.51
Faster R-CNN (Depth) [20]	50.71	57.77	0.50
Multi-modal Faster R-CNN (VGG-16) [17]	55.9	62.24	0.55
Multi-modal Faster R-CNN (ResNet-50) [17]	59.8	66.51	0.59
Variational Faster R-CNN (RGB)	56.46	63.6	0.56
Variational Faster R-CNN (Depth)	52.6	61.72	0.52
Proposed approach	**64.3**	**70.67**	**0.63**

Table 1: Experimental results in the RGB-D Washington dataset.

ferent modalities to overcome limitations of single modalities. It should be noted here that the implementation of the multi-modal Faster R-CNN follows the one presented in [17] (i.e., concatenation of the RGB and depth features prior to introducing them to the RPN network and the classifier of the Faster R-CNN) with the use of VGG-16 (originally presented in [17]) and ResNet-50 as backbone networks for fair comparison with the other tested Faster R-CNN implementations. From the comparison of the results in Table 1, it can be seen that our proposed approach outperforms both multi-modal Faster R-CNN implementations of [17], irrespective of the backbone network used, thus verifying the importance of employing a VAE framework and the ECA attention layer for improved object detection results.

In addition, a comparison between the original and the Variational Faster R-CNN algorithm shows that the proposed approach achieves more accurate and robust predictions when utilizing either a single modality or multi-modal information. More specifically, the proposed approach improves the object detection results by 5.53%, 1.89% and 4.5%, in terms of mAP, with respect to the original Faster R-CNN, when RGB, depth or RGB-D images are employed, respectively. Similar performance gains are observed for the other metrics as well. These results verify that using the proposed VAE network architecture and the ECA layer can significantly improve the generalization ability, robustness and accuracy of an object detection algorithm. To further illustrate the performance improvement of our proposed method, we present the Precision-Recall curves in Fig. 5.

Finally, we evaluate our proposed method in the new QCONPASS dataset in order to assess the ability of the method to accurately identify objects in an industrial setting. The results are presented in Table 2 and show that the proposed approach can achieve really accurate results, despite the fact that the QCONPASS dataset is smaller compared to the RGB-D Washington dataset and it does not allow a deep network to optimally tune its weights. Object detection results from the proposed method for different input modalities (i.e., RGB, Depth and RGB-D) are shown in Table 2 and are illustrated in Fig. 6 to verify the importance of fusing information from different modalities rather than employing a single modality (i.e., at least 3.24% improvement in mAP when RGB-D images are employed). More specifically, using multi-modal information can remove inaccuracies (e.g., false positive predictions and no predictions at all, as well as inaccurate object delineation) produced when employing a single modality.

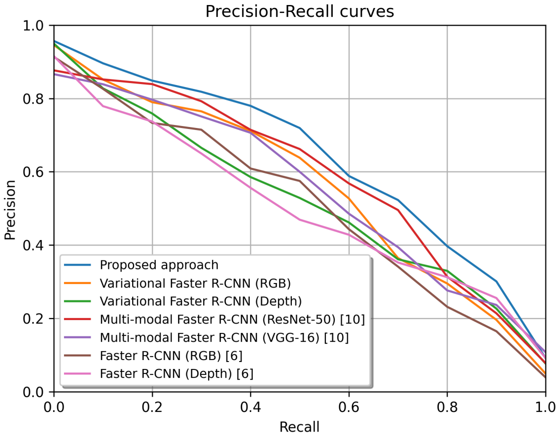


Figure 5: Precision-Recall curves for the compared methods.

Method	mAP	F1 Score	AUC
Variational Faster R-CNN RGB	89.92	87.6	0.85
Variational Faster R-CNN Depth	87.67	85.52	0.86
Proposed approach	**93.16**	**93.44**	**0.88**

Table 2: Experimental results in the QCONPASS dataset.

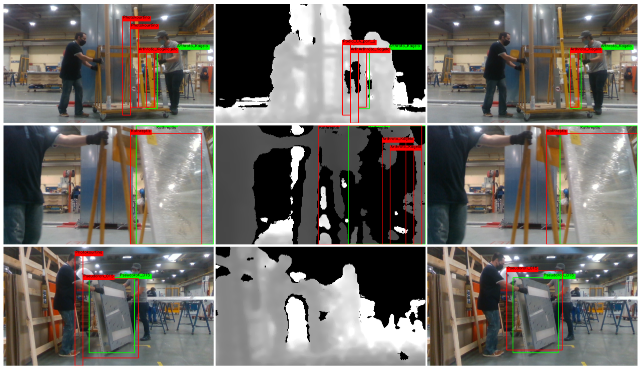


Figure 6: Object detection results from the proposed method in the QCONPASS dataset using RGB (left), Depth (middle) and RGB-D (right) information. Green and red boxes denote the ground truth and predicted results, respectively.

