# A Survey of Deep Learning-based Object Detection

| | |
|---|---|
| Journal: | *IEEE Access* |
| Manuscript ID | Access-2019-26723 |
| Manuscript Type: | Original Manuscript |
| Date Submitted by the Author: | 11-Jul-2019 |
| Complete List of Authors: | Jiao, Licheng; Xidian University, Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education<br>Zhang, Fan; Xidian University, School of Artificial Intelligence<br>Liu, Fang; Xidian University, Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation<br>Yang, Shuyuan; Xidian University, Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education<br>Li, Lingling; Xidian University, School of Artificial Intelligence<br>Feng, Zhixi; Xidian University<br>Qu, Rong; University of Nottingham, School of Computer Science |
| Keywords: | Computer vision, Object detection, Machine learning |
| Subject Category<br>Please select at least two subject categories that best reflect the scope of your manuscript: | Computational and artificial intelligence, Geoscience and remote sensing, Imaging |
| Additional Manuscript Keywords: | survey |
| | |

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.DOI*

# A Survey of Deep Learning-based Object Detection

**LICHENG JIAO[1], (Fellow, IEEE), FAN ZHANG[1], FANG LIU[1], (Senior Member, IEEE), SHUYUAN YANG[1], (Senior Member, IEEE), LINGLING LI[1], (Member,IEEE), ZHIXI FENG[1], (Member, IEEE), AND RONG QU[2], (Senior Member, IEEE)**

[1]Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Joint International Research Laboratory of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xian, Shaanxi Province 710071, China e-mail: (lchjiao@mail.xidian.edu.cn, zhangfan_1@stu.xidian.edu.cn, f63liu@163.com, syyang@xidian.edu.cn, linglingxidian@gmail.com, zxfeng@xidian.edu.cn)

[2]ASAP Research Group, School of Computer Science, University of Nottingham, Nottingham, NG8 1BB, UK email: (rong.qu@nottingham.ac.uk)

Corresponding author: Licheng Jiao (e-mail: lchjiao@mail.xidian.edu.cn).

This work was supported in part by the State Key Program of National Natural Science of China (No.61836009), Project supported the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No.61621005), the Major Research Plan of the National Natural Science Foundation of China (No.91438201, 91438103, 61801124), the National Natural Science Foundation of China (No.U1701267, 61871310, 61573267), the Fund for Foreign Scholars in University Research and Teaching Programs, the 111 Project (No.B07048) and the Program for Cheung Kong Scholars and Innovative Research Team in University (No.IRT_15R53).

**ABSTRACT** Object detection is one of the most important and challenging branches of computer vision, which has been widely applied in people's life, such as monitoring security, autonomous driving and so on, with the purpose of locating instances of semantic objects of a certain class. With the rapid development of deep learning networks for detection tasks, the performance of object detectors has been greatly improved. In order to understand the main development status of object detection pipeline, thoroughly and deeply, in this survey, we first analyze the methods of existing typical detection models and describe the benchmark datasets. Afterwards and primarily, we provide a comprehensive overview of a variety of object detection methods in a systematic manner, covering the one-stage and two-stage detectors. Moreover, we list the traditional and new applications. Some representative branches of object detection are analyzed as well. Finally, we discuss the architecture of exploiting these object detection methods to build an effective and efficient system and point out a set of development trends to better follow the state-of-the-art algorithms and further research.

**INDEX TERMS** Classification, deep learning, localization, object detection, typical pipelines.

## I. INTRODUCTION

OBJECT detection has been attracting increasing amounts of attention in recent years due to its wide range of applications and recent technological breakthroughs. This task is under extensive investigation in both academia and real world applications, such as monitoring security, autonomous driving, transportation surveillance, drone scene analysis, and robotic vision. Among many factors and efforts that lead to the fast evolution of object detection techniques, a notable contribution should be attributed to the development of deep convolution neural networks and GPUs computing power. At present, deep learning model has been widely used in the whole field of computer vision, including general object detection and domain-specific object detection. State-of-the-art object detectors almost use deep learning networks as their both backbone and detection network for extracting features from the input images (or videos), classification and localization respectively. Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of image object detection include multi-categories detection, edge detection, salient object detection, pose detection, scene text detection, face detection, and pedestrian detection. Because a rising number of applications need scene understanding, as an important part image object detection has been widely used in many areas of modern life. So far many benchmarks play an important role in object detection field, such as Caltech [1], KITTI [2], ImageNet [3], PASCAL VOC [4], and MS COCO [5]. In ECCV VisDrone 2018 contest, the organizer release a novel dataset benchmark contains a large amount of images and videos based on the drone platform.

Pre-existing domain-specific image object detectors usually can be divided into two categories, the one is two-stage detector, the most representative one, Faster R-CNN [6]. The other is one-stage detector, such as YOLO [7], SSD [8]. Two-stage detectors have high localization and object recognition accuracy, whereas the one-stage detectors achieve high inference speed. The two stages of two-stage detectors is divided by ROI (Region of Interest) pooling layer. For instance, in Faster R-CNN, the first stage, called RPN, a Region Proposal Network, proposes candidate object bounding boxes. The second stage, features are extracted by RoIPool operation from each candidate box for the following classification and bounding-box regression missions [9]. Fig.1 (a) shows the basic architecture of two-stage detectors. Furthermore, the one-stage detectors propose predicted boxes from input images directly without region proposal step, thus they are time efficient and can be used for real-time devices. Fig.1 (b) exhibits the basic architecture of one-stage detectors.

Our survey is focus on describing and analyzing deep learning based object detection. The existing surveys always cover a series of domain of general object detection and may not contain state-of-the-art methods which provide some novel solutions and newly directions of these tasks because of rapid development. We list very novel solutions proposed recently but neglect to discuss the basics so that readers can see the cutting edge of the field more easily. Different from previous object detection surveys, in this paper we systematically and comprehensively review deep learning based object detection methods and most importantly the up to date detection solutions while research trends. Our survey is featured by in-depth analysis and discussion in various aspects, many of which, to the best of our knowledge, are the first time in this field. It is our intention to provide an overview how different deep learning methods are being used rather than a full summary of all related papers. To get into the field, we recommend readers refer to [10] [11] [12] for more details of early methods.

The rest of the paper is organized as follows. Object detectors need a powerful backbone network for rich feature extracting. We discuss backbone networks in section 2 below. The typical pipeline domain-specific image detectors act as basics and milestone of the task. In section 3, we will elaborate the most representative and pioneering deep learning-based approaches proposed before June 2019. The common used datasets and metrics will be described in section 4. The analyses of general image object detection methods are systematically explained in section 5. In section 6, we describe five typical fields for object detection and several popular branches of object detection. The development trend is summarized in section 7.

## II. BACKBONE NETWORKS

Backbone network is acting as the basic feature extractor for object detection task which takes images as input and outputs feature maps of the corresponding input image. Most of these networks are the network for classification task taking out the last fully connected layers. The improved version of basic classification network is also available. For instance, Lin *et al.* [13] add or subtract layers or replace some layers with special designed layers. To better meet specific requirements, some works [7] [14] utilize the newly designed backbone for feature extracting.

For different requirements about accuracy vs. efficiency, people can choose deeper and densely connected backbones, like ResNet [9], ResNeXt [15], AmoebaNet [16] etc. or lightweight backbones like MobileNet [17], ShuffleNet [18], SqueezeNet [19], Xception [20], MobileNetV2 [21] etc. When applied to mobile devices, lightweight backbones can meet the requirements. Wang *et al.* [22] propose a novel real-time object detection system by combining PeleeNet with [8] and optimizing the architecture for fast processing speed. But the more precise applications need high accuracy thus complicated backbones. On the other hand, the real-time acquirements like video or webcam not only need high processing speed but high accuracy [7], which require finely designed backbone to adapt to the detection architecture also make a trade-off between speed and accuracy.

To explore more competitive detecting accuracy, deeper and densely connected backbone is adopting to replace the shallower and sparse connected counterpart. He *et al.* [9] utilize ResNet [23] rather than VGG [24] which is adopted in Faster R-CNN [6] for further accuracy gain because of its high capacity to capture rich features.

The newly high performance classification networks can improve the precision and reduce the complexity of object detection task. This is an effective way to further improve network performance because backbone network is acting as a feature extractor. As is known to all, the quality of the features determines the upper bound of network performance, thus it is an important step that needs further exploration. Please refer to [25] for more details.

## III. TYPICAL BASELINES

With the advent of deep learning and increasing computing power, great progress has been made in general object detection domain. When the first CNN-based object detector R-CNN was proposed, a series of significant contributions have been made which promote the development of general object detection. We introduce some representative object detection architectures for beginners to get started in this domain.

### A. R-CNN

R-CNN is a region based CNN detector. As Ross Girshick *et al.* [26] propose R-CNN which could be used in object detection tasks, their works are the first to show that a CNN could lead to dramatically higher object detection performance on PASCAL VOC datasets [4] than those systems based on simpler HOG-like features. Deep learning method is verified effective and efficient in the field of object detection.

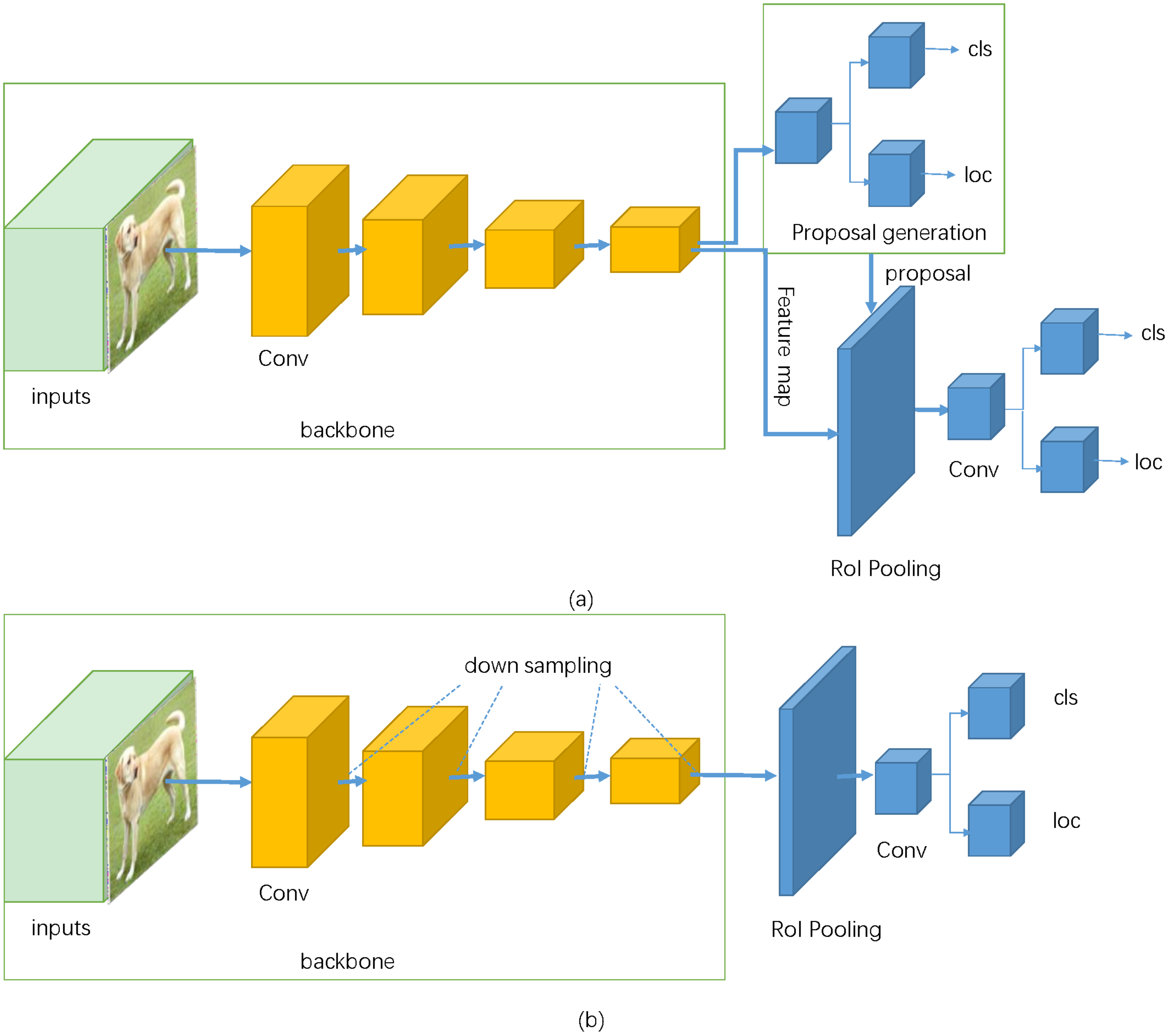


FIGURE 1: (a) exhibits the basic architecture of two-stage detectors, which consists of region proposal network to feed region proposals into classifier and regressor. (b) shows the basic architecture of one-stage detectors, which predicts bounding boxes from input images directly.Yellow cubes are a series of Conv layers (called a block) with the same resolution in backbone network, because of down-sampling operation after one block, the size of the following cubes gradually becoming small. Thick blue cubes are a series of Conv layers contains one or more convolutional layers. The flat blue cube demonstrates the RoI pooling layer to generate features of an object with the same size.

