Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

Article

Survey and Performance Analysis of Object Detection in Challenging Environments

Muhammad Ahmed [1,2,†], Khurram Azeem Hashmi [1,2,3,*,†], Alain Pagani [3], Marcus Liwicki [4], Didier Stricker [1,3], and Muhammad Zeshan Afzal [1,2,3,*]

1 Department of Computer Science, Technical University, 67663 Kaiserslautern, Germany; agk1248@gmail.com (M.A.); khurram_azeem.hashmi@dfki.de (K.A.H.); didier.stricker@dfki.de (D.S.);
2 Mindgrage, Technical University, 67663 Kaiserslautern, Germany
3 German Research Institute for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany alain.pagani@dfki.de (A.P.); muhammad_zeshan.afzal@dfki.de (M.Z.A.);
4 Luleå University of Technology, 971 87 Luleå, Sweden; marcus.liwicki@ltu.se (M.L.);
* Correspondence: khurram_azeem.hashmi@dfki.de and muhammad_zeshan.afzal@dfki.de
† These authors contributed equally to this work.

Abstract: Recent progress in deep learning has led to accurate and efficient generic object detection networks. Training of highly reliable models depends on large datasets with highly textured and rich images. However, in real-world scenarios, the performance of the generic object detection system decreases when (i) occlusions hide the objects, (ii) objects are present in low-light images, or (iii) they are merged with background information. In this paper, we refer to all these situations as challenging environments. With the recent rapid development in generic object detection algorithms, notable progress has been observed in the field of object detection in challenging environments. However, there is no consolidated reference to cover state-of-the-art in this domain. To the best of our knowledge, this paper presents the first comprehensive overview, covering recent approaches that have tackled the problem of object detection in challenging environments. Furthermore, we present the quantitative and qualitative performance analysis of these approaches and discuss the currently available challenging datasets. Moreover, this paper investigates the performance of current state-of-the-art generic object detection algorithms by benchmarking results on the three well-known challenging datasets. Finally, we highlight several current shortcomings and outline future directions.

Keywords: Object detection; challenging environments; low-light; image enhancement; complex environments; state-of-the-art; deep neural networks; computer vision; performance analysis.

check for updates

Citation: Ahmed, M.; Hashmi, K.A.; Pagani, A.; Liwicki, M ; Stricker, D., ;Afzal, M.Z. Survey and Performance Analysis of Object Detection in Challenging Environments. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Object detection is considered as one of the most important and elementary tasks in the field of computer vision. The problem of object detection deals with the identification and spatial localization of objects present in an image or a video [1]. The task of object detection covers wide-range of many other computer vision tasks, such as instance segmentation [2–4], visual question answering [5], image captioning [6,7], object tracking [8], activity recognition [9–11] and so on.

Object detection is a complex problem because of underlying high intra-class and low inter-class variance [12,13]. High intra-class variance is the consequence of different objects belonging to a single class, for instance, different poses of humans or humans wearing different clothes in an image. Low inter-class variance is the outcome of similar-looking objects belonging to different classes such as samples of class *chair* can easily be misclassified into the class *bench* and vice-versa.

One of the earlier approaches for object detection algorithms relied on sliding windows, applying classification on each window to find objects [14–16]. Later sliding window concept was replaced with region proposals to narrow the search before applying classification [17–21]. The recent surge in deep learning has given fresh blood for object detection systems along with other fields.

The prior published work in object detection can be further classified into three categories which are explained below. Figure 1 depicts the basic difference between them:

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

1. *Object Detection (OD)*: OD aims at detecting objects regardless of their class category [22,23]. OD algorithms [24–27] generally propose a large number of possible region proposals, from which later on, the best possible candidates are selected according to certain criteria.
2. *Salient Object Detection (SOD)*: SOD algorithms use the human attention mechanism concept to highlight and detect the objects in a picture, or video [28,29].
3. *Category-specific Object Detection (COD)* : COD aims at detecting multiple objects. Unlike OD and SOD, COD has to predict the category class and the location of the object in the image or video [30,31].

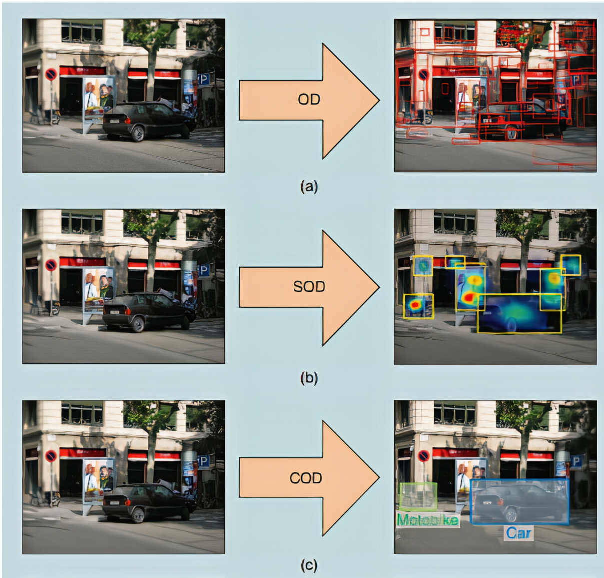

Figure 1. Visual demonstration of the possible three directions in the domain of object detection. Object detection (OD): objects are detected, Salient Object Detection (SOD): human attention mechanism is applied to detect objects, Category-specific object detection (COD): Object detection along with the class prediction. Image is courtesy of [32].

The deep learning-based object detection algorithms are categorized into two-stage object detectors and one-stage object detectors. Two-stage object detection architectures like R-CNN[33], Fast R-CNN[34] and Faster R-CNN[24] segregate the task of object localization from object classification task. They employ region proposal techniques to find possible regions where the likelihood of an object's existence is maximum. Later segmentation output and better detection pooling [35] techniques were introduced with Mask R-CNN [36]. On the other hand, one-stage object detection algorithms first generate candidate regions, and then these regions are classified as object/no-object. For instance, one-stage detectors like YOLO [37–40] and SSD [27] work with feature pyramid networks (FPNs) [41] as a backbone to detect objects at multiple scales in a single pass rather than first predicting regions and then classifying them.

Analogous to other domains of computer vision, the output of object detection is highly dependent on spatial features. Hence, the performance of the object detection system decreases when the occlusions hide the objects, objects have different sizes or are merged with background information. In real-time scenarios, it is frequent that the input images received by the object detection network are not rich with information, or they are captured under low-light conditions. This paper has referred to all these situations in a challenging environment. Figure 2 illustrates the problem of object detection in a challenging environment. This survey paper investigates the approaches that have worked on detecting objects in such environments by leveraging deep neural networks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

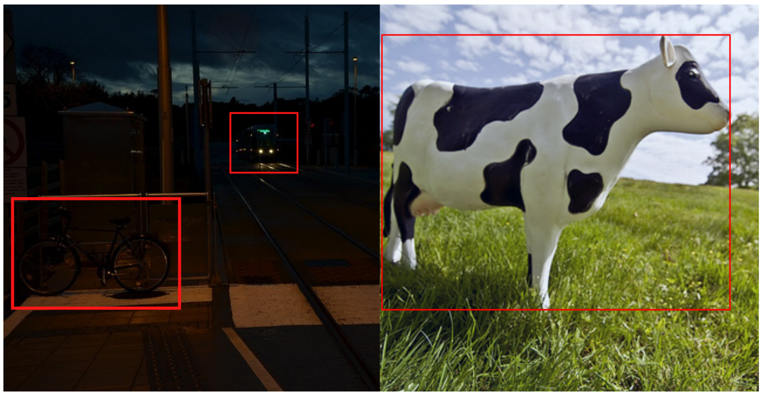

Figure 2. Graphical illustration of object detection in a challenging environment (left image), whereas object detection in a conventional environment is shown in the right image. There are two objects: bicycle and bus are present in the challenging left image while the cow is a targetted object in the right image.

Different pre-processing techniques have been applied to improve image quality to improve object detection performance in challenging environments over the years. Figure 3 illustrates the comparison between the flow of traditional approaches and deep learning-based methods. Traditional methods rely on improving image quality through image enhancement and manual feature selection methods [1,14,42,43]. Later these techniques are replaced with Deep Neural Networks (DNNs) because of their robust and generalization capabilities.

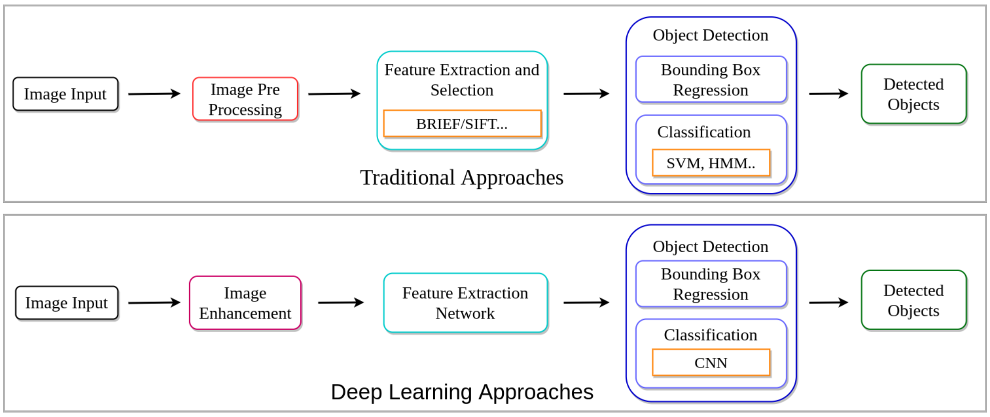

Figure 3. Pipeline comparison of traditional and deep learning approaches for object detection. In traditional approaches, generally, image enhancement is applied before feature extraction to improve object detection performance. Unlike traditional approaches, deep learning methods can find required features for detecting objects without relying on traditional rule-based methods.

1.1. Contributions

This survey paper focuses on the recent works that have tackled object detection in a challenging environment by employing deep neural networks. With the modern advancement of deep learning-based object detection approaches, a noticeable surge has been observed in the field of object detection under a complex environment, as illustrated in Figure 4.

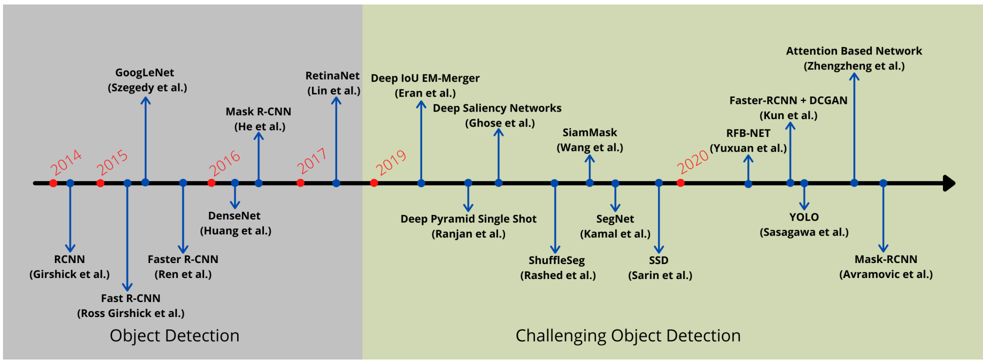

Figure 4. A timeline view of different deep learning approaches developed recently to improve object detection in challenging environments.