6. Conclusions

This work proposes a novel method for the identification of industrial objects based on the processing of RGB-D images. Inspired by the accuracy and low memory footprint of Faster R-CNN, we propose the Variational Faster R-CNN that utilizes a VAE framework to map the initial feature space into a new latent space, as well as the ECA attention layer to empower the network to diminish the influence of irrelevant features and enhance the impact of relevant ones. The goal is to develop an enhanced variant of the original Faster R-CNN algorithm that can achieve improved accuracy, robustness and generalization. Experimental results in a large publicly available object detection dataset verify the effectiveness of the proposed method. Moreover, motivated by the lack of RGB-D datasets for the identification of industrial objects, this study introduces the QCONPASS dataset that contains RGB-D images of elevator components and evaluates the proposed method on it.

As far as future work is concerned, we aim to investigate additional and more sophisticated ways to fuse multi-modal information using variational encoder-decoder network architectures. In addition, the proposed method could also be extended to integrate more than two modalities in order to deliver even higher accuracy and robustness in the object detection task, especially for industrial applications, where speed and accuracy are of utmost importance. Finally, we aim to enrich our QCONPASS dataset with more object instances and classes and make it publicly available in order to assist future industrial object detection methodologies.

Acknowledgment

This work has been supported from General Secretariat for Research and Technology under Grant agreement no. T6YBΠ-00238 "Q-CONPASS: Dynamic Quality CONtrol on Production lines using intelligent AutonomouS vehicleS".

References

[1] Pengchang Chen and Vinayak Elangovan. Object sorting using faster r-cnn. *arXiv preprint arXiv:2012.14840*, 2020.

[2] Xi Chen and Jan Guhl. Industrial robot control with object recognition based on deep learning. *Procedia CIRP*, 76:149–154, 2018.

[3] Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3412–3420, 2019.

[4] A. Dimou, D. Ataloglou, K. Dimitropoulos, F. Alvarez, and P. Daras. Lds-inspired residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2363–2375, 2019.

[5] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[6] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/, 2016. Version: X.Y.Z, Accessed: 02 April 2021.

[7] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE, 2015.

[8] SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E Hinton. Attend, infer, repeat: Fast scene understanding with generative models. *arXiv preprint arXiv:1603.08575*, 2016.

[9] Rinat Galin and Roman Meshcheryakov. Automation and robotics in the context of industry 4.0: the shift to collaborative robots. In *IOP Conference Series: Materials Science and Engineering*, volume 537, page 032073. IOP Publishing, 2019.

[10] Mingliang Gao, Jun Jiang, Guofeng Zou, Vijay John, and Zheng Liu. Rgb-d-based object recognition using multimodal convolutional neural networks: A survey. *IEEE Access*, 7:43110–43136, 2019.

[11] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *2011 IEEE international conference on robotics and automation*, pages 1817–1824. IEEE, 2011.

[16] Chungkeun Lee, H Jin Kim, and Kyeong Won Oh. Comparison of faster r-cnn models for object detection. In *2016 16th International Conference on Control, Automation and Systems (ICCAS)*, pages 107–110. IEEE, 2016.

[17] Irina Mocanu and Cosmin Clapon. Multimodal convolutional neural network for object detection using rgb-d images. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pages 1–5. IEEE, 2018.

[18] Christian Poss, Olim Ibragimov, Anoshan Indreswaran, Nils Gutsche, Thomas Irrenhauser, Marco Prueglmeier, and Daniel Goehring. Application of open source deep neural networks for object detection in industrial environments. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 231–236. IEEE, 2018.

[19] Wang Qilong, Wu Banggu, Zhu Pengfei, Li Peihua, Zuo Wangmeng, and Hu Qinghua. Eca-net: Efficient channel attention for deep convolutional neural networks. 2020.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[21] Yujin Robot. Kobuki. `http://kobuki.yujinrobot.com/about2/`, 2011. [Online; Accessed: 02 April 2021].

[22] Faisal Saeed, Anand Paul, and Seungmin Rho. Faster r-cnn based fault detection in industrial images. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 280–287. Springer, 2020.

[23] Yiyou Sun, Tonghua Su, and Zhiying Tu. Faster r-cnn based autonomous navigation for vehicles in warehouse. In *2017 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 1639–1644. IEEE, 2017.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[25] Xunmu Zhu, Changxin Chen, Bin Zheng, Xiaofan Yang, Haiming Gan, Chan Zheng, Aqing Yang, Liang Mao, and Yueju Xue. Automatic recognition of lactating sow postures by refined two-stream rgb-d faster r-cnn. *Biosystems Engineering*, 189:116–132, 2020.