R-CNN detector consists of four modules. The first module generates category-independent region proposals. The second module extracts a fixed-length feature vector from each region proposal. The third module is a set of class-specific linear SVMs to classify the objects in one image. The last module is a bounding-box regressor for precisely bounding-box prediction. For detailed, first, to generate region proposals, the authors adopt selective search method. Then, a CNN is used for extracting a 4096-dimensional feature vector from each region proposal. Because the fully connected layer needs input vectors of fixed length, the region proposal features should have the same size. The authors adopt a fixed 227 × 227 pixel as the input size of CNN. As we know, the objects in various images have different size and aspect ratio, which makes the region proposals extracted by the first module different in size. Regardless of the size or aspect ratio of the candidate region, the authors warp all pixels in a tight bounding box around it to the required size 227 × 227. The feature extraction network consists of five convolutional layers and two fully connected layers. And all CNN parameters are shared across all categories. Each category trains category-independent SVMs which don't share parameters between different SVMs.

Pre-training on lager dataset followed by fine-tuning on the specified dataset is a good training method for deep convolutional neural networks to achieve fast convergence. First, Ross Girshick *et al.* [26] pre-train the CNN on a large scale dataset (ImageNet classification dataset [3]). The last fully connected layer is replaced by the CNN's ImageNet specific 1000-way classification layer. The next step is fine-tuning the CNN parameters on the warped proposal windows uses SGD (stochastic gradient descent). The last fully connected layer is the (N+1)-way classification layer (N: object classes, 1: background) which is randomly initialized.

When setting positive examples and negative examples the authors divide into two parts. The one is defining the IoU (intersection over union) overlap threshold 0.5 in fine-tuning process, below which region proposals are defined as negatives while surpass which object proposals are defined as positives. As well, the object proposals whose maximum IoU overlap with a ground-truth class are assigned to the ground-truth box. The other is setting parameters when training SVMs. In contrast, only the ground-truth boxes are taken as positive examples for their respective classes and proposals have less than 0.3 IoU overlap with all ground-truth instances of one class as a negative proposal for that class. Because those proposals with overlap between 0.5 and 1 but not ground truth expand the number of positive examples by approximately 30×, the big set can avoid overfitting during fine-tuning the entire network effectively.

### B. FAST R-CNN

R-CNN proposed a year later, Ross Girshick [27] proposed a faster improved version of R-CNN, called Fast R-CNN [27]. Because R-CNN performs a ConvNet forward pass for each region proposal without sharing computation, R-CNN takes a long time on SVMs classification. Fast R-CNN extracts features from an entire input image and then passes the region of interest (RoI) pooling layer to get the fixed size features as the input of the followed classification and bounding box regression fully connected layer. The features are extracted from the entire image once and are sent to CNN for classification and localization at a time compared to R-CNN inputs each region proposals to CNN, which can save a lot of time used for CNN processing and large disk storage to store a great deal of features. As mentioned above, training R-CNN is a multi-stage process which covers pre-training stage, fine-tuning stage, SVMs classification stage and bounding box regression stage. Fast R-CNN is a one-stage end-to-end training process using a multi-task loss on each labeled RoI to jointly train for classification and bounding box regression.

Another improvement is Fast R-CNN uses a RoI pooling layer to extract a fixed size feature map from region proposals have different size. This operation with no need for warping regions and reserves the spatial information of features of region proposals. For fast detection, Ross Girshick uses truncated SVD which accelerates the forward pass of computing the fully connected layers.

Experiment results show that Fast R-CNN has 66.9% mAP while R-CNN has 66.0% on PASCAL VOC 2007 dataset [4]. Training time drops to 9.5 hours as compared to R-CNN with 84h, 9 times faster. For test rate (s/image), Fast R-CNN with truncated SVD (0.32s) is 213× faster than R-CNN (47s). Those experiments were proceeding on an Nvidia K40 GPU, all of which demonstrated that Fast R-CNN did accelerate object detection.

### C. FASTER R-CNN

Faster R-CNN [6] makes an improvement in region-based CNN baseline after Fast R-CNN proposed 3 months. Fast R-CNN uses selective search for proposing RoI, which is slow and needs the same running time as the detection network. Faster R-CNN replaces it with a novel RPN (region proposal network) that is a fully convolutional network to efficiently predict region proposals with a wide range of scales and aspect ratios. RPN accelerates the generating speed of region proposals as well as shares fully-image convolutional features and a common set of convolutional layers with the detection network. The procedure is simplified in Fig.3 (b). Another novel method for different sized object detection is using multi-scale anchors as reference. The anchors can greatly simplify the process of generating various sized region proposals with no need of multiple scales of input images or features. On the outputs (feature maps) of the last shared convolutional layer, sliding a fixed size window (3 × 3), the center point of each feature window is relative to a point of the original input image which is the center point of k (3 × 3) anchor boxes. The author set anchor boxes have 3 different scales and 3 aspect ratios. The region proposal is parameterized relative to a reference anchor box. Then measure the distance between predicted box and the corresponding ground truth to optimize the location of the predicted boxes.

Experiments indicated that Faster R-CNN has greatly improved both precision and detection efficiency. On PASCAL VOC 2007 test set, Faster R-CNN achieved mAP of 69.9% as compared to Fast R-CNN of 66.9% with shared convolutional computations. As well, total running time of Faster R-CNN (198ms) is nearly 10 times lower than Fast R-CNN (1830ms) with the same VGG [24] backbone, and processing rate is 5fps vs. 0.5fps.

### D. MASK R-CNN

Mask R-CNN [9] is an extending work to Faster R-CNN mainly for instance segmentation task. Regardless of the adding parallel mask branch, Mask R-CNN can be seen a more accurate object detector. He *et al.*use Faster R-CNN with a ResNet [23]-FPN [13] (feature pyramid network, a backbone extracts RoI features from different levels of the feature pyramid according to their scale) backbone for feature extraction achieves excellent precision and processing speed. FPN contains a bottom-up pathway and a top-down pathway with lateral connections. The bottom-up pathway is a backbone ConvNet which computes a feature hierarchy consisting of feature maps at several scales with a scaling step of 2. The top-down pathway produces higher resolution features by upsampling spatially coarser, but semantically stronger, feature maps from higher pyramid levels. At the beginning, the top pyramid feature maps are captured by the output of the last convolutional layer of the bottom-up pathway. Each lateral connection merges feature maps of the same spatial size from the bottom-up pathway and the top-down pathway. While the dimensions of feature maps are different, the $1 \times 1$ convolutional layer can change the dimension. Once undergoing a lateral connection operation, there will form a new pyramid level and predictions are independently made on each level. Because higher-resolution feature maps are important for detecting small objects while lower-resolution feature maps are rich in semantic information, feature pyramid network extracts significant features.

Another way to improve accuracy is replacing RoI pooling with RoIAlign for extracting a small feature map from each RoI, as shown in Fig.2. Traditional RoI pooling quantizes floating-number in two steps to get approximate feature values in each bin. First, quantization was applied for calculating the coordinate of each RoI in feature maps, given the coordinates of RoIs in the input images and down sampling stride. Then the authors divide RoI feature maps into bins to generate feature maps with the same size, which is also quantized during the process. These two quantization operations cause misalignments between the RoI and the extracted features. To address this, at those two steps, RoIAlign avoids any quantization of the RoI boundaries or bins. First it computes the floating-number of the coordinates of each RoI feature map followed by a bilinear interpolation operation to compute the exact values of the features at four regularly sampled locations in each RoI bin. Then it aggregates the results using max or average pooling to get values of each bin. Fig. 2 is an example of RoIAlign operation.

Experiments showed that with the above two improvements the precision got promotion. Using ResNet-FPN backbone improved 1.7 points box AP and RoIAlign operation improved 1.1 points box AP on MS COCO detection dataset.

### E. YOLO

YOLO [7] (you only look once) is a one-stage object detector proposed by Joseph Redmon *et al.*after Faster R-CNN [6]. The main contribution is real-time detecting full images and webcam. Firstly, it is due to this pipeline only predicts less than 100 bounding boxes per image while Fast R-CNN using selective search predicts 2000 region proposals per image. Secondly, YOLO frames detection as a regression problem, so a unified architecture can extract features from input images straightly for predicting bounding boxes and class probabilities. YOLO base network runs at 45 frames per second with no batch processing on a Titan X GPU as compared to Fast R-CNN at 0.5fps and Faster R-CNN at 7fps.

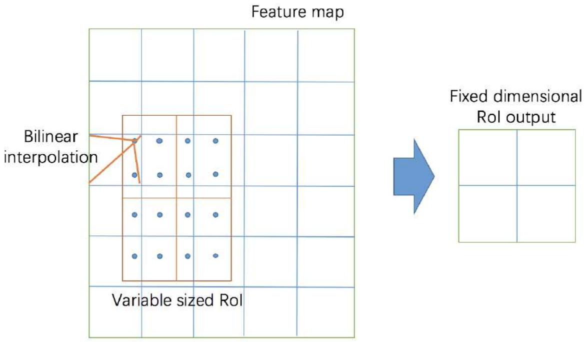


FIGURE 2: RoIAlign operation. The first step calculates floating number coordinates of an object in the feature map. Next step utilizes bilinear interpolation to compute the exact values of the features at four regularly sampled locations in the separated bin.

YOLO pipeline first divides the input image into an $S \times S$ grid, where a grid cell is responsible for detecting the object whose center falls into. The confidence scores multiplied by two parts, $P(object)$ denoting the probability of the box contains an object and IOU (intersection over union) showing how accurate the box contain that object. Each grid cell predicts B bounding boxes $(x, y, w, h)$ and confidence scores for them and C-dimension conditional class probabilities for C categories. The feature extraction network contains 24 convolutional layers followed by 2 fully connected layers. When pre-training on ImageNet dataset, the authors use the first 20 convolutional layers and an average pooling layer followed by a fully connected layer. For detection, the whole network is used for better performance. In order to get fine-grained visual information improving detection precision, in detection stage double the input resolution of $224 \times 224$ in pre-training stage.

The experiments showed that YOLO was not good at accurate localization and localization error was the main component of prediction error. Fast R-CNN makes many background false positives mistakes while YOLO is 3 times less than it. Training and testing on PASCAL VOC dataset, YOLO achieves 63.4% mAP with 45 fps as compared to Fast R-CNN (70.0% mAP, 0.5fps) and Faster R-CNN (73.2% mAP, 7fps).