To summarize, our contributions are as follows:

1. We present a unified framework that explains object detection in a challenging environment.
2. We give an overview of all the publicly available datasets that have been published to detect objects in challenging scenarios.
3. We summarize the advantages and limitations of the discussed methods in order to improve object detection in a challenging environment.
4. We benchmark current state-of-the-art generic object detection algorithms on the three well-known challenging datasets.

The remainder of the paper is organized as follows: Section 2 briefly discusses previous surveys conducted on object detection in challenging environments. Section 3 provides an exhaustive discussion on recently applied approaches to solving problems of object detection in difficult environments. Figure 5 explains the structural flow of mentioned methodologies. Section 4 provides details about publicly available datasets for challenging object detection tasks. Section 6 explains well known evaluation metrics and provides performance of all the discussed approaches in Section 3. Section 7 highlights the current challenges and suggest future directions, whereas Section 8 concludes the paper.

2. Related Surveys on Object Detection

There are many surveys carried out on the topic of object detection [44–47]. This section covers some of the prior surveys.

Han et al. [32] organized the survey in which deep learning techniques for salient and category-specific object detection are reviewed. In 2019, *Zou et al.*[48] performed an extensive survey on object detection methods that have been proposed in the last 20 years. The authors discussed all the types of object detection algorithms proposed over the years and highlighted their improvements.

Another survey organized by *Jiao et al.* [49] discussed various deep learning-based methods for object detection. The proposed work provided a comprehensive overview of traditional and modern applications of object detection. Moreover, The authors discussed methods for building better and efficient object detection methods by exploiting existing architectures. *Arnold et al.* [50] surveyed 3D object detection methods for autonomous driving. The proposed work compared various 3D object detection-based approaches.

It is vital to mention that all of the prior surveys have focused on the general problem of object detection. Although these surveys explain how object detection has improved over the years, they do not cover the challenges and solutions to improve object detection performance in a challenging environment such as low light, occlusions, hidden objects, and so on. To the best of our knowledge, we provide the first survey that reviews the performance of deep learning-based approaches in the field of object detection in a challenging environment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

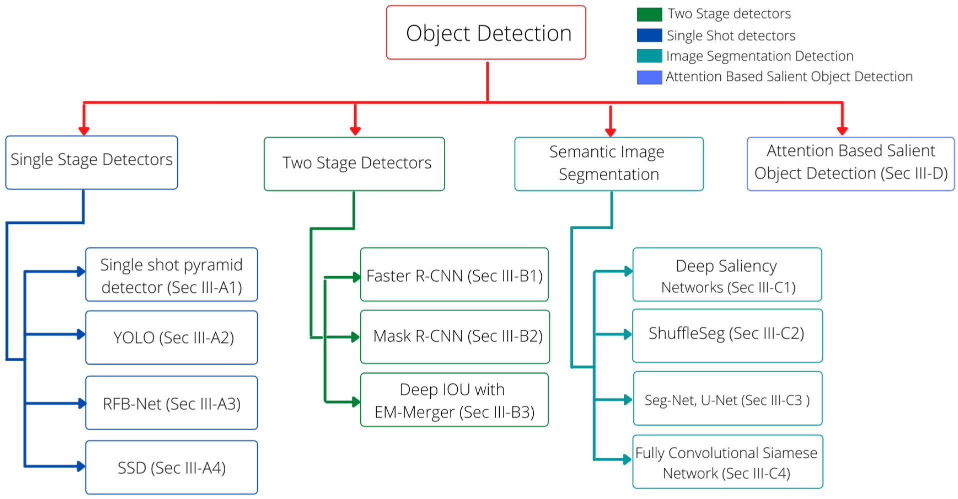

Figure 5. Organization of explained methodologies in the paper. The problem of object detection in challenging environments has been tackled by employing various deep learning concepts. In this paper, we categorize these methods according to the utilized architectures.

3. METHODOLOGIES

The process of detecting an object is divided into the following steps:

1. Find regions as object/no-object.
2. Classify the detected regions where objects exist.

This section discusses the various approaches that have tackled object detection problems under complex scenarios by employing deep neural networks. As explained earlier in Section 1, Various object detection-based algorithms and a few semantic segmentation-based approaches have been exploited in the recent past to develop object detection systems under challenging situations. For convenience, we have categorized the methodologies according to the employed architectures. Figure 5 exhibits the structure of this section, whereas the categories of the explained methods are illustrated in Figure 6. Furthermore, Table 1 summarizes the main advantages and limitations of the mentioned approaches. Moreover, Figure 7 exhibits the generic mechanism of the recently proposed approaches in this domain.

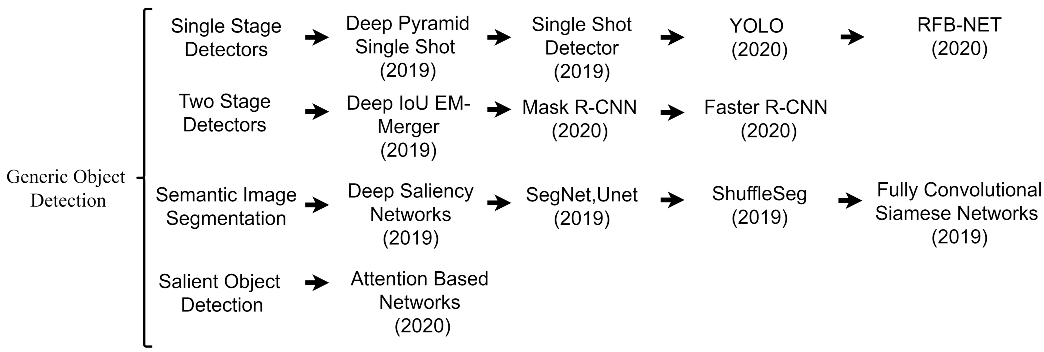

Figure 6. Different object detection algorithms approaches are categorized into four categories. One-stage detectors are fast but lack accuracy, whereas two-stage detectors are slow but accurate. Semantic image segmentation deals with pixel-level detection, and salient object detection deals with detecting objects by applying an attention mechanism.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

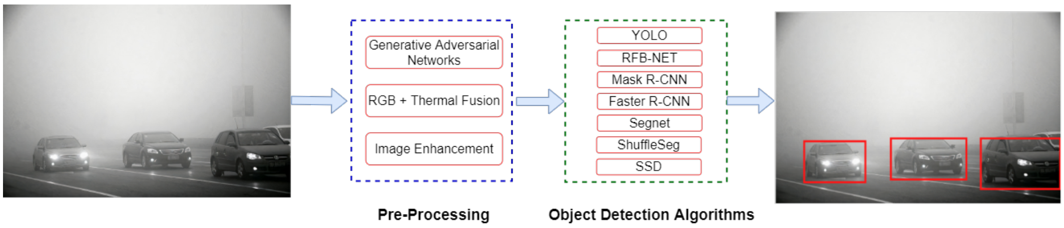

Figure 7. The generic mechanism of the proposed approaches that have tackled the problem of object detection in a challenging environment. Image-enhancement networks extract important features from the input images. Object detectors employ these features to retrieve final predictions.

3.1. One-Stage Detectors

One-stage detectors predict the candidate regions without taking aid from region proposal networks. Compared to two-stage detectors [24,26], they are efficient and widely exploited in developing real-time object detection systems [51–53]. The architectural overview of one-stage detectors is depicted in Figure 8. Researchers have proposed approaches to detect objects in arduous environments by adopting one-stage detectors. This section further arranges the approaches according to the operated one-stage object detectors.

3.1.1. YOLO

You Only Look Once (YOLO) [25] has been widely applied to develop real-time object detection systems in various domains [51–53]. Unlike region-based convolutional neural networks, YOLO only takes a single look and divides the image into $S \times S$ grid of cells. For each object that exists on the image, one grid cell is supposed to be responsible for having that object at its center. Figure 9 illustrates the main idea of YOLO. This section covers the techniques that have exploited YOLO to detect objects in a complex environment.

By leveraging the power of transfer learning, *Sasagawa et al.* [54] proposed an approach to detect objects under low illumination. The work proposed combining two trained models from different but related domains through glue layers and a generative model. The authors have used the knowledge distillation technique to train the proposed approach. Initially, encoder-decoder network [55] is used to extract spatial features from the given image by using convolutional [56] and pooling layers [57]. The pooling layer is used to capture features at different frequencies of information. Once the encoder-decoder network learns the latent representation, it is propagated to the glue layer. The glue layer is composed of batch normalization [58], pooling, and concatenation. The authors empirically established that the concatenation of all latent features produces the optimal result. The glue layer uses RGB data generated by encoder g_2 using the knowledge distillation method. YOLO [25] is utilized to localize and identify objects. Since the proposed technique relies on pre-trained networks, the backbone network is pre-trained on MS-COCO [12] and See in the dark(SID) [59] dataset. By combining pre-trained models using glue layers, the proposed method could detect objects in scenes illuminated by less than 1 lux. Further, it is 2.4 times more sensitive than the original YOLO model [25]. Moreover, the use of the glue layer further reduces the computational resources.

Utilizing YOLO, *Mate et al.* [60] employed thermal images to improve object detection performance in challenging conditions such as bad weather, nighttime, and densely packed areas. The authors argued that thermal images could easily show the difference between objects and the surrounding environment based on temperature values as compared to RGB images. This aids the process of detecting objects in difficult situations like low light scenarios, rainy or foggy weather. This work consists of two different experiments. The first network was fine-tuned on the thermal dataset UNIRI-TID [61] whereas the second model was trained from scratch on the same UNIRI-TID thermal dataset. The authors found that the model trained from scratch performed better than the fine-tuned model. The authors also reported that the model's performance drops in the case of bird-eye view images. Figure 17 shows a failure case of proposed method.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

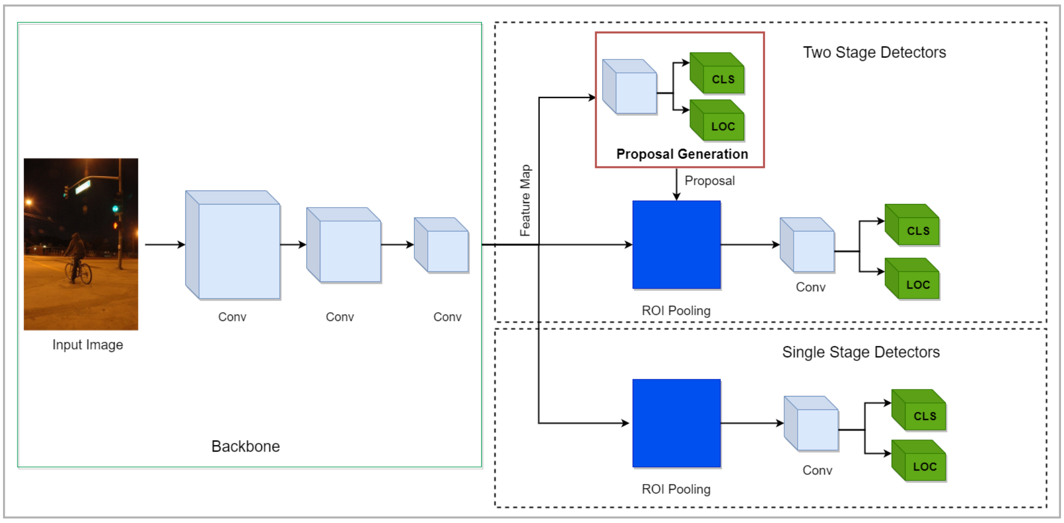

Figure 8. Graphical illustration of the object detection algorithms. The upper part of a figure exhibits two-stage detectors in which the backbone network extracts features that are passed to RPN to generate region proposals. These region proposals are refined to detect the objects in images. The lower part illustrates one-stage detectors in which objects are localized and classified in a single forward pass. The grey cubes in the backbone demonstrate the convolutional operation.