### F. YOLOV2

YOLOv2 [28] is a second version of YOLO [7], which adopts many design decisions from past works with novel concepts to improve YOLO's speed and precision.

Batch Normalization. Fixed distribution of inputs to a ConvNet layer would have positive consequences for the layers. It is impractical to normalize the entire training set because the optimization step uses stochastic gradient descent. Since SGD uses mini-batches during training, each mini-batch produces estimates of the mean and variance of each activation. Computing the mean and variance value of the mini-batch of size m, then normalize the activations of number m to have mean zero and variance 1. Finally the elements of each mini-batch are sampled from the same distribution. This operation can be seen as a BN layer [29] outputs activations with the same distribution. YOLOv2 add a BN layer ahead of each convolutional layer which accelerates the network to get convergence and helps regularize the model. Batch normalization gets more than 2% improvement in mAP.

High Resolution Classifier. In YOLO backbone, the classifier adopts an input resolution of $224 \times 224$ then increases the resolution to 448 for detection. This process needs the network adjust to a new resolution inputs when switches to object detection task. To address this, YOLOv2 adds a fine-tuning process on the classification network at $448 \times 448$ for 10 epochs on ImageNet dataset which increases the mAP at $4\%$.

Convolutional With Anchor Boxes. In original YOLO networks, coordinates of predicted boxes are directly generating by fully connected layers. Faster R-CNN uses anchor boxes as reference to generate offsets with predicted boxes. YOLOv2 adopts this prediction mechanism and firstly removes fully connected layers. Then it predicts class and objectness for every anchor box. This operation increases 7% recall while mAP decreases 0.3%.

Predicting the size and aspect ratio of anchor boxes using dimension clusters. In Faster R-CNN the size and aspect ratio of anchor boxes is identified empirically. For easier learning to predict good detections, YOLOv2 uses K-means clustering on the training set bounding boxes to automatically get good priors. Using dimension clusters along with directly predicting the bounding box center location improves YOLO by almost 5% over the above version with anchor boxes.

Fine-Grained Features. For localizing smaller objects, high-resolution feature maps can provide useful information. Similar to the identity mappings in ResNet, YOLOv2 concatenates the higher resolution features with the low resolution features by stacking adjacent features into different channels which gives a modest 1% performance increase.

Multi-Scale Training. For networks to be robust to run on images of different sizes, every 10 batches the network randomly chooses a new image dimension size from $\{320, 352, ..., 608\}$. This means the same network can predict detections at different resolutions. At high resolution detection, YOLOv2 achieves 78.6% mAP and 40fps as compared to YOLO with 63.4% mAP and 45fps on VOC 2007.

As well, YOLOv2 proposes a new classification backbone namely Darknet-19 with 19 convolutional layers and 5 max-pooling layers which requires less operations to process an image yet achieves high accuracy. The more competitive YOLOv2 version has 78.6% mAP and 40fps as compared to Faster R-CNN with ResNet backbone of 76.4% mAP and 5fps, and SSD500 has 76.8% mAP and 19fps. As mentioned above, YOLOv2 can achieve high detecting precision while high processing rate which benefit from 7 main improvements and a new backbone.

TABLE 1: AP scores (%) on the MS COCO dataset,$AP_S$:AP of small objects, $AP_M$:AP of medium objects, $AP_L$:AP of large objects

| Model | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|
| DSSD513 | 13.0 | 35.4 | 51.1 |
| RetinaNet | 24.1 | 44.2 | 51.2 |

### G. YOLOV3

YOLOv3 [30] is an improved version of YOLOv2. First, YOLOv3 uses multi-label classification (independent logistic classifiers) to adapt to more complex datasets containing many overlapping labels. Second, YOLOv3 uses three different scales feature maps to predict the bounding box. The last convolutional layer predicts a 3-d tensor encoding class predictions, objectness, and bounding box. Third, YOLOv3 proposes a deeper and robust feature extractor, called Darknet-53, inspired by ResNet to get deeper.

According to results of experiments on MS COCO dataset, YOLOv3 (AP:33%) performs on par with the SSD variant (DSSD513:AP:33.2%) on MS COCO metrics yet 3 times faster than it while quite a bit behind RetinaNet [31] (AP:40.8%). But uses the "old" detection metric of mAP at IOU= 0.5 (or $AP_{50}$), YOLOv3 can achieve 57.9% mAP as compared to DSSD513 of 53.3% and RetinaNet of 61.1%. Due to the advantages of multi-scale predictions, YOLOv3 can detect small objects even more but has comparatively worse performance on medium and larger size objects.

### H. RETINANET

RetinaNet [31] is a one-stage object detector with focal loss as classification loss function proposed by Lin *et al.* [31] in February 2018. The architecture of RetinaNet is shown in Fig.4 (c). R-CNN is a typical two-stage object detector. The first stage generates a sparse set of region proposals and the second stage classifies each candidate location. Owing to the first stage filters out the majority of negative locations, two-stage object detectors can achieve higher precision than one-stage detectors which propose a dense set of candidate locations. The main reason is the extreme foreground-background class imbalance when one-stage detectors train networks to get convergence. So the authors proposed a loss function, called focal loss, which can down-weight the loss assigned to well-classified or easy examples, focusing on the hard training examples and avoiding the vast number of easy negative examples overwhelming the detector during training. RetinaNet inherits the fast speed of previous one-stage detectors while greatly overcomes one-stage detectors difficult to training for unbalanced positive and negative examples.

Experiments showed that RetinaNet with ResNet-101-FPN backbone got 39.1% AP as compared to DSSD513 of 33.2% AP on MS COCO test-dev dataset. With ResNeXt-101-FPN, it made 40.8% AP far surpassing the-state-of-the-art one-stage detector–DSSD513. RetinaNet improved the detection precision on small and medium objects by a large margin.

### I. SSD

SSD [8], a single-shot detector for multiple categories within one-stage which directly predicting category scores and box offsets for a fixed set of default bounding boxes of different scales at each location in several feature maps with different scales, as shown in Fig.4 (a). The default bounding boxes have different aspect ratios and scales in each feature map. In different feature maps, the scale of default bounding boxes is computed with regularly space between the highest layer and the lowest layer where each specific feature map learns to be responsive to the particular scale of the objects. For each default box, it predicts both the offsets and the confidences for all object categories. Fig.3 (c) shows the method. At training time, matching these default bounding boxes to the ground truth boxes where the matched default boxes as positive examples and the rest as negatives. For the large amount of default boxes are negatives, the authors adopt hard negative mining using the highest confidence loss for each default box then picking the top ones to make the ratio between the negatives and positives at most 3:1. As well, the authors implement data augmentation which is proved an effective way to enhance precision by a large margin.

Experiments showed that SSD512 had a competitive result both mAP and speed with VGG-16 [24] backbone. SSD512 (input image size: $512 \times 512$) achieved mAP of 81.6% on PASCAL VOC 2007 test set and 80.0% on PASCAL VOC 2012 test set as compared to Faster R-CNN (78.8%, 75.9%) and YOLO (VOC2012: 57.9%). On MS COCO DET dataset, SSD512 was better than Faster R-CNN in all criteria.

### J. DSSD

DSSD [32] (Deconvolutional Single Shot Detector) is a modified version of SSD (Single Shot Detector) which adding prediction module and deconvolution module also using ResNet-101 as backbone. The architecture of DSSD is shown in Fig.4 (b). For prediction module, Fu *et al.*add a residual block to each predicting layer, then do element-wise addition of the outputs of prediction layer and residual block. Deconvolution module is for increasing the resolution of feature maps to strengthen features. Each deconvolution layer followed by a prediction module is to predict a variety of objects with different sizes. At training process, first the authors pre-train ResNet-101 based backbone network on the ILSVRC CLS-LOC dataset, and then use $321 \times 321$ inputs or $513 \times 513$ inputs training the original SSD model on detection dataset, finally train the deconvolution module freezing all the weights of SSD module. Experiments on both PASCAL VOC dataset and MS COCO dataset show the effectiveness of DSSD513 model, while the added prediction module and deconvolution module bring 2.2% enhancement on PASCAL VOC 2007 test dataset.

### K. REFINEDET

The whole network [33] contains two inter-connected modules, the anchors refinement module and the object detection module. These two modules are connected by a transfer connection block to transfer and enhance features from the former module to better predict objects in the latter module. The training process is in an end-to-end way, conducted by three stages, preprocessing, detection (two inter-connected modules) and NMS.

Classical one-stage detectors such as SSD, YOLO, RetinaNet, etc. all use one-step regression method to obtain the final results. The authors find that use two-step cascaded regression method can better predict hard detected objects, especially for small objects and provide more accurate locations of objects.

### L. RELATION NETWORKS FOR OBJECT DETECTION

Hu *et al.* [34] propose an adapted attention module for object detection called object relation module which considers the interaction between different targets in an image including their appearance feature and geometry information. This object relation module is added in the head of detector before two fully connected layers to get enhanced features for accurate classification and localization of objects. The relation module not only feeds enhanced features into classifier and regressor, but replaces NMS post-processing step also gain higher accuracy than it. By using Faster R-CNN, FPN and DCN as the backbone network respectively on the COCO test-dev dataset, adding the relationship module increases the accuracy in 0.2, 0.6 and 0.2, respectively.

### M. DCNV2

For learning to adapt to geometric variation reflected in the effective spatial support region of targets, deformable convolutional networks DCN [35] was proposed by Dai *et al.*Regular ConvNets can only focus on features of fixed square size (according to the kernal), thus the receptive field does not properly cover the entire pixel of a target object to represent it. The deformable ConvNets can produce deformable kernel and the offset from the initial convolution kernel (of fixed size) are learned from the networks. Deformable RoI Pooling can also adapt to part localization for objects with different shapes. DCNv1 achieves significant accuracy improvements almost 4% enhancement than three plain ConvNets on COCO test-dev set. The best mean average-precision under the strict COCO evaluation criteria (mAP @[0.5:0.95] ) is 37.5%.

Deformable ConvNets v2 [36] utilizes more deformable convolutional layers than DCNv1 from only the conv layers in the conv5 stage to all the conv layers in the conv3-conv5

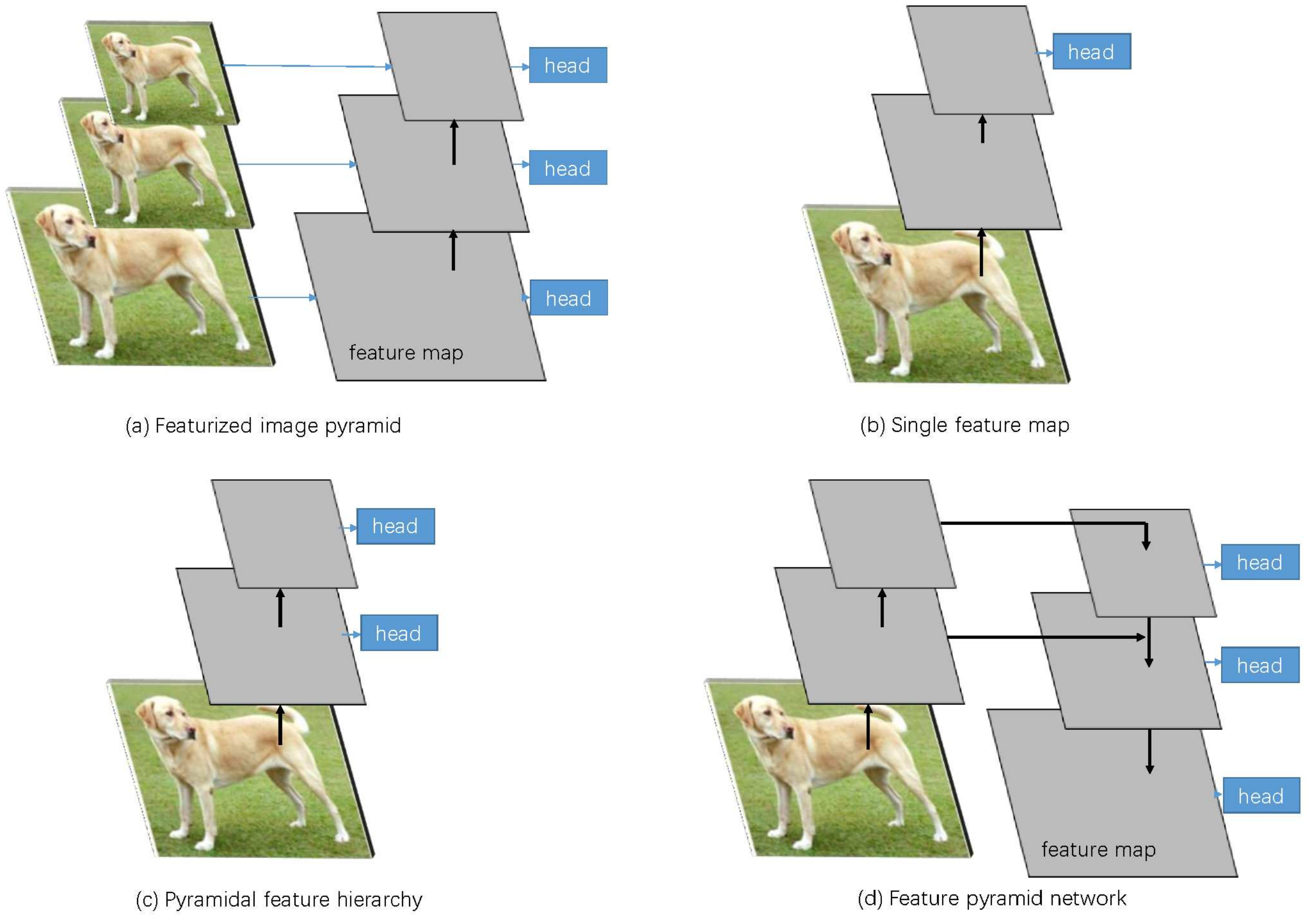

head
head
head
feature map
(a) Featurized image pyramid
head
(b) Single feature map
head
head
(c) Pyramidal feature hierarchy
head
head
head
feature map
(d) Feature pyramid network

FIGURE 3: Four methods utilize features for different sized object prediction. (a) Using an image pyramid to build a feature pyramid. Features are computed on each of the image scales independently, which is slow. (b) Detection systems [6] [27] use only single scale features (the outputs of the last conv layer) for faster detection. (c) Predicting each of the pyramidal feature hierarchy from a ConvNet as if it is a featurized image pyramid like SSD [8]. (d) Feature Pyramid Network (FPN) [13] is fast like (b) and (c), but more accurate. In this figure, feature maps are indicate by blue outlines and thicker outlines denote semantically stronger features.

stages to replace the regular conv layers. All the deformable layers are modulated by a learnable scalar, which obviously enhance the deformable effect and accuracy. The authors adopt feature mimicking to further improve detection accuracy by incorporating a feature mimic loss on the per-RoI features of DCN to be similar to good features extracted from cropped images. DCNv2 achieves 45.3% mAP under COCO evaluation criteria on the COCO 2017 test-dev set, while DCNv1 with 41.7% and regular Faster R-CNN with 40.1% on ResNext-101 backbone. On other strong backbones, DCNv2 surpasses DCNv1 by $3\% \sim 5\%$ mAP and regular Faster R-CNN by $5\% \sim 8\%$.

### N. NAS-FPN

In recent days, the authors from Google Brain adopt neural architecture search to find some new feature pyramid architecture, named NAS-FPN [16], consisting of both top-down and bottom-up connections to fuse features with a variety of different scales. By repeating FPN architecture N times and then concatenating them into a large architecture during the search, the high level feature layers pick which level features for them to imitate. All of the highest accuracy architectures have the connection between high resolution input feature maps and output feature layers, which indicate that it is necessary to generate high resolution features for detecting small targets. Stacking more pyramid networks, adding feature dimension, adopting high capacity architecture all increase detection accuracy by a large margin. Experiments show that adopting ResNet-50 as backbone with 256 feature dimension, NAS-FPN surpass the original FPN 2.9% mean average-precision on COCO test-dev dataset. The superlative configuration of NAS-FPN is utilizing AmoebaNet as backbone network and stacking 7 FPN with 384 feature dimension, which achieves 48.0% on COCO test-dev.

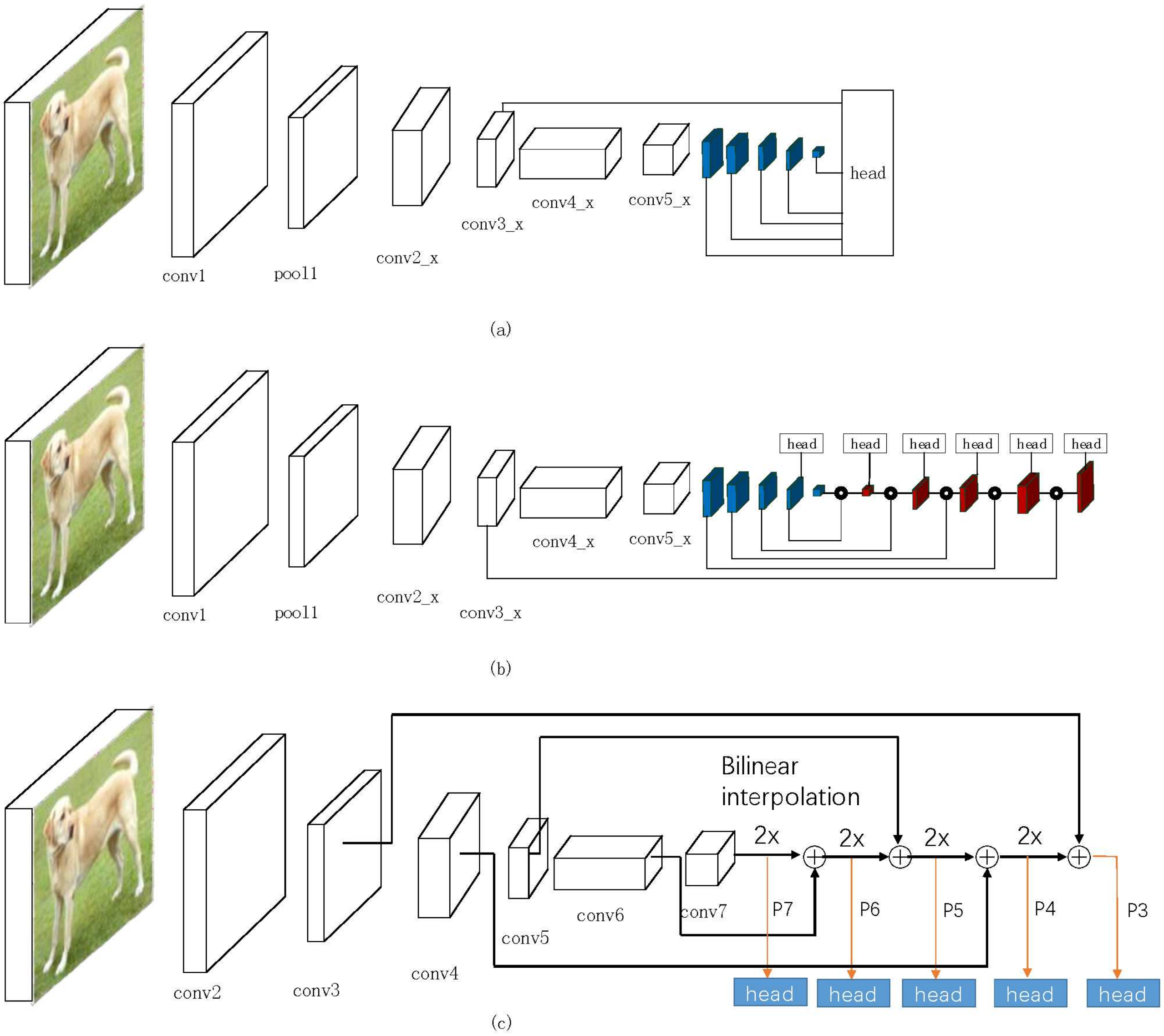


FIGURE 4: Networks of SSD, DSSD and RetinaNet on residual network. (a) The blue modules are the layers added in SSD framework whose resolution gradually drop because of down sampling. In SSD the prediction layer is acting on fused features of different levels. Head module consists of a series of convolutional layers followed by several classification layers and localization layers. (b) The red modules are the layers added in DSSD framework denoting deconvolution operation. In DSSD the prediction layer is following every deconvolution module. (c) RetinaNet utilizes ResNet-FPN as its backbone network, which generates P3-P7 5 level feature pyramid corresponding to C3-C7 (the feature map of conv3-conv7 respectively) to predict different sized objects.

### *O. M2DET*

To meet a large variety of scale variation across object instances, Zhao *et al.* [37] propose a multi-level feature pyramid network (MLFPN) constructing more effective feature pyramids. The authors adopt three steps to obtain final enhanced feature pyramids. First, like FPN, multi-level features extracted from multiple layers in the backbone are fusing as the base feature. Second, the base feature is fed into a block, composing of alternating joint Thinned U-shape Modules and Feature Fusion Modules, and obtains the decoder layers of TUM as the features for next step. Finally, the decoder layers with equivalent scales are gathered up to construct a feature pyramid containing multi-level features. So far, features with multi-scale and multi-level are prepared. The remaining part is to follow the SSD architecture to obtain bounding box localization and classification results in an end-to-end manner. For M2Det is an one-stage detector, it achieves AP of 41.0 at speed of 11.8 FPS with single-scale inference strategy and AP of 44.2 with multi-scale inference strategy utilizing VGG-16 on COCO test-dev set. It outperforms RetinaNet800 (Res101-FPN as backbone) by 0.9% with single-scale inference strategy, but is twice slower than RetinaNet800.

In conclusion, the typical baselines enhance accuracy by extracting richer features of objects and adopting multi-level and multi-scale features for different sized object detection. To achieve higher speed and precision, the one-stage detectors utilize newly designed loss function to filter out easy samples which drops the number of proposal targets by a large margin. To address geometric variation, adopting deformable convolution layers is an effective way. Modeling the relationship between different objects in an image is also necessary to improve performance. Detection results on MS COCO test-dev dataset of the above typical baselines are listed on table 2.

## IV. DATASETS AND METRICS

Detecting an object has to state that an object belongs to a specified class and localize it in the image. The localization of an object is typically represented by a bounding box as in Fig. 5. Using challenging datasets as benchmark is significant in many areas of research, because they are able to draw a standard comparison between different algorithms and set goals for solutions. Early algorithms focused on face detection using various ad hoc datasets. Later, more realistic and challenging face detection datasets were created. Another popular challenge is the detection of pedestrians for which several datasets have been created. The Caltech Pedestrian Dataset [1] contains 350,000 labeled instances with bounding boxes. General object detection datasets like PASCAL VOC [4], MS COCO [5], ImageNet-loc [3] are the mainstream benchmarks of object detection task. The official metrics are mainly adopted to measure the performance of detectors with corresponding dataset.