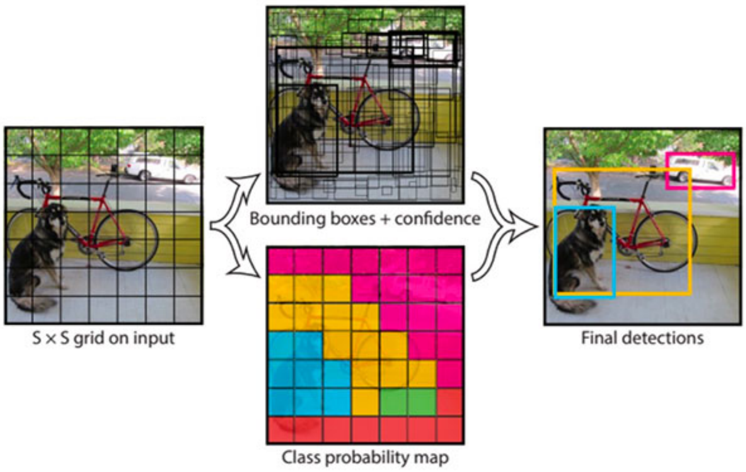

Figure 9. The explained architecture of YOLO [25]. YOLO sees the complete image at once and segments an image into S $\times$ S grids. Each cell predicts bounding boxes and a confidence score for every bounding box. The cell in which object is present at the center, is liable to detect the object. Figure is from [25].

3.1.2. RFB-Net

Currently, existing state-of-the-art detection algorithms use pre-trained backbones based on Feature Pyramid Network (FPN) to extract low-level and high-level features. Instead of relying on a pre-trained backbone, RFB-Net is inspired by the Receptive Fields (RFs) structure in the human visual system. RFB-Net considers the relationship between size and eccentricity of RFs to enhance features. Figure 10 illustrates the main building block of RFB-Net.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

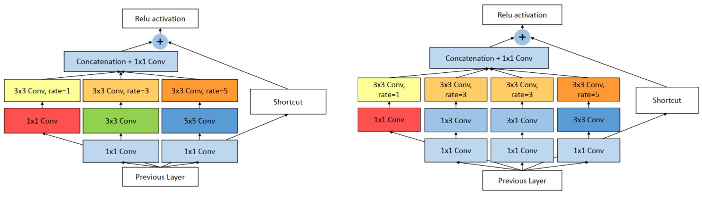

Figure 10. The explained architecture of RFB [62]. RFB is inspired by Receptive Fields (RFs) in human visual system. RFB considers the relationship between the size and eccentricity of Receptive fields into account to enhance the features. (a) RFB is identical to Inception, responsible for simulating the RFs of multiple sizes, and (b) part reproduces the relation between the RF size and eccentricity in the human visual system by using smaller kernels. Figure is from [62].

Existing state-of-the-art generic object detection algorithms do not perform well on images captured at low light because most low-level features are merged with the background. To tackle this problem, *Yuxuan et al.* [63] fused contextual information in the backbone to avoid loss of low-level contextual features. Lower spatial feature maps with higher spatial feature maps are fused to preserve low-level features, maximizing pre-trained channel information.

The second problem with images captured in low light is that when these images are passed through conventional hierarchical convolutions, the resulting edges and features are not very rich with information. Therefore context fusion is incorporated in the backbone part of the network. The aim is to compensate for the information loss during the lower to higher-level convolution process. Low-level feature maps of the network are selected and interpolated to the identical shape of its successive feature maps. The resulting feature map is rich in contextual information combining high-level features and low-level features. The authors merged RFB-Net [62] as a base architecture with the proposed modifications.

3.1.3. SSD

SSD [27] is another extensively employed real-time object detection algorithm belonging to the family of one-stage detectors. SSD eliminates the region proposal network from the Faster R-CNN [24] to increase the network's speed. SSD includes multi-scale features and default boxes to compensate for the performance trade-off. The architecture of SSD is illustrated in Figure 11.

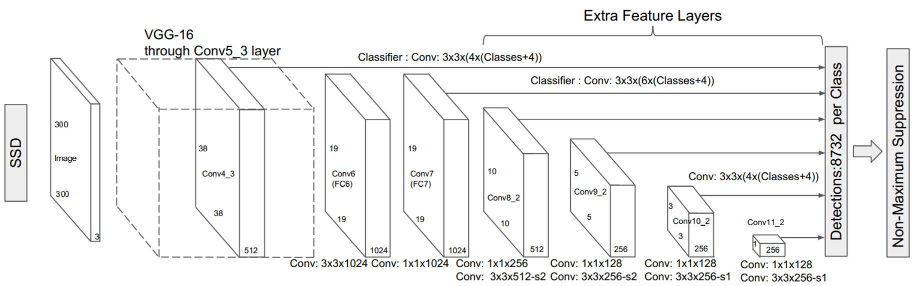

Figure 11. The explained architecture of SSD [27]. The architecture of SSD is based on the VGG-16 [64] as base network by replacing VGG fully connected layers with auxiliary convolutional layers to provide feature extraction at multiple scales. Later, the multibox detector [65] performs the task of object detection. Image is from [27].

Sarin et al. [66] proposed a convolutional neural network-based human detection and SSD-based face detection system to improve the results of surveillance systems in challenging environments. The first human detection network is used to identify whether an input image contains a human. Once a human is detected, the method exploits SSD

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

to localize the area of the face on a predicted image. Since the conventional method tends to over-fit on the provided dataset, the proposed work leverages dropout [67] and data augmentation techniques to avoid overfitting. For the enhancement of dark images, brightness normalization with the zero-center method is adopted. The model operated on AlexNet [56] and was modified for the problem at hand. Two datasets KAIST [68] and UFDD [69] datasets, were used to train human and face detection networks, respectively.

Deep Pyramid Single Shot Face Detector

For face detection in a harsh environment, *Ranjan et al.* [70] proposed a Deep Pyramid Single Shot face Detector (DPSSD). The authors argued that deep pyramidal architecture in SSD [27] can even detect small-scale faces in the images. In this work, the input image is passed through up-sampling layers to extract rich contextual features and then passed through pooling layers. Final detection boxes are generated from the up-sampling layers through the anchor box matching technique. The output prediction divides into two branches: regression and classification. Furthermore, to optimize network convergence, crystal loss is introduced for face verification and classification tasks. The proposed crystal loss minimizes the angular distance between similar subject pairs and maximizing the angular distance between different subject pairs. For the training purpose, the WIDER Face dataset [71] was employed. The proposed approach is extensively evaluated on unconstrained face detection datasets of IJB-A [72], IJB-B [73], and IJB-C [74].

3.2. Two Stage Detectors

Along with one-stage detectors, two-stage detectors have been applied to tackle object detection in difficult situations. Figure 8 depicts the basic architecture of two-stage detectors.

3.2.1. Faster R-CNN

Faster R-CNN [24] is one of the most widely exploited object detection networks [83,84]. It is the extension of Fast R-CNN [34] with Region Proposal Network (RPN). RPN aims to generate proposals with different sizes and aspect ratios. These proposals and the spatial features are propagated to the Fast R-CNN [34] module, which performs object detection. The basic architecture of Faster R-CNN [24] is depicted in Figure 12. This section discusses the methodologies that employ Faster R-CNN [24] as an object detection algorithm in their system of identifying objects under various difficult scenarios.

Generative Adversarial Networks(GAN) [85] have also been applied to enhance the visual quality of images captured under low light. *Kun et al.* [77] proposed the Deep Convolution Generative Adversarial Networks(DCGAN) [86] combined with Faster R-CNN [24] to solve the problem of object detection in low light. The first step is to use DCGAN to convert nighttime scenes into daytime scenes by keeping the important aspects maintained between two input images. The second step is Multi-scale convolution feature fusion. This work applies up-sampling and down-sampling convolutions to fuse features extracted from starting layers.

The third contribution of this work is the proposed modification in the ROI Pooling layer. According to the coordinate vector, the standard ROI pooling layer uses bilinear interpolation to obtain pixel value. Although conventional ROI pooling reduces parameters, the target object's critical features are lost. Therefore, the authors employed ROI pooling of different sizes to capture more detailed information. Subsequently, the final ROI pooling output is passed to the classifier and regressor to obtain final predictions.

3.2.2. Mask R-CNN

In the family of region-based convolutional neural networks, Mask R-CNN [26] has been exploited by the state-of-the-art object detection and instance segmentation approaches [87,88]. Mask R-CNN [26] extends Faster R-CNN [24] with an addition of another branch that retrieves the mask for each detected object. The complete framework

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

Table 1. A summary of advantages and limitations of methods tackling object detection in challenging environments. The double horizontal line separates the approaches according to the categories explained in Section 3.

Literature	Methods	Advantages	Limitations
Yuxan et al. [63]	Images are transformed and then fed into the RFB-Net (Sec 3.1.2).	Context information fusion allows detection of object in low-light.	Relies on prior information about type of object, shape etc for detecting them in night time.
Mate et al. [60]	Thermal images incorporated with YOLO (Sec 3.1.1).	Thermal images give better information then RGB images in difficult conditions.	Fails at capturing from bird-eye view. Further, thermal images are not useful in environments where background and object temperature is same. As illustrated in Figure 17.
Ranjan et al. [70]	Single shot deep pyramid face detector (DPSSD) (Sec 3.1.3.1).	Extract rich contextual information and multi-scale features with help of pooling layers.	Requires a fixed input size.
Sasagawa et al. [54]	Fusion of pre-trained models using Glue layer and information distillation (Sec 3.1.1).	Domain joining with help of glue layer reduces in computation and provides more information for models to learn from different domains.	Relies on prior domain knowledge.
Sarin et al. [66]	Single shot human and face detector (Sec 3.1.3).	Simple and effective approach.	Fails in low light and gets fooled by human look alike pictures. As illustrated in Figure 18.
Avramovic et al. [75]	Region of interest(ROI) integration with Mask-RCNN and YOLO (Sec 3.2.2).	Only choose regions of interest where objects of interest can occur instead of applying object detection to whole image.	Relies on prior knowledge of where objects can show.
Eran et al. [76]	Modification of CNN with Soft-IOU layer and custom EM-Merger layer.	Reduces overlapping detections (Sec 3.2.3).	Treats overlapping predictions as clustering problem. Not very real run time.
Kun et al. [77]	Generative adversarial network with Faster R-CNN (Sec 3.2.1).	Networks learns both day and night time features.	Relies on prior information of converting night time images to day time.
Ghose et al. [78]	Fusion of thermal images and their saliency maps using deep saliency networks (Sec 3.3.1).	Provides rich contextual and depth information.	Relying on thermal images causes poor performance in day time or similar conditions.
Rashed et al. [79]	Fusion of RGB images with LiDAR sensors information with encoder-decoder architecture (Sec 3.3.2).	Fusion of RGB, rgbFLow and lidarFlow provides greater information.	LiDAR sensors data and rgbFlow data need to be merged. There are multiple methods of fusion.
Kamal et al. [80]	Combination of SegNet and U-Net to detect traffic signs (Sec 3.3.3).	Less computation cost as four corners of image are cropped and merged together before passing through network	Relies on prior information regarding where objects are most likely to occur. Fails when object location changes. As illustrated in Figure 19.
Wang et al. [81]	Fully convolutional Siamese networks with modified binary segmentation task (Sec 3.3.4).	Pre-frame binary segmentation mask is used for low-level object representation instead of relying on feature extractor backbone.	Relies on prior information while generating binary segmentation mask. Fails when faced with motion blur and non-object pattern. As illustrated in Figure 20.
Zhengzheng et al. [82]	Two Stream convolutional neural network with attention mechanism (Sec 3.4).	Fusion of rgb and thermal image to generate features and noise reduction with convolutional block attention module. A new dataset for benchmarking.	Thermal images are not efficient for every environment use.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

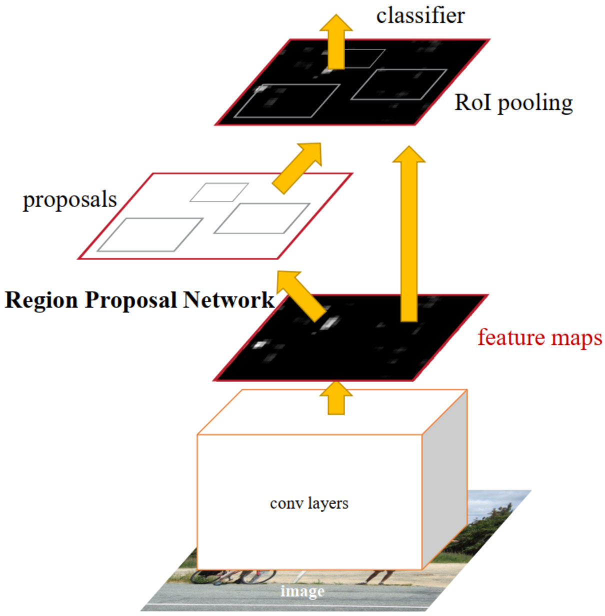

Figure 12. The explained architecture of Faster R-CNN [24]. The base network generates the feature maps from input images and propagates them to RPN and ROI pooling. The RPN generates region proposals which are refined by ROI pooling to perform regression and classification. Figure is from [24].

of Mask R-CNN is illustrated in Figure 13. This section talks about the approaches that have tackled object detection under strenuous conditions by exploiting the capabilities of Mask R-CNN.

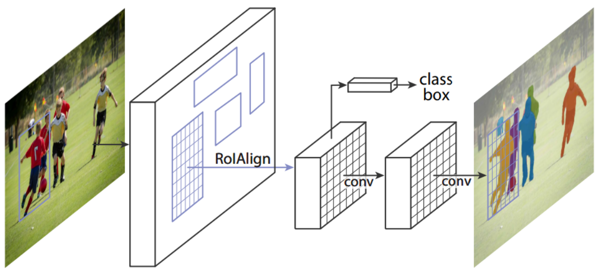

Figure 13. The explained architecture of Mask R-CNN [26]. The architecture of Mask R-CNN extends the architecture of Faster R-CNN [24] with two modifications: 1) Replacing the RoI pooling with RoIAlign layer. 2) Addition of segmentation branch along with regression and classification. The figure is from [26].

Another approach for detecting traffic signs is proposed by *Avramovic et al.* [75]. The authors discussed that a driver could only focus in front of him and beside him through (side mirrors) during driving. Hence, to detect traffic signs, the model should perceive an input image in the same way. Therefore, instead of applying object detection to the whole image, the proposed work focus on a specific region where traffic sign is likely to be seen. Selecting a limited amount of Regions Of Interests (ROIs) reduces the overall computation cost and propagates smaller images through the network efficiently. Furthermore, the original aspect ratio of traffic signs is preserved by using a small number of ROIs. They

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

evaluated the approach equipped with Mask-RCNN [26] and YOLO[25] on the DFG traffic sign dataset [89].

3.2.3. Deep IOU with EM-Merger

In a challenging environment, generic object detectors predict multiple bounding boxes for a single object. To address this issue, *Eran et al.* [76] proposed a Soft-IOU layer for estimating Jackkard distance as a quality detector between the predicted bounding box and the ground truth. The presented approach aims to minimize overlap regions between multiple detections. Once the network predicts the bounding boxes for the input image, the predictions are treated as a clustering problem. The method applies a custom EM-Merger layer to group similar predictions to form a single detection.

Although conventional non-max suppression can remove overlapping detection, the authors argued that non-max suppression does not provide a satisfactory result in densely packed situations. This work empirically established that it is better to use the EM-Merger layer to combine extra detections. Their proposed approach achieved better results than the state-of-the-art object detection-based algorithm on SKU-110K benchmark dataset [76].

3.3. Semantic Image Segmentation

Besides generic object detection algorithms, researchers have investigated the capabilities of Semantic Image Segmentation (SIS) [90] in identifying objects in arduous conditions. Instead of localizing an object in an image, SIS predicts a class label for each pixel in an image. This task is also termed a dense prediction. Once we have the labeled pixels, we segment the objects by grouping the pixels belonging to the same class. The methods based on SIS mainly operate on Fully Convolutional Networks (FCNs) [91]. Unlike other deep neural networks, FCNs consist of only convolutional layers and no fully connected layers. Therefore, FCNs can work with images of various sizes. A visual illustration of FCNs are exhibited in Figure 14. This section addresses some of the methods that operate on the concept of SIS to deal with object detection in perplexing scenarios.

3.3.1. Deep Saliency Networks

In 2019, *Ghose et al.* [78] proposed the combination of saliency maps with thermal images to detect pedestrians in poor lighting conditions. This work suggested that instead of relying on thermal and RGB images, the addition of saliency map information to thermal images facilitates the network to identify the objects. At first, the method augments the thermal images by replacing one duplicate channel with a corresponding saliency map. This combination helps illuminate salient parts of the image while preserving textural information. To extract saliency maps, the authors experimented with two state-of-the-art deep saliency networks (PiCA-Net [92], and R3Net [93]).

PiCA-Net generates an attention map for each pixel in an image and employs Long Short Term Memory (LSTM) [94] to scan the image and obtain global context. U-Net[95] is then applied to detect salient objects from the generated attention maps. R3Net [93] uses residual refinement block to learn residuals between the ground truth and saliency maps. To assess the performance of the proposed method, the authors exploited KAIST multi-spectral pedestrian dataset [68].

3.3.2. ShuffleSeg

Following the concept of applying image segmentation [78], *Rashed et al.* [79] proposed FCNs based architecture for moving object detection by concatenating spatial information with LiDAR sensors. The presented work employed three encoders to handle RGB, rgbFlow, and lidarFlow information separately. The whole network follows encoder-decoder architecture. The encoder extracts the latent features while the decoder up-samples them to generate the final image. To analyze the capabilities of the proposed approach, the authors used KITTI dataset [96] and a custom-made Dark-KITTI dataset to simulate a low-light

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

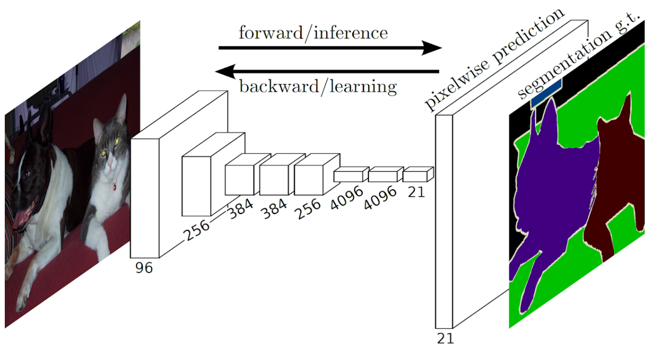

Figure 14. The explained architecture of Fully Convolutional Networks (FCNs). FCN first down-samples and then transforms the image back to the original spatial size using up-convolutions. Using dense predictions, they try to predict a class label for each pixel in an image and segment the objects based on the classes of pixels. The image is courtesy of [91].

environment. The proposed method achieved a 4.25% improvement on standard KITTI and 10.1% relative improvement on the Dark-KITTI dataset.

3.3.3. SegNet, U-Net

Kamal et al. [80] integrated two different network architectures based on FCNs [91] to localize and classify traffic signs. This approach operated on a combination of SegNet [97] and U-Net [95] to detect segment signs and a VGG-16[64] network for the classification.

Initially, the method extracts patches from all the corners of the input images. The models (SegNet and U-Net) train on these patches to predict the segmented traffic signs. Then predictions of the four patches are merged to make an output mask for the original image. VGG-16[64] categorizes the detected sign into 14 various classes. To increase the detection of small traffic signs, the authors used the L1 constraint term to modify Tversky Loss [98] instead of the conventional intersection over union loss. To appraise the system, the author used the CURE-TSD dataset [99] and achieved a precision of 94.60% and recall of 80.21% beating previous state-of-the-art results.

3.3.4. Fully Convolutional Siamese Networks

Utilizing fully convolutional Siamese networks [100] *Wang et al.* [81] proposed *SiamMask*, a modified network with an addition of binary segmentation task. This work explained the importance of producing per-frame binary segmentation masks instead of relying on low-level object representations extracted from existing spatial extraction networks [64].

As siamese networks [100] can conveniently encode the required information to produce a pixel-wise binary mask, the method adds a different branch and loss to the existing architecture. The first modification generates a binary mask by adding a simple two-layer neural network with learnable parameters during training. Secondly, the authors proposed binary logistic regression loss for the new branch.

Unlike Mask R-CNN [26] and other segmentation architectures that rely on maintaining spatial information throughout the network, the proposed approach generates masks starting from a flattened representation of the object. For evaluation, two networks SiamFC [101] and SiamRPN [102] are modified to perform experiments on the VOT-2018 dataset [103]. SiamMask was able to perform better than prior existing algorithms. Possible failure cases are shown in Figure 20.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

3.4. Attention Based Salient Object Detection

Employing salient object detection *Zhengzheng et al.* [82] fused RGB images with thermal images to extract multi-level features to detect objects in adverse conditions. The proposed work used a two-stream convolution neural network architecture to generate RGB and thermal image features.

The proposed work exploited multiple attention mechanisms to extract weighted features from both RGB and thermal input images. The extracted weighted features are then fused to form single input for later convolutions. These feature maps are passed through lower-level to high-level convolution blocks in a standard forward pass fashion. Although high-level features are vital, mid-level features are essential to refine object details. Therefore, this work added Pyramid Pooling Module and a feature aggregation Module to sharpen the object details and facilitate localizing the object.

To remove noise from features, the approach adds the Convolutional Block Attention Module(CBAM) [104] to apply channel-wise and spatial-wise attention. Later, average pooling aggregates the spatial information from feature maps while max-pooling generates two different spatial context information. Subsequently, the approach merged the spatial context features and applied the attention mechanism with the sigmoid. During the training, edge loss and cross-entropy loss are added together. The proposed method was tested on the VT5000 dataset [82].

4. Datasets

The performance of deep neural network-based approaches directly relates to the size of datasets. In any area of research, standard datasets are essential to benchmark the performance and draw fair comparisons between various proposed approaches. Several challenging large-scale datasets have been introduced that consist of images or videos captured in a harsh environment. In this section, we take a look at publicly available challenging datasets. Moreover, Table 2 summarizes the essential information of the explained datasets.