### A. PASCAL VOC DATASET

#### 1) Dataset

For the detection of basic object categories, a multi-year effort from 2005 to 2012 was devoted to the creation and maintenance of a series of benchmark datasets that were widely adopted. The PASCAL VOC datasets [4] contain 20 object categories (in VOC2007, such as person, bicycle, bird, bottle, dog, etc.) spread over 11,000 images. The 20 categories can be considered as 4 main branches-vehicles, animals, household objects and people. Some of them increase semantic specificity of the output, such as car and motorbike, different types of vehicle, but not look similar. In addition, the visually similar classes increase the difficulty of detection, e.g. “dog” vs. “cat”. Over 27,000 object instance bounding boxes are labeled, of which almost 7,000 have detailed segmentations. Imbalanced datasets exist in the VOC2007 dataset, while the class “person” is definitely the biggest one, which is nearly 20 times more than the smallest class “sheep” in the training set. This problem is widespread in the surrounding scene, how can detectors solve this well? Another issue is viewpoint, such as, front, rear, left, right and unspecified, the detectors need to treat different viewpoints separately. Some annotated examples are showed in the last two lines of Fig. 5.

#### 2) Metric

For the VOC2007 criteria, the interpolated average precision (Salton and McGill 1986) was used to evaluate both classification and detection. It is designed to penalize the algorithm for missing object instances, for duplicate detections of one instance, and for false positive detections.

$$Recall(t) = \frac{\sum_{ij} 1[s_{ij} \geq t] z_{ij}}{N}$$

$$Precision(t) = \frac{\sum_{ij} 1[s_{ij} \geq t] z_{ij}}{\sum_{ij} 1[s_{ij} \geq t]}$$

where $t$ is threshold to judge the IoU between predicted box and ground truth box. In VOC metric, $t$ is set to 0.5. $i$ is the index of the i-th image while $j$ is the index of the j-th object. if detection is matched to a ground truth box according to the threshold criteria, and 0 otherwise. $N$ is the number of predicted boxes. The indicator function $1[s_{ij} \geq t] = 1$ if $s_{ij} \geq t$ is true, 0 otherwise. For a given task and class, the precision/recall curve is computed from a method’s ranked output. Recall is defined as the proportion of all positive examples ranked above a given rank. Precision is the proportion of all examples above that rank which are from the positive class. The mean average precision across all categories is the ultimate results.

### B. MS COCO BENCHMARK

#### 1) Dataset

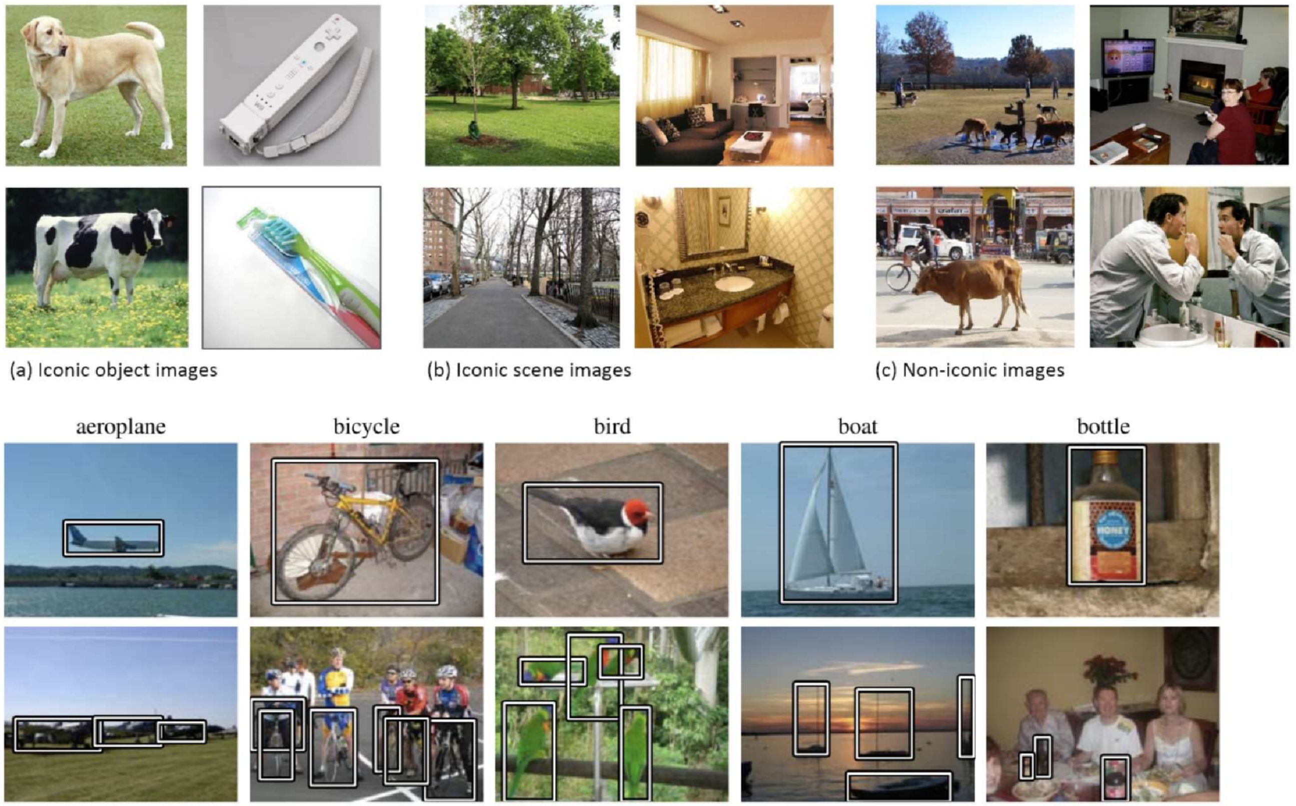
The Microsoft Common Objects in Context (MS COCO) dataset [5] for detecting and segmenting objects found in everyday life in their natural environments contains 91 common object categories with 82 of them having more than 5,000 labeled instances. These categories cover the 20 categories in PASCAL VOC dataset. In total the dataset has 2,500,000 labeled instances in 328,000 images. MS COCO dataset also pays attention to varied viewpoints and all objects of it are in natural environments which gives us rich contextual information.

In contrast to the popular ImageNet dataset [3] COCO has fewer categories but more instances per category. The dataset is also significantly larger in the number of instances per category (27k on average) than the PASCAL VOC datasets [4] (about 10 more times less than MS COCO dataset) and ImageNet object detection dataset (1k) [3]. MS COCO contains considerably more object instances per image (7.7) as compared to PASCAL VOC (2.3) and ImageNet (3.0). Furthermore, MS COCO dataset contains 3.5 categories per image as compared to PASCAL (1.4) and ImageNet (1.7) on average. In addition, 10% images in MS COCO have

TABLE 2: Detection results on the MS COCO test-dev dataset of some typical baselines. AP, $AP_{50}$ , $AP_{75}$ scores (%). $AP_S$:AP of small objects, $AP_M$:AP of medium objects, $AP_L$:AP of large objects. *DCNv2+Faster R-CNN models are trained on the 118k images of the COCO 2017 train set.

| Method | Data | Backbone | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|---|
| Fast R-CNN [27] | train | VGG-16 | 19.7 | 35.9 | – | – | – | – |
| Faster R-CNN [6] | trainval | VGG-16 | 21.9 | 42.7 | – | – | – | – |
| OHEM [38] | trainval | VGG-16 | 22.6 | 42.5 | 22.2 | 5.0 | 23.7 | 37.9 |
| ION [39] | train | VGG-16 | 23.6 | 43.2 | 23.6 | 6.4 | 24.1 | 38.3 |
| OHEM++ [38] | trainval | VGG-16 | 25.5 | 45.9 | 26.1 | 7.4 | 27.7 | 40.3 |
| R-FCN [40] | trainval | ResNet-101 | 29.9 | 51.9 | - | 10.8 | 32.8 | 45.0 |
| CoupleNet [41] | trainval | ResNet-101 | 34.4 | 54.8 | 37.2 | 13.4 | 38.1 | 52.0 |
| Faster R-CNN G-RMI [42] | – | Inception-ResNet-v2 | 34.7 | 55.5 | 36.7 | 13.5 | 38.1 | 52.0 |
| Faster R-CNN+++ [23] | trainval | ResNet-101-C4 | 34.9 | 55.7 | 37.4 | 15.6 | 38.7 | 50.9 |
| Faster R-CNN w FPN [13] | trainval35k | ResNet-101-FPN | 36.2 | 59.1 | 39.0 | 18.2 | 39.0 | 48.2 |
| Faster R-CNN w TDM [43] | trainval | Inception-ResNet-v2-TDM | 36.8 | 57.7 | 39.2 | 16.2 | 39.8 | 52.1 |
| Deformable R-FCN [35] | trainval | Aligned-Inception-ResNet | 37.5 | 58.0 | 40.8 | 19.4 | 40.1 | 52.5 |
| $umd_$det [44] | trainval | ResNet-101 | 40.8 | 62.4 | 44.9 | 23.0 | 43.4 | 53.2 |
| Cascade R-CNN [45] | trainval35k | ResNet-101-FPN | 42.8 | 62.1 | 46.3 | 23.7 | 45.5 | 55.2 |
| SNIP [46] | trainval35k | DPN-98 | 45.7 | 67.3 | 51.1 | 29.3 | 48.8 | 57.1 |
| Fitness-NMS [47] | trainval35k | ResNet-101 | 41.8 | 60.9 | 44.9 | 21.5 | 45.0 | 57.5 |
| Mask R-CNN [9] | trainval35k | ResNeXt-101 | 39.8 | 62.3 | 43.4 | 22.1 | 43.2 | 51.2 |
| DCNv2+Faster R-CNN [36] | train118k* | ResNet-101 | 44.8 | 66.3 | 48.8 | 24.4 | 48.1 | 59.6 |
| G-RMI [42] | trainval32k | Ensemble of Five Models | 41.6 | 61.9 | 45.4 | 23.9 | 43.5 | 54.9 |
| YOLOv2 [28] | trainval35k | DarkNet-53 | 33.0 | 57.9 | 34.4 | 18.3 | 35.4 | 41.9 |
| YOLOv3 [30] | trainval35k | DarkNet-19 | 21.6 | 44.0 | 19.2 | 5.0 | 22.4 | 35.5 |
| $SSD$300* [8] | trainval35k | VGG-16 | 25.1 | 43.1 | 25.8 | 6.6 | 22.4 | 35.5 |
| RON384+++ [48] | trainval | VGG-16 | 27.4 | 49.5 | 27.1 | – | – | – |
| SSD321 [32] | trainval35k | ResNet-101 | 28.0 | 45.4 | 29.3 | 6.2 | 28.3 | 49.3 |
| DSSD321 [32] | trainval35k | ResNet-101 | 28.0 | 46.1 | 29.2 | 7.4 | 28.1 | 47.6 |
| SSD512* [8] | trainval35k | VGG-16 | 28.8 | 48.5 | 30.3 | 10.9 | 31.8 | 43.5 |
| SSD513 [32] | trainval35k | ResNet-101 | 31.2 | 50.4 | 33.3 | 10.2 | 34.5 | 49.8 |
| DSSD513 [32] | trainval35k | ResNet-101 | 33.2 | 53.3 | 35.2 | 13.0 | 35.4 | 51.1 |
| RetinaNet500 [31] | trainval35k | ResNet-101 | 34.4 | 53.1 | 36.8 | 14.7 | 38.5 | 49.1 |
| RetinaNet800 [31] | trainval35k | ResNet-101-FPN | 39.1 | 59.1 | 42.3 | 21.8 | 42.7 | 50.2 |
| M2Det512 [37] | trainval35k | VGG-16 | 37.6 | 56.6 | 40.5 | 18.4 | 43.4 | 51.2 |
| M2Det512 [37] | trainval35k | ResNet-101 | 38.8 | 59.4 | 41.7 | 20.5 | 43.9 | 53.4 |
| M2Det800 [37] | trainval35k | VGG-16 | 41.0 | 59.7 | 45.0 | 22.1 | 46.5 | 53.8 |
| RefineDet320 [33] | trainval35k | VGG-16 | 29.4 | 49.2 | 31.3 | 10.0 | 32.0 | 44.4 |
| RefineDet512 [33] | trainval35k | VGG-16 | 33.0 | 54.5 | 35.5 | 16.3 | 36.3 | 44.3 |
| RefineDet320 [33] | trainval35k | ResNet-101 | 32.0 | 51.4 | 34.2 | 10.5 | 34.7 | 50.4 |
| RefineDet512 [33] | trainval35k | ResNet-101 | 36.4 | 57.5 | 39.5 | 16.6 | 39.9 | 51.4 |
| RefineDet320+ [33] | trainval35k | VGG-16 | 35.2 | 56.1 | 37.7 | 19.5 | 37.2 | 47.0 |
| RefineDet512+ [33] | trainval35k | VGG-16 | 37.6 | 58.7 | 40.8 | 22.7 | 40.3 | 48.3 |
| RefineDet320+ [33] | trainval35k | ResNet-101 | 38.6 | 59.9 | 41.7 | 21.1 | 41.7 | 52.3 |
| RefineDet512+ [33] | trainval35k | ResNet-101 | 41.8 | 62.9 | 45.7 | 25.6 | 45.1 | 54.1 |
| CornerNet512 [49] | trainval35k | Hourglass | 40.5 | 57.8 | 45.3 | 20.8 | 44.8 | 56.7 |
| NAS-FPN [16] | trainval35k | RetinaNet | 45.4 | - | - | - | - | - |
| NAS-FPN [16] | trainval35k | AmoebaNet | 48.0 | - | - | - | - | - |