4.1. ExDARK

One of the very few publicly available low light datasets is the ExDARK[1] dataset. Patil et al.[105] published this dataset in 2020. The dataset contains 7363 real-life low-light pictures from different low-light environments, such as indoors and at nighttime. There are a total of 12 classes in the dataset. All the images in the dataset are manually collected, and afterward, different image enhancement techniques like de-hazing and blurring are applied to them. The classes of this dataset are **table, cat, people, motorbike, dog, cup, chair, bicycle, boat, bottle, bus, car, and cat**. As highlighted in Table 2, this dataset only contains information for the boundaries of objects. Figure 15 (a) exhibits few samples from this dataset. We have exploited this dataset to evaluate the performance of current state-of-the-art object detection algorithms on this dataset.

4.2. CURE-TSD

CURE-TSD [99][2] is a large-scale publicly available dataset for the task of traffic sign detection. The dataset contains video sequences manually gathered by driving a car around different areas at different times of the day to capture diverse scenarios. Furthermore, the manually collected images are augmented to provide different kinds of occlusions such as rainy weather, decolorization, blur, darkening, dirty lens, exposure, codex error, snow, and haze. The authors of this dataset have suggested 34 videos for training, whereas 15 videos are allocated for testing purposes. Given five different levels of challenges to each type of occlusions, there is a total of 1.72 million frames. There are 14 types of traffic signs in this dataset labeled as **speed limit, goods vehicles, no overtaking, no stopping, no parking,**

[1] https://github.com/cs-chan/Exclusively-Dark-Image-Dataset
[2] https://github.com/olivesgatech/CURE-TSD

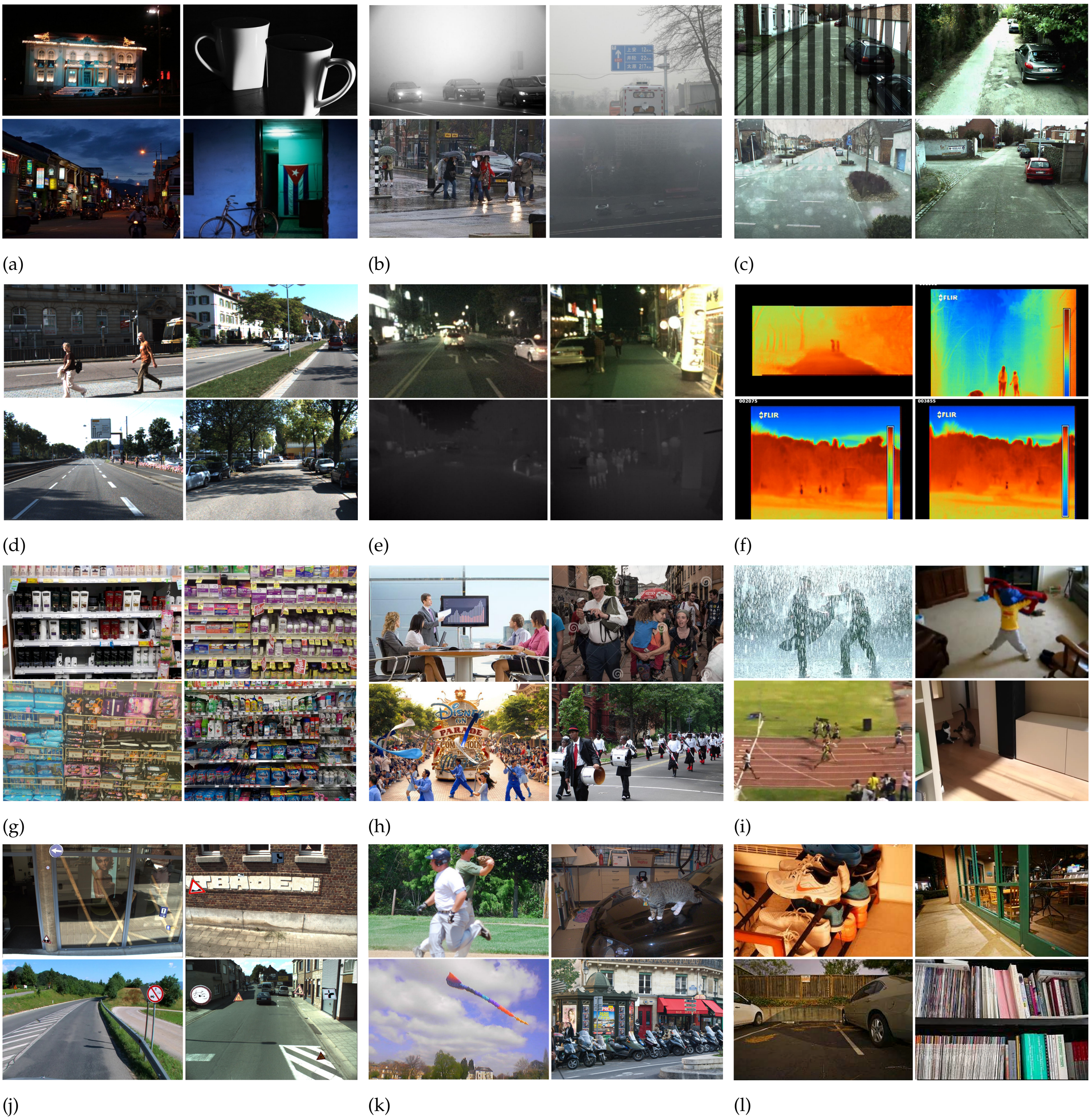

Figure 15. Sample images belonging to challenging datasets. (a) represents samples taken from ExDark [105] showing low-light examples, (b) depicts samples from Reside [106] explaining harsh weather, (c) describes samples from CURE-TSD [99] illustrating example of camera distortion, lens flare, (d) denotes samples from KITTI [96] depicting various objects to be detected, (e) represents samples from Kaist [68] explaining saliency maps with their night time images, (f) depicts samples taken from UNIRI-TID [61] showing example of thermal images, (g) highlights samples from SKU-110K [76] representing example of cluttered objects, (h) represents samples taken from Wider Face [71] showing faces at various angles, (i) represents samples taken from VOT-2018 [103] presenting example of complex indoor scenes, (j) shows samples taken from DFG [89] illustrating traffic signs at various places, (k) represents images taken from MS-COCO [12] describing example of objects in daily life and (l) outlines samples taken from See in the dark [59] dataset exhibiting examples captured at low illumination and high exposure.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

stop, bicycle, hump, no left, no right, priority to, no entry, yield, parking. Figure 15 (c) illustrates few samples of this dataset. We have included this dataset in our benchmarking experiments.

4.3. RESIDE

RESIDE [106] is a publicly available dataset employed for the task of object detection in difficult weathers. The dataset contains multiple subdivisions, each for a separate task. However, we have worked on the Real-world Task-driven Testing Set (RTTS) subset. The subset RTTS contains 4,332 real-world hazy images representing different scenarios in a day. Images were collected manually through video cameras and annotated with bounding boxes localizing objects. The dataset contains various real-world occlusions such as hazy, rainy, snowy weather, and so on. There are five classes annotated as **bicycle, bus, motorbike, car, and person**. Figure 15 (b) depicts few samples from this dataset. We incorporate this dataset along with ExDARK [105] and CURE-TSD [99] to benchmark the results on state-of-the-art object detection algorithms.

4.4. SKU-110K

SKU-100K [3] is a new dataset collected by authors of the approach "Precise Detection in Densely Packed Scenes [76]" where they have explained that for challenging scenarios, overlapping detections can occur.

The publishers of this dataset collected data from supermarkets and stores worldwide using cell cameras. The dataset only contains bounding box annotations for each image. The dataset contains 11,762 images. For training, the authors have suggested 8,233 images, 588 images for validation, and 2,941 images for testing. To ensure the same shop does not appear in more than one set, the authors recommended random selection of the sample images for training, validation, and testing. A few samples from SKU-110K dataset are visible in Figure 15 (g).

4.5. UNIRI-TID

UNIRI-TID [61] [4] is a novel dataset published in [60] for the task of detecting human/non-human objects in difficult weather. The dataset contains thermal images collected manually using the thermal camera in different weather conditions. The dataset was created to provide a balanced ratio of images for each kind of difficult weather like rainy, fog, clear, and hazy. Furthermore, the creators of this dataset applied custom augmentation like grayscale to increase the number of images. The dataset has a total of 35,974 images with 2663 images for clear weather, 2313 for rainy weather, 1135 for foggy weather, 18333 are grayscale images, and the remaining 6111 images contain a mix of all possible weather conditions. Figure 15 (f) illustrates few samples of this dataset.

4.6. KAIST Multispectral Pedestrian Detection

KAIST Multispectral Pedestrian [68] [5] is a publicly available dataset containing both RGB and thermal images for the task of pedestrian detection. The dataset contains 95000 images with thermal-RGB pairs captured by a camera mounted on a vehicle. The dataset contains challenging images captured during nighttime and in harsh weather. Furthermore, the dataset provides scale, occlusions, and other challenges for object detection. The dataset only has two classes of pedestrians and background. Few Samples of the dataset are illustrated in Figure 15 (e).

[3] https://github.com/eg4000/SKU110K_CVPR19
[4] https://ieee-dataport.org/open-access/thermal-image-dataset-person-detection-uniri-tid
[5] http://rcv.kaist.ac.kr/multispectral-pedestrian/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

Table 2. Important statistics of challenging object detection datasets. OD denotes Object Detection, SOD represents Salient Object Detection, and COD is for Category Specific object Detection.

Dataset	OD	SOD	COD	Samples	Type	Location
ExDark [105]	✓	✗	✓	7K	Camera Captured Image	https://github.com/cs-chan/Exclusively-Dark-Image-Dataset
CURE-TSD[99]	✓	✗	✓	1.72M	Camera Captured Videos	https://github.com/olivesgatech/CURE-TSD
UNIRI-TID [61]	✓	✗	✓	36K	Thermal Camera Captured Image	https://ieee-dataport.org/open-access/thermal-image-dataset-person-detection-uniri-tid
KAIST Multispectral Pedestrian Detection [68]	✓	✓	✗	95K	Camera Captured Image	https://soonminhwang.github.io/rgbt-ped-detection/
DFG Traffic Sign [89]	✓	✗	✓	7K	Camera Captured Image	http://www.vicos.si/Downloads/DFGTSD
VT5000 [82]	✓	✓	✗	5.5K	Camera Captured Image	https://github.com/lz118/RGBT-Salient-Object-Detection
Wider Face [71]	✓	✗	✗	32K	Camera Captured Image	http://shuoyang1213.me/WIDERFACE/
UFDD [69]	✓	✗	✗	6.4K	Camera Captured Image	https://ufdd.info/
See-in-the-Dark [59]	✓	✗	✓	5K	Camera Captured Raw	https://github.com/cchen156/Learning-to-See-in-the-Dark
MS-COCO [12]	✓	✓	✓	328K	Camera Captured Image	https://cocodataset.org/#download
VOT-2018 [103]	✓	✗	✗	60	Video captured sequences	https://zenodo.org/record/3257319
Kitti [96]	✓	✓	✓	1.5K	Camera Captured Image	http://www.cvlibs.net/datasets/kitti/index.php
SKU-110K [76]	✓	✗	✗	11.7K	Camera Captured Image	https://github.com/eg4000/SKU110K_CVPR19
Reside [106]	✓	✗	✓	4.3K	Camera Captured Image	https://github.com/Boyiliee/RESIDE-dataset-link

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

4.7. DFG Traffic Sign Dataset

DFG traffic sign dataset [89] [6] is another dataset that provides ground truth for both traffic sign detection and recognition tasks. There is a total of 7 thousand images collected with 200 different traffic signs. The complete dataset is divided into a training set with 5254 images and a test set with 1703 images. There is also an extended version of the dataset available with artificially augmented traffic signs. The extended version contains 15K traffic signs in natural images. Figure 15 (j) exhibits few examples of this dataset.

4.8. VT5000 Dataset

VT5000 dataset [7] is the custom dataset curated by Tu et al. [82] for the task salient object detection. The dataset has 11 challenges collected in different scenes. There are 5500 pairs of RGB images with their corresponding thermal images. The main advantage of this dataset is that the sample images consist of objects with different sizes and varying illumination conditions.

4.9. Wider Face

Wider face dataset [71] [8] is another dataset curated for detecting faces in complex scenarios. The dataset contains 32,203 images with high variance in scales, poses, and occlusions. Due to several tiny faces in the images, this dataset is one of the most challenging datasets for face detection. Few samples are visible in Figure 15 (h).

4.10. UFDD

UFDD [69] is a new face detection dataset that captures real issues not present in other datasets. The dataset contains images captured in different weather scenarios with rain, snow, and haze weather. Motion blur and focus blur have been manually added to images to increase the total number of images in the datasets. There is a total of 6,452 images with 10,897 annotations.

4.11. See-in-the-Dark

Another dataset that operates under low light is published by *Chen et al.* [59]. The dataset contains 5094 raw short-exposure images of both indoor and outdoor scenarios. Outdoor images are taken at night time to add further complications. There are only bounding box annotations in the dataset. Therefore, it is an ideal dataset to employ object detection-based approaches. Figure 15 (l) explains the dataset visually.

4.12. MS-COCO

Common Objects in Context (MS-COCO) [9] [12] is one of the biggest publicly available datasets introduced by Microsoft in 2015. The dataset contains 91 objects with 2.5 million labeled instances in 328000 images. The dataset contains annotations for object detection tasks, instance segmentation, keypoints detection, panoptic segmentation, and salient object detection. The authors have split the dataset into three parts, with 82,783 images belonging to the training set, 40,504 images for the validation set, and 40,775 images for the test set. Few samples of the dataset are visible in Figure 15 (k).

4.13. VOT-2018

VOT-2018 [103] is a publicly available dataset for tracking challenges. The dataset has 60 sequences captured with the help of a camera. The data captured in the dataset has various challenges such as occlusion, illumination change, motion change, size change,

6 http://www.vicos.si/Downloads/DFGTSD
7 https://github.com/lz118/RGBT-Salient-Object-Detection
8 http://shuoyang1213.me/WIDERFACE/
9 https://cocodataset.org/#download

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

and camera motion. The dataset facilitates the researchers to perform object detection and instance segmentation as well. Few instances of the dataset are shown in Figure 15 (i).

4.14. Kitti

KITTI [96][10] is one of the largest publicly available datasets. KITTI consists of several sub-datasets that perform various tasks such as image segmentation, 2D object detection, and 3D object detection. The data is collected with the help of LiDAR and various cameras mounted on vehicles and driving around the mid-size city of Karlsruhe, in rural areas, and on highways. More or less, each image has 20 annotated objects. There is a total of eight object classes. Figure 15 (d) depicts few instances of this dataset.

5. Experiments

We have investigated the performance of current state-of-the-art object detection algorithms on the three most challenging datasets. The idea is to draw an analysis that explains how well object detection algorithms can perform under harsh conditions. We employed Faster R-CNN [24], Mask R-CNN [26], YOLO V3 [39], Retina-Net [107], and Cascade Mask R-CNN [108] to benchmark their performance on the datasets of ExDARK [105] , CURE-TSD [99], and RESIDE [106].

We have leveraged the capabilities of transfer learning in our experiments. All the object detection networks are incorporated with a backbone of ResNet50 [109] pre-trained on COCO dataset [12]. We fine-tuned all the models for 15 epochs with a learning rate of $2 \times e^{-5}$ and used Adam [110] as an optimizer. We resized images to 800×800 during the training and testing phases.

6. Evaluation

This section discusses the well-known evaluation criteria essential to standardize state-of-the-art results for object detection in difficult situations. Moreover, this section analyzes the performance of the approaches discussed in Section 3 with quantitative and qualitative illustrations. Finally, we will present the outcome of our experiments on the three most widely exploited challenging datasets.

6.1. Evaluation Criteria

The standardization of how to assess the performance of approaches on unified datasets is imperative. Since object detection in a challenging environment is identical to generic object detection, the approaches appraise similar evaluation metrics.

6.1.1. Precision

Precision [111] defines as the percentage of a predicted region that belongs to the ground truth. Figure 16 illustrates an the difference between precise object detection and imprecise object detection. The formula for precision is explained below:

$$\frac{\text{Predicted area in ground truth}}{\text{Total area of predicted region}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

where TP denotes True Positives and FP represents False positives.

6.1.2. Recall

Recall [111] is calculated as the percentage of the ground truth region that is present in the predicted region. The formula for the recall is given by:

$$\frac{\text{Ground truth area in predicted region}}{\text{Total area of ground truth region}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

[10] http://www.cvlibs.net/datasets/kitti/index.php

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

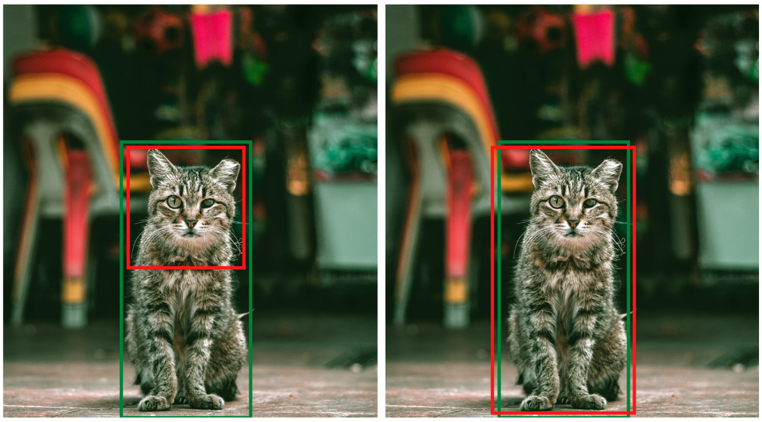

Figure 16. The image explains the visual difference between precise and imprecise prediction in object detection. The green color represents the ground truth, and the red color depicts the predicted boundary. Considering the IOU threshold value equals 0.5, the left prediction is not precise because the IOU between the ground truth and the inferred bounding box is less than 0.5. The bounding box prediction on the right side is precise because it covers almost the complete ground truth area.

Where TP is True Positives and FN represents False Negatives.

6.1.3. F-Measure

F-measure [111] is computed by taking the harmonic mean of precision and Recall. Mathematically, it is explained below:

$$\frac{2\times \text{ Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \tag{3}$$

6.1.4. Intersection Over Union

Intersection Over Union (IOU) [112] is one of the most important evaluation metrics that is regularly employed to determine the performance of object detection algorithms. It is the measure of how much the predicted region is overlapping with the actual ground truth region. IOU is defined as follows:

$$\frac{\text{Area of Overlap region}}{\text{Area of Union region}} \tag{4}$$

6.1.5. Average Precision(AP)

Average Precision computes average value of precision over different levels of recall. Higher the value of AP, the better performance and vice versa. Formula for calculating average precision is mentioned below:

$$\text{AP} = \sum_{n}(R_n - R_{n-1})P_n \tag{5}$$

where $\mathrm{R_n}$ and $\mathrm{P_n}$ are the precision and recall at the $\mathrm{n_{th}}$ threshold.

6.1.6. Mean Absolute Error(MAE)

Mean absolute error measures average pixel-level absolute difference between predicted value and ground truth. It can be calculated as follows:

$$\sum_{i=1}^{D} |x_i - y_i| \tag{6}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

6.1.7. Mean Average Precision

Mean Average Precision (mAP) is another extensively applied evaluation metric for category-specific object detection. The mAP is the mean of average precision computed over all the classes. Mathematically, it is explained by:

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{7}$$

where AP_i is the average precision for a given class explained in Section 6.1.5 and N depics the total number of classes.

6.2. Evaluations for Object Detection in a Challenging Environment

Object detection in a challenging environment is to distinguish objects from the background and regress their boundaries. Table 3 and 4 summarizes the performance comparison of various methodologies that are explained in Section 3. Different approaches have reported results on specific evaluation metrics. Therefore, we have placed a "-" on the unused metric. For the sake of readability, we have presented the results of [70] and [71] in a separate table (Table 4) because they reported results on their own introduced evaluation metrics.

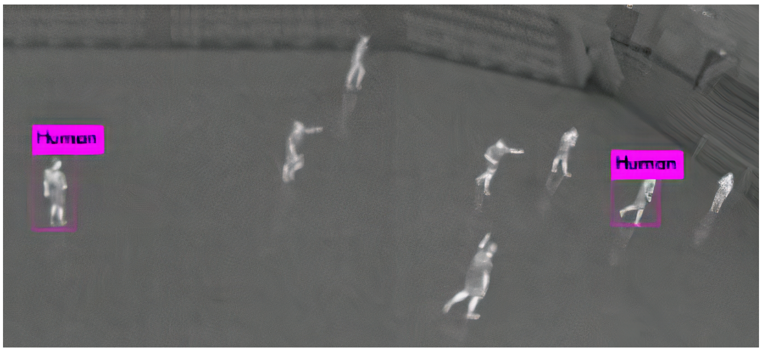

Figure 17. Failure case of the method proposed by *Mate et al.*. Fails at detecting persons from the bird-eye view. Image taken from [60].

Along with the quantitative analysis, we examine a few of the discussed approaches by paying attention to their failure cases. In Figure 17, we observe that from the bird-eye view, the method proposed by [60] fails to detect persons. Similarly, the work of [66] gets confused with non-living human-like objects in images as exhibited in Figure 18.

Figure 18. Failure case of the method proposed by *Sarin et al.*. Human and face detector fails in low light and gets fooled by non-living human-like objects in images. Image is taken from [66].

The system of *Kamal et al.* [80] is vulnerable to detect traffic signs that are not close to the camera. A visual illustration is visible in Figure 19. Another approach presented by *Wang et al.* [81] unable to detect objects in motion blur images. On top of that, the work

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

Table 3. Object Detection Performance Comparison. Outstanding results in all the respective methods are highlighted.