only one category, while in ImageNet and PASCAL VOC all have more than 60% of images contain a single object category. As we know, small objects need more contextual reasoning to recognize. Images among MS COCO dataset are rich in contextual information. The biggest class is also the "person", nearly 800,000 instances, while the smallest class is "hair driver", about 600 instances in the whole dataset. Another small class is "hair brush" whose number is nearly 800. Except for 20 classes with many or few instances, the number of instances in the remaining 71 categories is roughly


(a) Iconic object images

(b) Iconic scene images

(c) Non-iconic images

FIGURE 5: The first two lines are examples from the MS COCO dataset [5]. The images show three different types of images sampled in the dataset, including iconic objects, iconic scenes and non-iconic objects. In addition, the last two lines are annotated sample images from the PASCAL VOC dataset [4].

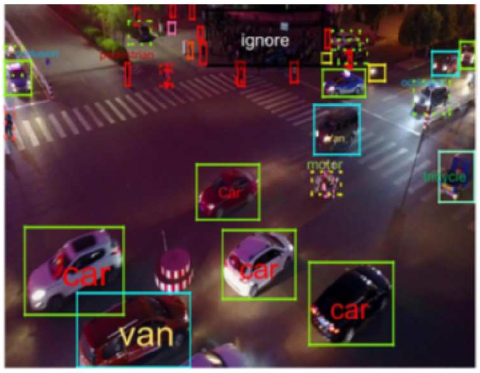


FIGURE 6: A drone-based image with bounding box and category labels of objects. Image from VisDrone 2018 dataset [50].

the same. Three typical categories of images in MS COCO dataset are showed in the first two lines of Fig. 5.

#### 2) Metric

MS COCO metric is under a strict manner and thoroughly judge the performance of detections. The threshold in PASCAL VOC is set to a single value, 0.5, but is belong to [0.5,0.95] with an interval 0.05 that is 10 values to calculate the mean average precision in MS COCO. Also the special average precision for small, medium and large objects are calculated separately.

### C. IMAGENET BENCHMARK

#### 1) Dataset

Challenging datasets can encourage a step forward of vision tasks and practical applications. Another important large-scale benchmark dataset is ImageNet dataset [3]. The ILSVRC task of object detection evaluates the ability of an algorithm to name and localize all instances of all target objects present in an image. ILSVRC2014 has 200 object classes and nearly 450k training images, 20k validation images and 40k test images. More comparisons with PASCAL VOC are in Table 3.

#### 2) Metric

The PASCAL VOC metric uses the threshold t = 0.5. However, for small objects even deviations of a few pixels would be unacceptable according to this threshold. ImageNet uses a loosen threshold calculated as:

$$t = min(0.5, \frac{wh}{(w+10)(h+10)})$$

TABLE 3: Comparison between ILSVRC object detection dataset and PASCAL VOC dataset

| Dataset | Classes | Fully annotated training images | Training objects | Val images | Val objects | Annotated obj/im |
|---|---|---|---|---|---|---|
| PASCAL VOC | 20 | 5717 | 13609 | 5823 | 15787 | 2.7 |
| ILSVRC | 200 | 60658 | 478807 | 20121 | 55501 | 2.8 |

where $w$ and $h$ are width and height of a ground truth box respectively. This threshold allow for the annotation to extend up to 5 pixels on average in each direction around the object.

### D. VISDRONE2018 BENCHMARK

Last year, a new dataset consists of images and videos captured by drones. VisDrone2018 [50], a large-scale visual object detection and tracking benchmark dataset, which is aiming at advancing visual understanding tasks on the drone platform. The images and video sequences in the benchmark were captured over various urban/suburban areas of 14 different cities across China from north to south. Specifically, VisDrone2018 consists of 263 video clips and 10,209 images (no overlap with video clips) with rich annotations, including object bounding boxes, object categories, occlusion, truncation ratios, etc. This benchmark has more than 2.5 million annotated instances in 179,264 images/video frames. Being the largest such dataset ever published, the benchmark enables extensive evaluation and investigation of visual analysis algorithms on the drone platform. VisDrone2018 has a large amount of small objects, such as dense cars, pedestrians and bicycles, which will cause difficult detection about certain categories. Moreover, a large proportion of the images in this dataset have more than 20 objects per image, 82.4% in training set, and the average number of objects per image is 54 in 6471 images of training set. This dataset contains dark night scenes so the brightness of these images lower than those in day time, which complicates the correct detection of small and dense objects, as shown in Fig. 6. This dataset adopts MS COCO metric.

### E. OPEN IMAGES V5

#### 1) Dataset

Open Images [51] is a dataset of 9.2M images annotated with image-level labels, object bounding boxes, object segmentation masks, and visual relationships. Open Images V5 contains a total of 16M bounding boxes for 600 object classes on 1.9M images, which makes it the largest existing dataset with object location annotations. First, the boxes in this dataset have been largely manually drawn by professional annotators (Google-internal annotators) to ensure accuracy and consistency. Second, the images in it are very diverse and mostly contain complex scenes with several objects (8.3 per image on average). Third, this dataset offers visual relationship annotations, indicating pairs of objects in particular relations (e.g. "woman playing guitar", "beer on table"). In total it has 329 relationship triplets with 391,073 samples. Fourth, V5 provides segmentation masks for 2.8M object instances in 350 classes. Segmentation masks mark the outline of objects, which characterizes their spatial extent to a much higher level of detail. Finally, the dataset is annotated with 36.5M image-level labels spanning 19,969 classes.

#### 2) Metric

On the basis of PASCAL VOC 2012 mAP evaluation metric, Kuznetsova *et al.*propose several modifications to consider thoroughly of some important aspects of the Open Images Dataset. First, for fair evaluation, the unannotated classes are ignored for avoiding wrongly counted as false negatives. Second, if an object belongs to a class and a subclass, an object detection model should give a detection result for each of the relevant classes. The absence of one of these classes would be considered a false positive in that class. Third, in Open Images Dataset, there exists group-of boxes which contain a group of (more than one which are occluding each other or physically touching) object instances but unknown a single object localization inside them. If a detection inside a group-of box and the intersection of the detection and the box divided by the area of the detection is larger than 0.5, the detection will be counted as a true positive. Multiple correct detections inside the same group-of box only count one valid true positive.

### F. PEDESTRIAN DETECTION DATASETS

Table 4 and table 5 list the comparison between several people detection benchmarks and pedestrian detection datasets, respectively.

## V. ANALYSIS OF GENERAL IMAGE OBJECT DETECTION METHODS

Deep neural network based object detection pipelines have four steps in general, image pre-processing, feature extracting, classification and regression, post-processing. Firstly, raw images from the dataset can't be fed into the network directly. Thus, we need to resize them to any special sizes and make them clearer, such as enhancing brightness, color, contrast, etc. Data augmentation is also available for some requirements, such as flipping, rotation, scaling, cropping, translation, adding Gaussian noise. Furthermore, GANs [59] (generative adversarial networks) can generate new images as you want to enrich diversity of inputs. For more details about data augmentation, please refer to [60] for more details. Secondly, feature extracting is a key step for further detection. The feature quality directly determines the upper bound of subsequent tasks which contain classification and regression. Thirdly, the detector head is responsible for proposing and refining bounding box concluding classification scores and bounding box coordinates. Fig. 1 illustrates the basic procedure of the second and the third step. At last,

TABLE 4: Comparison of person detection benchmarks,* Images in EuroCity Persons benchmark have day and night collections, which use "/" to split the number of day and night. Table information from Markus Braun *et al*. IEEE TPAMI2019 [52]

| Dataset | countries | cities | seasons | images | pedestrians | resolution | weather | train-cal-test-split(%) |
|---|---|---|---|---|---|---|---|---|
| Caltech [1] | 1 | 1 | 1 | 249884 | 289395 | 640 × 480 | dry | 50-0-50 |
| KITTI [2] | 1 | 1 | 1 | 14999 | 9400 | 1240 × 376 | dry | 50-0-50 |
| CityPersons [53] | 3 | 27 | 3 | 5000 | 31514 | 2048 × 1024 | dry | 60-10-30 |
| TDC [54] | 1 | 1 | 1 | 14674 | 8919 | 2048 × 1024 | dry | 71-8-21 |
| EuroCity Persons [52] | 12 | 31 | 4 | 40217/7118* | 183004/35309* | 1920 × 1024 | dry, wet | 60-10-30 |

TABLE 5: Comparison of pedestrian detection datasets. The 3rd, 4th, 5th are training set. The 6th, 7th, 8th are test set. Table information from Piotr *et al.* IEEE TPAMI2012 [1]

| Dataset | imaging setup | pedestrians | neg. images | pos. images | pedestrians | neg. images | pos. images |
|---|---|---|---|---|---|---|---|
| Caltech [1] | mobile | 192k | 61k | 67k | 155k | 56k | 65k |
| INRIA [55] | photo | 1208 | 1218 | 614 | 566 | 453 | 288 |
| ETH [56] | mobile | 2388 | - | 499 | 12k | - | 1804 |
| TUD-Brussels [57] | mobile | 1776 | 218 | 1092 | 1498 | - | 508 |
| Daimler-DB [58] | mobile | 192k | 61k | 67k | 155k | 56k | 65k |

the post-processing step deletes any weak detecting results. For example, NMS is a widely used method in which the highest scoring object deletes its nearby objects with inferior classification scores.

To obtain precise detection results, there exists several ways of which one can be used alone or in combination with the other.

### A. ENHANCED FEATURES

Extracting effective features from input images is a vital prerequisite for further accurate classification and localization steps. To fully utilize the output feature maps of consecutive backbone layers, Lin *et al.* [13] aim to extract richer features by dividing them into different levels to detect objects of different sizes, as shown in Fig. 3 (d). Some works [9] [31] [61] [62] utilize FPN as their multi-level feature pyramid backbone. Furthermore, a series of improved FPN [16] [37] [63] enriching features for detecting task. Kim *et al.* [64] propose a parallel feature pyramid (FP) network (PFPNet), where the FP is constructed by widening the network width instead of increasing the network depth. The additional feature transformation operation is to generate a pool of feature maps with different sizes, which yields the feature maps with similar levels of semantic abstraction across the scales. Li *et al.* [65] concatenate features from different layers with different scales and then generates new feature pyramid to feed into multibox detectors predicting the final detection results. Chen *et al.* [66] propose WeaveNet iteratively weaves context information from adjacent scales together to enable more sophisticated context reasoning. Zheng *et al.* [67] extend better context information for the shallow layers of one-stage detector [8].