Literature	Year	Dataset	IOU	mAP	AP	F-Measure	Highlights
Eran et al. [76]	2019	SKU-110K	0.50:0.95	-	0.49	-	Deep IoU with EM-Merger (Section3.2.3)
Yuxuan et al. [63]	2020	ExDark	0.50:0.95	-	0.34	-	RFB-Net (Section 3.1.2)
Zhengzheng et al. [82]	2020	VT5000	-	-	-	0.81	Attention based SOD (Section 3.4)
Mate et al. [60]	2020	KAIST	0.50:0.95	0.35	-	0.36	YOLO (Section 3.1.1)
Ghose et al. [78]	2019	KAIST	0.50:0.95	0.68	-	-	Deep Saliency Networks (Section 3.3.1)
Avramovic et al. [75]	2020	DFG	0.50	**0.94**	-	-	Mask R-CNN (Section 3.2.2)
Sasagawa et al. [54]	2020	SID	-	0.55	-	-	YOLO (Section 3.1.1)
Rashed et al. [79]	2019	KITTI	**0.75**	-	-	-	ShuffleSeg Segmentation Network (Section 3.3.2)
Wang et al. [81]	2019	VOT-2016	0.50	**0.90**	-	-	FC Siamese Networks (Section 3.3.4)
Kun et al. [77]	2019	Night-Dataset	-	**0.82**	-	-	GAN + Faster R-CNN (Section 3.2.1)
Kamal et al. [80]	2019	CURE-TSD	0.50:0.95	-	**0.94**	-	SegNet + U-Net (Section 3.3.3)

Table 4. Object Detection Performance. The results mentioned in this table are not directly comparable because different datasets have been used.

Literature	Year	Dataset	Evaluation Metric	Score	Highlight
Ranjan et al. [70]	2019	WIDER Face	True Accept Rate [70]	91.4%	Single Shot deep pyramid Face Detector (Section 3.1.3.1)
Sarin et al. [66]	2019	KAIST	True Positive Rate [66]	92.80%	Single Shot Human and Face Detector (Section3.1.3)

tends to produce false positives on the images where no object is present. Figure 20 exhibits an example of such cases.

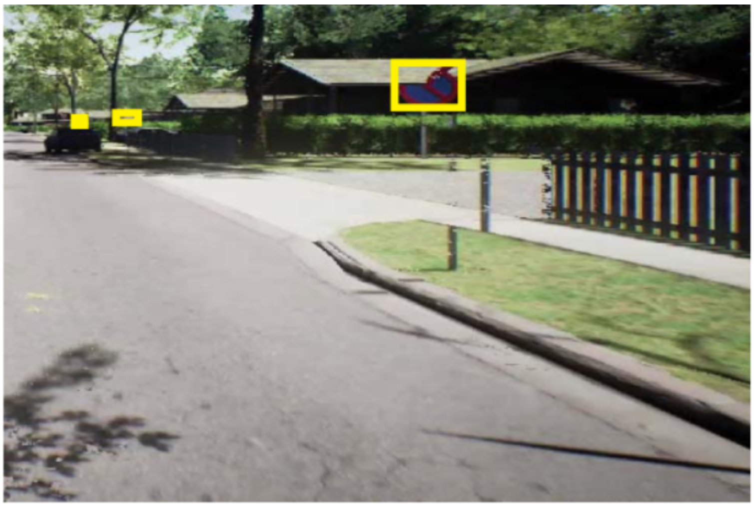

Figure 19. Failure case of the proposed method by *Kamal et al.*. The proposed method fails if traffic signs are placed at different places. Image is courtesy of [80]

6.3. Evaluation of Our Experiments

The performance of current state-of-the-art object detection algorithms on challenging datasets is presented in Table 5. By looking at Table 5, it is evident that there is significant room for improvement on all of the employed datasets. For the ExDark dataset, YOLO

V3 [39] produced the best results with an AP of 0.67. For CURE-TSD, Cascade Mask R-CNN [108] yields the best score with an AP of 0.28. In the case of RESIDE dataset, Mask R-CNN [26] shows the best results with an AP of 0.51.

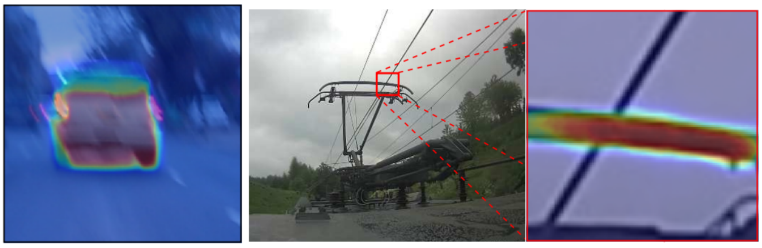

Figure 20. Failure case of the proposed method by *Wang et al.*. The proposed method fails when faced with motion blur or a "non-object" pattern. Image is courtesy of [81].

The main reason for the low performance of these state-of-the-art generic object detection algorithms is that they are not trained on challenging datasets that include low-light images or occluded images. Furthermore, the backbone network of these architectures cannot optimally extract the spatial features necessary for detecting objects in challenging environments. Hence, it is empirically established that generic object detection algorithms are not ideal for resolving object detection in challenging images.

7. Open Challenges and Future Directions

After reviewing many methodologies and benchmarking on object detection in a challenging environment, we have noticed several open issues that need to be highlighted to be tackled in future directions.

Standard Evaluation Criteria: We observe that there has not been a defined standard while benchmarking on object detection in challenging situations. Some approaches have employed f1-score [60,82], whereas some have used mAP and AP to report their results [76,78,80]. Moreover, we notice that different IOU thresholds have been exploited to present the result, making the approaches incomparable to each other. Therefore, it is vital to establish a generic standard on which all future approaches working in this field can report the results.

Real-time Applications: Similar to generic object detection, real-time challenging object detection is the growing need in practical scenarios. Generally, deep neural architectures require high computing power, which is unavailable in embedded devices. Therefore, it is essential to build methods that can optimally work on low computational resources. Furthermore, future research should discuss the computation of their proposed method, including the quantitative and qualitative analysis.

Weak/Unsupervised Approaches: To the best of our knowledge, all the current approaches tackling object detection in tricky environments with deep neural networks require large-scale annotated datasets. The creation of such datasets is a timely expensive, and laborious process. Hence, there is a growing need to introduce challenging object detection systems that can train on no labels (unsupervised) or very few labels (weak supervised learning).

Intelligent Domain Adaptation: Current object detection systems under challenging environments work in similar types of challenges. For instance, a network trained on low-light images will not detect objects present in underwater images. However, we humans can adapt our learned knowledge into various domains. In the near future, we can expect a flurry of innovative ideas towards this promising direction.

Leveraging GANs: The deep neural network-based object detection systems demand enormous datasets for training the network. The ability of GANs to produce fake images can be highly beneficial to create large-scale challenging datasets. Moreover, most of the challenging datasets contain low-light, blurry, or not spatially rich images [68,105,106]. In future researches, powers of recently published cycle GANs [113,114] could be exploited to transform images that are better suited for object detection algorithms.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

Table 5. Object Detection Performance Comparison on three challenging datasets. AP_s denotes average precision for small area at IoU=0.50:0.95. AP_m represents average precision for medium area at IoU=0.50:0.95 and AP_l depicts average precision for large area at IoU=0.50:0.95.

Our Models	ExDark					CURE-TSD					RESIDE				
	AP	**AP^{50}**	**AP_s**	**AP_m**	**AP_l**	**AP**	**AP^{50}**	**AP_s**	**AP_m**	**AP_l**	**AP**	**AP^{50}**	**AP_s**	**AP_m**	**AP_l**
Mask R-CNN	0.54	0.84	0.22	0.46	0.59	0.20	0.35	0.019	0.17	0.37	**0.51**	**0.79**	**0.4**	0.11	**0.57**
Faster R-CNN	0.53	0.82	0.22	0.46	0.58	0.25	**0.43**	0.03	0.14	**0.41**	0.49	0.78	0.07	0.70	0.56
Yolo V3	**0.67**	**0.93**	**0.5**	**0.61**	**0.71**	0.16	0.32	0.05	0.03	0.26	0.37	0.78	0.07	**0.87**	0.56
Retina-Net	0.36	0.67	0.12	0.30	0.52	0.14	0.25	**0.12**	0.19	0.35	0.48	0.75	0.002	0.07	0.55
Cascade Mask R-CNN	0.49	0.78	0.27	0.37	0.55	**0.28**	0.38	0.06	**0.23**	0.34	0.50	0.76	0.006	0.12	0.56

Combining Spatial and Textual Information: Modern object detection approaches under complex scenarios operate on CNN-based architectures that mainly depend on spatial features. However, images present in the challenging datasets are not spatially rich, resulting in suboptimal extracted features. Therefore, fusing textual information along with spatial information will facilitate the network to localize the objects. Nevertheless, developing efficient multi-modal systems to detect objects in difficult conditions would be an exciting future direction.

8. Conclusion

With the rapid computational advances, noticeable improvements have been observed by employing deep learning-based object detection algorithms in challenging environments. This paper comprehensively reviews the relevant approaches that have attempted to identify objects under challenging scenarios through deep neural networks and provides an empirical performance analysis of currently employed object detection algorithms. Moreover, the advantages and limitations of all explained techniques are provided. Furthermore, this work describes all the relevant datasets currently exploited to resolve the problem of object detection under challenging environments. This work discusses the well-known evaluation criteria and presents the performance analysis of explained methodologies and object detection algorithms by benchmarking recent object detection algorithms on challenging datasets of ExDark [105], CURE-TSD [99], and Reside [106]. In the end, we highlight the current open issues and present future direction that can enhance the research of object detection under challenging environments.

Author Contributions: writing—original draft preparation, M.A., K.A.H.; writing—review and editing, K.A.H., M.Z.A.; supervision and project administration, A.P., M.L., D.S. All authors have read and agreed to the submitted version of the manuscript.

Funding: The work leading to this publication has been partially funded by the European project INFINITY under Grant Agreement ID 883293.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* **2009**, *32*, 1627–1645.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

2. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3150–3158.
3. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Hypercolumns for object segmentation and fine-grained localization. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 447–456.
4. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Simultaneous detection and segmentation. European Conference on Computer Vision. Springer, 2014, pp. 297–312.
5. Alberti, C.; Ling, J.; Collins, M.; Reitter, D. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054* **2019**.
6. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. International conference on machine learning. PMLR, 2015, pp. 2048–2057.
7. Wu, Q.; Shen, C.; Wang, P.; Dick, A.; Van Den Hengel, A. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence* **2017**, *40*, 1367–1381.
8. Kang, K.; Li, H.; Yan, J.; Zeng, X.; Yang, B.; Xiao, T.; Zhang, C.; Wang, Z.; Wang, R.; Wang, X.; others. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology* **2017**, *28*, 2896–2907.
9. Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; Zheng, N. Semantics-guided neural networks for efficient skeleton-based human action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1112–1121.
10. Vaswani, N.; Chowdhury, A.R.; Chellappa, R. Activity recognition using the dynamics of the configuration of interacting objects. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. IEEE, 2003, Vol. 2, pp. II–633.
11. Motwani, T.S.; Mooney, R.J. Improving Video Activity Recognition using Object Recognition and Text Mining. ECAI. Citeseer, 2012, Vol. 1, p. 2.
12. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollar, P. Microsoft COCO: common objects in context (2014). *arXiv preprint arXiv:1405.0312* **2019**.
13. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision* **2010**, *88*, 303–338.
14. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, Vol. 1, pp. 886–893.
15. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D. Cascade object detection with deformable part models. 2010 IEEE Computer society conference on computer vision and pattern recognition. IEEE, 2010, pp. 2241–2248.
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
17. Alexe, B.; Deselaers, T.; Ferrari, V. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence* **2012**, *34*, 2189–2202.
18. Carreira, J.; Sminchisescu, C. CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2011**, *34*, 1312–1328.
19. Rahtu, E.; Kannala, J.; Blaschko, M. Learning a category independent object detection cascade. 2011 international conference on Computer Vision. IEEE, 2011, pp. 1052–1059.
20. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *International journal of computer vision* **2013**, *104*, 154–171.
21. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. European conference on computer vision. Springer, 2014, pp. 391–405.
22. Kuo, W.; Hariharan, B.; Malik, J. Deepbox: Learning objectness with convolutional networks. Proceedings of the IEEE international conference on computer vision, 2015, pp. 2479–2487.
23. Alexe, B.; Deselaers, T.; Ferrari, V. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence* **2012**, *34*, 2189–2202.
24. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* **2015**.
25. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
26. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
27. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. European conference on computer vision. Springer, 2016, pp. 21–37.
28. Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; Shum, H.Y. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence* **2010**, *33*, 353–367.
29. Cheng, M.M.; Mitra, N.J.; Huang, X.; Torr, P.H.; Hu, S.M. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence* **2014**, *37*, 569–582.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

30. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
31. Ouyang, W.; Wang, X.; Zeng, X.; Qiu, S.; Luo, P.; Tian, Y.; Li, H.; Yang, S.; Wang, Z.; Loy, C.C.; others. Deepid-net: Deformable deep convolutional neural networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2403–2412.
32. Han, J.; Zhang, D.; Cheng, G.; Liu, N.; Xu, D. Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Processing Magazine* **2018**, *35*, 84–100.
33. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
34. Girshick, R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
35. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* **2015**.
36. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
37. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
38. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263–7271.
39. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.
40. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* **2020**.
41. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
42. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* **2004**, *60*, 91–110.
43. Lienhart, R.; Maydt, J. An extended set of haar-like features for rapid object detection. Proceedings. international conference on image processing. IEEE, 2002, Vol. 1, pp. I–I.
44. Agarwal, S.; Terrail, J.O.D.; Jurie, F. Recent advances in object detection in the age of deep convolutional neural networks. *arXiv preprint arXiv:1809.03193* **2018**.
45. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; others. Speed/accuracy trade-offs for modern convolutional object detectors. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7310–7311.
46. Grauman, K.; Leibe, B. Visual object recognition. *Synthesis lectures on artificial intelligence and machine learning* **2011**, *5*, 1–181.
47. Andreopoulos, A.; Tsotsos, J.K. 50 years of object recognition: Directions forward. *Computer vision and image understanding* **2013**, *117*, 827–891.
48. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055* **2019**.
49. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A survey of deep learning-based object detection. *IEEE Access* **2019**, *7*, 128837–128868.
50. Arnold, E.; Al-Jarrah, O.Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems* **2019**, *20*, 3782–3795.
51. Bharati, P.; Pramanik, A. Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey. In *Computational Intelligence in Pattern Recognition*; Springer, 2020; pp. 657–668.
52. Tian, Y.; Yang, G.; Wang, Z.; Wang, H.; Li, E.; Liang, Z. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Computers and electronics in agriculture* **2019**, *157*, 417–426.
53. Lan, W.; Dang, J.; Wang, Y.; Wang, S. Pedestrian detection based on YOLO network model. 2018 IEEE international conference on mechatronics and automation (ICMA). IEEE, 2018, pp. 1547–1551.
54. Sasagawa, Y.; Nagahara, H. Yolo in the dark-domain adaptation method for merging multiple models. European Conference on Computer Vision. Springer, 2020, pp. 345–359.
55. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215* **2014**.
56. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*, 1097–1105.
57. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* **2015**, *37*, 1904–1916.
58. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning. PMLR, 2015, pp. 448–456.
59. Chen, C.; Chen, Q.; Xu, J.; Koltun, V. Learning to see in the dark. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3291–3300.
60. Krišto, M.; Ivasic-Kos, M.; Pobar, M. Thermal Object Detection in Difficult Weather Conditions Using YOLO. *IEEE Access* **2020**, *8*, 125459–125476.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

61. Krišto, M.; Ivašić-Kos, M. Thermal imaging dataset for person detection. 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, 2019, pp. 1126–1131.
62. Liu, S.; Huang, D.; others. Receptive field block net for accurate and fast object detection. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 385–400.
63. Xiao, Y.; Jiang, A.; Ye, J.; Wang, M.W. Making of night vision: Object detection under low-illumination. *IEEE Access* **2020**, *8*, 123075–123086.
64. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
65. Szegedy, C.; Reed, S.; Erhan, D.; Anguelov, D.; Ioffe, S. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441* **2014**.
66. Sarin, M.; Chandrakar, S.; Patel, R. Face and Human Detection in Low Light for Surveillance Purposes. 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE). IEEE, 2019, pp. 614–620.
67. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*, 1097–1105.
68. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral pedestrian detection: Benchmark dataset and baseline. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1037–1045.
69. Nada, H.; Sindagi, V.A.; Zhang, H.; Patel, V.M. Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, 2018, pp. 1–10.
70. Ranjan, R.; Bansal, A.; Zheng, J.; Xu, H.; Gleason, J.; Lu, B.; Nanduri, A.; Chen, J.C.; Castillo, C.D.; Chellappa, R. A fast and accurate system for face detection, identification, and verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science* **2019**, *1*, 82–96.
71. Yang, S.; Luo, P.; Loy, C.C.; Tang, X. Wider face: A face detection benchmark. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5525–5533.
72. Klare, B.F.; Klein, B.; Taborsky, E.; Blanton, A.; Cheney, J.; Allen, K.; Grother, P.; Mah, A.; Jain, A.K. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1931–1939.
73. Whitelam, C.; Taborsky, E.; Blanton, A.; Maze, B.; Adams, J.; Miller, T.; Kalka, N.; Jain, A.K.; Duncan, J.A.; Allen, K.; others. Iarpa janus benchmark-b face dataset. proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 90–98.
74. Maze, B.; Adams, J.; Duncan, J.A.; Kalka, N.; Miller, T.; Otto, C.; Jain, A.K.; Niggel, W.T.; Anderson, J.; Cheney, J.; others. Iarpa janus benchmark-c: Face dataset and protocol. 2018 International Conference on Biometrics (ICB). IEEE, 2018, pp. 158–165.
75. Avramović, A.; Sluga, D.; Tabernik, D.; Skočaj, D.; Stojnić, V.; Ilc, N. Neural-Network-Based Traffic Sign Detection and Recognition in High-Definition Images Using Region Focusing and Parallelization. *IEEE Access* **2020**, *8*, 189855–189868.
76. Goldman, E.; Herzig, R.; Eisenschtat, A.; Goldberger, J.; Hassner, T. Precise detection in densely packed scenes. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5227–5236.
77. Wang, K.; Liu, M.Z. Object Recognition at Night Scene Based on DCGAN and Faster R-CNN. *IEEE Access* **2020**, *8*, 193168–193182.
78. Ghose, D.; Desai, S.M.; Bhattacharya, S.; Chakraborty, D.; Fiterau, M.; Rahman, T. Pedestrian detection in thermal images using saliency maps. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.
79. Rashed, H.; Ramzy, M.; Vaquero, V.; El Sallab, A.; Sistu, G.; Yogamani, S. Fusemodnet: Real-time camera and lidar based moving object detection for robust low-light autonomous driving. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.
80. Kamal, U.; Tonmoy, T.I.; Das, S.; Hasan, M.K. Automatic traffic sign detection and recognition using SegU-net and a modified tversky loss function with L1-constraint. *IEEE Transactions on Intelligent Transportation Systems* **2019**, *21*, 1467–1479.
81. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H. Fast online object tracking and segmentation: A unifying approach. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1328–1338.
82. Tu, Z.; Ma, Y.; Li, Z.; Li, C.; Xu, J.; Liu, Y. RGBT salient object detection: A large-scale dataset and benchmark. *arXiv preprint arXiv:2007.03262* **2020**.
83. Fan, Q.; Brown, L.; Smith, J. A closer look at Faster R-CNN for vehicle detection. 2016 IEEE intelligent vehicles symposium (IV). IEEE, 2016, pp. 124–129.
84. He, Z.; Zhang, L. Multi-adversarial faster-rcnn for unrestricted object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6668–6677.
85. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine* **2018**, *35*, 53–65.
86. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028* **2017**.
87. Kopelowitz, E.; Engelhard, G. Lung Nodules Detection and Segmentation Using 3D Mask-RCNN. *arXiv preprint arXiv:1907.07676* **2019**.
88. Zhang, Q.; Chang, X.; Bian, S.B. Vehicle-damage-detection segmentation algorithm based on improved mask RCNN. *IEEE Access* **2020**, *8*, 6997–7004.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2021 doi:10.20944/preprints202106.0590.v1

89. Tabernik, D.; Skočaj, D. Deep learning for large-scale traffic-sign detection and recognition. *IEEE transactions on intelligent transportation systems* **2019**, *21*, 1427–1440.
90. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* **2014**.
91. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
92. Liu, N.; Han, J.; Yang, M.H. Picanet: Learning pixel-wise contextual attention for saliency detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3089–3098.
93. Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; Heng, P.A. R3net: Recurrent residual refinement network for saliency detection. Proceedings of the 27th International Joint Conference on Artificial Intelligence. AAAI Press, 2018, pp. 684–690.
94. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation* **1997**, *9*, 1735–1780.
95. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
96. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 3354–3361.
97. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* **2017**, *39*, 2481–2495.
98. Salehi, S.S.M.; Erdogmus, D.; Gholipour, A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. International workshop on machine learning in medical imaging. Springer, 2017, pp. 379–387.
99. Temel, D.; Chen, M.H.; AlRegib, G. Traffic sign detection under challenging conditions: A deeper look into performance variations and spectral characteristics. *IEEE Transactions on Intelligent Transportation Systems* **2019**, *21*, 3663–3673.
100. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. European conference on computer vision. Springer, 2016, pp. 850–865.
101. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. European conference on computer vision. Springer, 2016, pp. 850–865.
102. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High performance visual tracking with siamese region proposal network. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8971–8980.
103. Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Pflugfelder, R.; ˇCehovin Zajc, L.; Vojir, T.; Bhat, G.; Lukezic, A.; Eldesokey, A.; others. The sixth visual object tracking vot2018 challenge results. Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.
104. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
105. Loh, Y.P.; Chan, C.S. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding* **2019**, *178*, 30–42.
106. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing* **2018**, *28*, 492–505.
107. Jaeger, P.F.; Kohl, S.A.; Bickelhaupt, S.; Isensee, F.; Kuder, T.A.; Schlemmer, H.P.; Maier-Hein, K.H. Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. Machine Learning for Health Workshop. PMLR, 2020, pp. 171–183.
108. Cai, Z.; Vasconcelos, N. Cascade R-CNN: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence* **2019**.
109. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, Vol. 31.
110. Zhang, Z. Improved adam optimizer for deep neural networks. 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE, 2018, pp. 1–2.
111. Powers, D.M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* **2020**.
112. Blaschko, M.B.; Lampert, C.H. Learning to localize objects with structured output regression. European conference on computer vision. Springer, 2008, pp. 2–15.
113. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.
114. Nidadavolu, P.S.; Villalba, J.; Dehak, N. Cycle-gans for domain adaptation of acoustic features for speaker recognition. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6206–6210.