Semantic relationships between different objects or regions of an image can help detect occluded and small objects. Bae *et al.* [68] utilize the combined and high-level semantic features for object classification and localization which is combining the multi-region features stage by stage. Zhang *et al.* [33] utilize a semantic segmentation branch and a global activation module to enrich the semantics of object detection features within a typical deep detector. Scene contextual relations [69] can provide some useful information for accurate visual recognition, Liu *et al.* [70] adopt scene contextual information to further improve accuracy. Modeling relations between objects can help object detection. Singh *et al.* [71] process context regions around the ground-truth object on an appropriate scale. Hu *et al.* [34] propose a relation module that processes a set of objects simultaneously considering both appearance and geometry features through interaction. Mid-level semantic properties of objects can benefit object detection containing visual attributes [72].

Attention mechanism is an effective method for networks focusing on the most significant region part. Some typical works [73] [74] [75] [76] [77] [78] [79] are focusing on attention mechanism so as to capture more useful features what detecting objects need. Kong *et al.* [80] design an architecture combining both global attention and local reconfigurations so as to gather task-oriented features across different spatial locations and scales.

Fully utilizing the effective region of one object can promote the accuracy. Original ConvNets can only focus on features of fixed square size (according to the kernal), thus the receptive field does not properly cover the entire pixel of a target object to represent it. The deformable ConvNets can produce deformable kernel and the offset from the initial convolution kernel (of fixed size) are learned from the networks. Deformable RoI Pooling can also adapt to part localization for objects with different shapes. In [35] [36], network weights and sampling locations jointly determine the effective support region.

Above all, richer and proper representations of an object can promote the detecting accuracy remarkably. Brain-inspired mechanism is a powerful way to further enhance detection performance.

### B. INCREASING LOCALIZATION ACCURACY

Localization and classification are two missions of object detection. Under object detection evaluation metrics, the precision of localization is a vital measurable indicator, thus increasing localization accuracy can promote detection performance remarkably. Designing a novel loss function to measure the accuracy of predicted boxes is an effective way to increase localization accuracy. Considering intersection over union is the most commonly used evaluation metric of object detection, estimating regression quality can judge the IoU between predicted bounding box and its corresponding assignment ground truth box. For two bounding boxes, IoU can be calculated as the intersection area divided by the union area.

$$IoU = \frac{bbox \cap gt}{bbox \cup gt}$$

A typical work [81] adopts IoU loss to measure the degree of accuracy the network predicting. This loss function is robust to varied shapes and scales of different objects and can converge well in a short time. Rezatofighi *et al.* [82] incorporate generalized IoU as a loss function and a new metric into existing object detection pipeline which makes a consistent improvement than the original smooth L1 loss counterpart. Tychsen *et al.* [47] adopt a novel bounding box regression loss for localization branch. IoU loss in this research considers the intersection over union between predicted box and assigned ground truth box which is higher than a preset threshold but not concludes only the highest one. He *et al.* [83] propose a novel bounding box regression loss for learning bounding box localization and transformation variance together. He *et al.* [84] propose a novel bounding box regression loss which has a strong connection to localization accuracy. Pang *et al.* [63] propose a novel balanced L1 Loss to further improving localization accuracy. Cabriel *et al.* [85] propose Axially Localized Detection method to achieve a very high localization precision at the cellular level.

In general, researchers design new loss function of localization branch to make the retained predictions more accurate.

### C. SOLVING NEGATIVES-POSITIVES IMBALANCE ISSUE

The two-stage detectors have a mainly well designed step that is the first stage producing proposals and filtering out a large number of negative samples. When feed into the detector the proposal bounding boxes belong to a sparse set. However, in a one-stage detector, the network has no steps to filter out bad samples, thus the dense sample sets are difficult to train. The proportion of positive and negative samples is extremely unbalanced as well. The typical solution is hard negative mining [86] The popularized hard mining methods OHEM [38] can help driving the focus towards hard samples. Liu *et al.* [8] adopt hard negative mining method which sorts all of the negative samples using the highest confidence loss for each pre-defined boxes and picking the top ones to make the ratio between the negative and positive samples at most 3:1. Considering hard samples is more effective to improve the detection performance when training an object detector. Pang *et al.* [63] propose a novel hard mining method called IoU-balanced sampling. Yu *et al.* [87] concentrate on real-time requirements.

Another effective way is adding some items in classification loss function. Lin *et al.* [31] propose a loss function, called focal loss, which can down-weight the loss assigned to well-classified or easy examples, focusing on the hard training examples and avoiding the vast number of easy negative examples overwhelming the detector during training. Chen *et al.* [88] consider design a novel ranking task to replace the conventional classification task and a newly Average-Precision loss for this task, which can alleviate the extreme negative-positive class imbalance issue remarkably.

### D. IMPROVING POST-PROCESSING NMS METHODS

Only one detected object can be successfully matched to a ground truth object which will be preserved as a result, while others matched to it are classified as duplicate. NMS (non-maximum suppression) is a heuristic method which selects only the object of the highest classification score otherwise will be ignored. Hu *et al.* [34] can use its intermediate results produced by relation module to better determine which objects will be saved while it doesn't need NMS. NMS considers the classification score but the localization confidence is absent, which causes less accurate in deleting weak results. Jiang *et al.* [89] propose IoU-Net learning to predict the IoU between each detected bounding box and the matched ground-truth. Because of its consideration of localization confidence, it improves the NMS method by preserving accurately localized bounding boxes. Tychsen *et al.* [47] propose a novel fitness NMS method which considers both greater estimated IoU overlap and classification score of predicted bounding boxes. Liu *et al.* [90] propose adaptive-NMS which applies a dynamic suppression threshold to an instance decided by the target density. Bodla *et al.* [44] propose an improved NMS method without any extra training and is simple to implement. He *et al.* [84] further improve soft-NMS method. Jan *et al.* [91] feed network score maps resulting from NMS at multiple IoU thresholds. Hosang *et al.* [92] design a novel ConvNets which does NMS directly without a subsequent post-processing step. Yu *et al.* [87] utilize the final feature map to filter out easy samples so the network concentrates on hard samples.

### E. COMBINING ONE-STAGE AND TWO-STAGE DETECTORS TO MAKE GOOD RESULTS

In general, pre-existing object detectors are divided into two categories, the one is two-stage detector, the representative one, [6]. The other is one-stage detector, such as [7], [8]. Two-stage detectors have high localization and object recognition precision, while the one-stage detectors achieve high inference and test speed. The two stages of two-stage detectors are divided by ROI (Region of Interest) pooling layer. In Faster R-CNN detector, the first stage, called RPN, a Region Proposal Network, proposes candidate object bounding boxes. The second stage, the network extracts features using RoIPool from each candidate box and performs classification and bounding-box regression.

To fully inherit the advantages of one-stage and two-stage detectors while overcoming their disadvantages, Zhang *et al.* [33] propose a novel RefineDet which achieves better accuracy than two-stage detectors and maintains comparable efficiency of one-stage detectors.

### F. COMPLICATED SCENE SOLUTIONS

Object detection always meets some challenges like small objects hard to detect and heavy occluded situation. Due to low resolution and noisy representation, detecting small objects is a very hard problem. Object detection pipelines [8] [31] detect small objects through learning representations of the objects at multiple scales. Some works [93] [94] [95] improve detection accuracy on the basis of [8]. Li *et al.* [96] utilize GAN model in which generator transfer perceived poor representations of the small objects to super-resolved ones that are similar enough to real large objects to fool a competing discriminator. This makes the representation of small objects similar to the large one thus improves accuracy without heavy computing cost. Some methods [45] [97] improve detection accuracy of small objects by enhancing IoU thresholds to train multiple localization modules. Hu *et al.* [98] utilize feature fusion to better detect small faces which is produced by image pyramid. Xu *et al.* [99] fuse high level features with rich semantic information and low level features via Deconvolution Fusion Block to enhance representation of small objects.

Target occlusion is another difficult problem in the field of object detection. Wang *et al.* [100] improve the recall of the face detection problem in the occluded case without speed decay. Wang *et al.* [101] propose a novel bounding box regression loss specifically designed for crowd scenes, called repulsion loss. Zhang *et al.* [102] propose a newly designed occlusion-aware R-CNN (OR-CNN) to improve the detection accuracy in the crowd. Baqu *et al.* [103] combine Convolutional Neural Nets and Conditional Random Fields that model potential occlusions.

As for the size of different objects in a dataset varies greatly, to address it, there are three commonly used methods. Firstly, input images are resized at multiple specified scales and feature maps are computed for each scale, called multi-scale training. Typical examples [27] [104] [105] [46] use this method. Singh *et al.* [71] adaptively sample regions from multiple scales of an image pyramid, conditioned on the image content. Secondly, researchers use convolutional filters of multiple scales on the feature maps. For instance, in [106], models of different aspect ratios are trained separately using different filter sizes (such as $5 \times 7$ and $7 \times 5$ ). Thirdly, pre-defined anchors with multi-scales and multiple aspect ratios are reference boxes of the predicted bounding boxes. Faster R-CNN [6] and SSD [8] are used in two-stage and one-stage detectors for the first time, respectively. Fig. 7 is a schematic diagram of the above three cases.

### G. ANCHOR-FREE

While there are constellation anchor-based object detectors being mainstream method which contain both one-stage and two-stage detectors making significant performance improvements, such as SSD, Faster R-CNN, YOLOv2, YOLOv3, they still suffer some drawbacks.

(1) The pre-defined anchor boxes have a set of hand-crafted scales and aspect ratios which are sensitive to dataset and affect the detection performance by a large margin.

(2) The scales and aspect ratios of pre-defined anchor boxes are kept fixed during training, thus the next step can't get adaptively adjust boxes. Meanwhile, detectors have trouble handling objects of all sizes.

(3) For densely place anchor boxes to achieve high recall, especially on large-scale dataset, the computation cost and memory requirements bring huge overhead during processing procedure.

(4) Most of pre-defined anchors are negative samples, which causes great imbalance between positive and negative sample during training.

To address that, recently propose a series of anchor-free methods [61] [62] [49] [107] [108] [109] [110] [111] [112] [113]. CenterNet [108] localizes the center point, top-left and bottom-right point of an object. Tian *et al.* [61] propose a localization method which is based on the four distance values between the predicted center point and four sides of a bounding box. The general structure of the anchor-based approach is shown in Fig. 8. It is still a novel direction for further research.

### H. TRAINING FROM SCRATCH

Almost all of state-of-the-art detectors utilize off-the-shelf classification backbone pre-trained on large scale classification dataset [3] as their initial parameter set then fine-tune parameters to adapt to the new detection task. Another way to implement training procedure is that all parameters are assigned from scratch. Zhu *et al.* [114] train detector from scratch thus don't need pre-trained classification backbone because of stable and predictable gradient brought by batch normalization operation. Some works [115] [116] [117] [118] train object detectors from scratch by dense layer-wise connections.

### I. DESIGNING NEW ARCHITECTURE

Because of different propose of classification and localization task, there exists a gap between classification network and detection architecture. Localization needs fine-grained representations of objects while classification needs high semantic information. Li *et al.* [14] propose a newly designed object detection architecture to specially focus on detection task

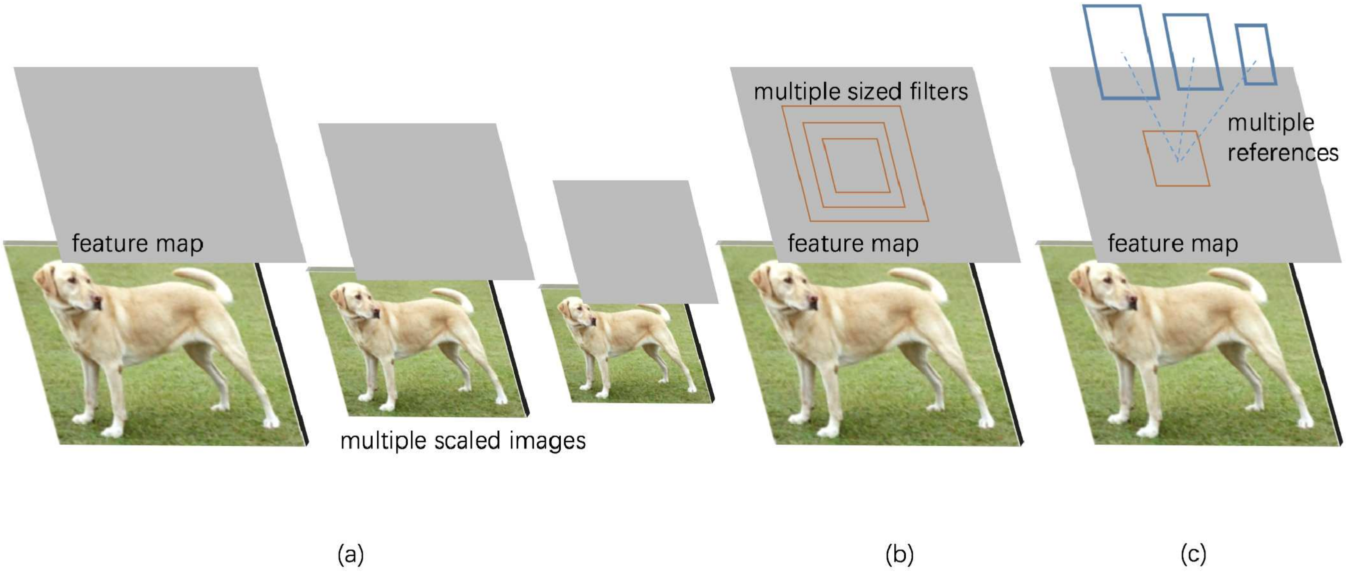


FIGURE 7: To meet various scales of objects issue, there are three ways. (a) multiple scaled images detector trains each of them. (b) multiple sized filters separately act on the same sized image. (c) multiple pre-defined boxes are the reference of predicted boxes.

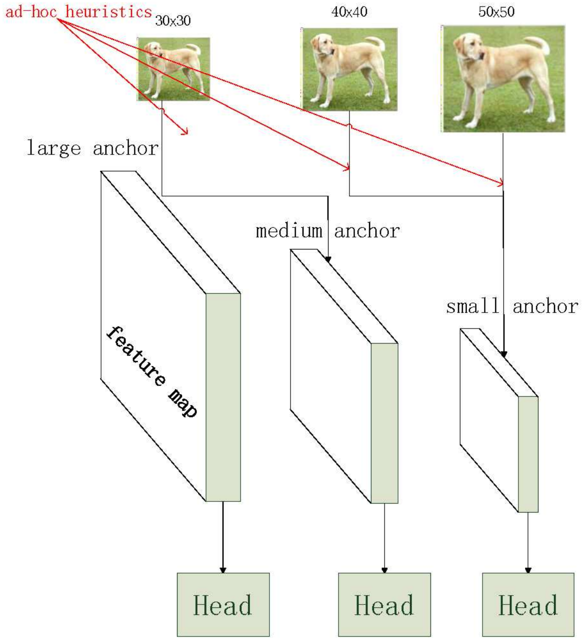


FIGURE 8: An anchor-based architecture require heuristics to determine which size level anchors are responsible for what scale range of objects.

which maintains high spatial resolution in deeper layers and doesn't need to pre-train on large scale classification dataset.

The two-stage detectors always slower than one-stage detectors. By studying the structure of the two-stage network, researchers find two-stage detectors like Faster R-CNN and R-FCN have a heavy head which slows it down. Li *et al.* [119] propose a light head two-stage detector to keep time efficiency.

### *J. SPEEDING UP DETECTION*

For limited computing power and memory resource such as mobile devices, real-time devices, webcam, automatic driving encourage studies on efficient detection architecture design. The most typical real-time detector is the [7] [28] [30] series and [8] [32] and their improved architecture [66] [67] [95] [120]. Some methods [22] [87] [121] [122] [123] [124] are aim to reach real-time detecting effect recently.

### *K. ACHIEVING FAST AND ACCURATE DETECTIONS*

The best object detector needs both high efficiency and high accuracy which is the ultimate goal of this task. Lin *et al.* [31] aim to surpass the accuracy of existing two-stage detectors as well as maintain fast speed. Zhou *et al.* [125] combine an accurate (but slow) detector and a fast (but less accurate) detector adaptively determining whether an image is easy or hard to detect and choosing an appropriate detector to detect it. Liu *et al.* [126] build a fast and accurate detector by strengthening lightweight network features using receptive fields block.

## VI. APPLICATIONS AND BRANCHES

### *A. TYPICAL APPLICATION AREAS*

Object detection has been widely used in some fields to assist people to complete some tasks, such as security field, military field, transportation field, medical field and life field etc. We describe the typical and recent methods utilized in these fields in detail.

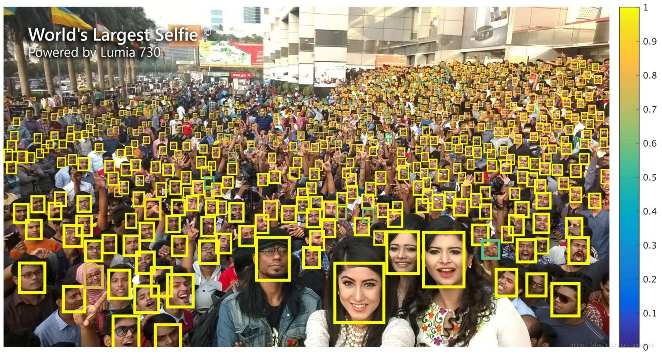


FIGURE 9: A challenging densely tiny human faces detection results. Image from Hu *et al.* [98].

### 1) Security field

The most well known applications in the security field are face detection, pedestrian detection, fingerprint identification, fraud detection, anomaly detection etc.

Face detection aims at detecting people faces in an image, as shown in Fig. 9. Because of extreme poses, illumination and resolution variations, face detection is still a difficult mission. Many works focus on precise detector designing. Ranjan *et al.* [127] learn correlated tasks (face detection, facial landmarks localization, head pose estimation and gender recognition) simultaneously to boost the performance of individual tasks. He *et al.* [128] propose a novel Wasserstein convolutional neural network approach for learning invariant features between near-infrared (NIR) and visual (VIS) face images. Designing appropriate loss functions can enhance discriminative power of DCNNs based large-scale face recognition. The cosine-based softmax losses [129] [130] [131] [132] achieve great success in deep learning based face recognition. Deng *et al.* [133] propose an Additive Angular Margin Loss (ArcFace) to get highly discriminative features for face recognition. Please refer to [134] for more details.

Pedestrian detection focuses on detecting pedestrians in the natural scenes. Braun *et al.* [52] release an EuroCity Persons dataset containing pedestrians, cyclists and other riders in urban traffic scenes. Complexity-aware cascaded pedestrian detectors [135] [136] [137] focus on real time pedestrian detection. Please refer to a survey [138] for more details.

Anomaly detection plays an significant role in fraud detection, climate analysis, and healthcare monitoring. Existing anomaly detection techniques [139] [140] [141] [142] analyze the data on a point-wise basis. To point the expert analysts to the interesting regions (anomalies) of the data, Barz *et al.* [143] propose a novel unsupervised method called “Maximally Divergent Intervals” (MDI), which searches for contiguous intervals of time and regions in space.

### 2) Military field

In military field, remote sensing object detection, topographic survey, flyer detection, etc. are representative applications.

Remote sensing object detection aims at detecting objects on remote sensing images or videos, which meets some challenges. Firstly, the extreme large input size but small targets makes the existing object detection procedure too slow for practical use and too hard to detect. Secondly, the massive and complex backgrounds cause serious false detection. To address these issues, researchers adopt data fusion and focus on detecting small objects for their less information and small deviation causing huge inaccuracy. Remote sensing images have some characteristics far from natural images, thus those strong pipelines such as Faster R-CNN, FCN, SSD, YOLO etc. can’t transfer well in the new data domain. Designing remote sensing dataset adapted detectors remains a research hot spot in this domain.

Cheng *et al.* [144] propose a CNN-based Remote Sensing Image (RSI) object detection model dealing with the rotation problem by proposing a rotation-invariant layer. Zhang *et al.* [145] propose a rotation and scaling robust structure to address lacking rotation and scaling invariance in RSI object detection domain. Li *et al.* [146] propose a rotatable region proposal network and a rotatable detection network considering the orientation of vehicles. Deng *et al.* [147] propose an accurate-vehicle-proposal-network (AVPN) for small object detection. Audebert *et al.* [148] utilize accurate semantic segmentation results to obtain detection of vehicles. Li *et al.* [149] address large range of resolutions of ships (ranging from dozens of pixels to thousands) issue in ship detection. Pang *et al.* [150] propose a real-time remote sensing method. Pei *et al.* [151] propose a deep learning framework on synthetic aperture radar (SAR) automatic target recognition. Long *et al.* [152] concentrate on automatic accurate localization of detected objects. Shahzad *et al.* [153] propose a novel framework containing automatic labeling and recurrent neural network for detection.

Typical methods [154] [155] [156] [157] [158] [159] all utilize deep neural networks to achieve detection missions on remote sensing datasets. NWPU VHR-10 [160], HRRSD [145], DOTA [161], DLR 3K Munich [162] and VEDAI [163] are remote sensing object detection benchmarks. We recommend readers refer to [164] for more details on remote sensing object detection domain.

### 3) Transportation field

As we known that, license plate recognition, automatic driving and traffic sign recognition etc. greatly facilitate people’s life.

With the widespread use of vehicles, license plate recognition is required in tracking crime, residential access, traffic violations tracking etc. edge information, mathematical morphology, texture features, sliding concentric windows, connected component analysis etc. can bring license plate recognition system more robust and stable. Recently, deep learning-based methods [165] [166] [167] [168] [169] provide a variety of solutions for license plate recognition. Please refer to [170] for more details.

An autonomous vehicle (AV) needs an accurate perception of its surroundings to operate reliably. The perception system of an AV normally employs machine learning (e.g., deep learning) and transforms sensory data into semantic information that enables autonomous driving. Object detection is a fundamental function of this perception system. 3D object detection methods involve a third dimension that reveals more detailed object's size and location information, which are divided into three categories, monocular, point-cloud and fusion. First, monocular image based methods predict 2D bounding boxes on the image then extrapolate them to 3D, which lacks explicit depth information so limits the accuracy of localization. Second, point-cloud based methods project point clouds into a 2D image to process or generate a 3D representation of the point cloud directly in a voxel structure, where the former loses information and the latter is time consuming. Third, fusion based methods fuse both front view images and point clouds to generate a robust detection, which represent state-of-the-art detectors while computationally expensive. Recently, Lu *et al.* [171] utilize a novel architecture contains 3D convolutions and RNNs to achieve centimeter-level localization accuracy in different real-world driving scenarios. Song *et al.* [172] release a 3D car instance understanding benchmark for autonomous driving. Banerjee *et al.* [173] utilize sensor fusion to obtain better features. Please refer to an recently survey [174] for more details.

Both unmanned vehicles and autonomous driving systems require solving the problem of traffic sign recognition. For the sake of safety and obeying the rules, real-time accurate traffic sign recognition assists in driving by acquiring the temporal and spatial information of the potential signs. Deep learning methods [175] [176] [177] [178] [179] [180] [181] solve this problem with high performance.

#### 4) Medical field

In medical field, medical image detection, cancer detection, disease detection, skin disease detection and healthcare monitoring etc. have become a means of supplementary medical treatments increasingly. Computer Aided Diagnosis (CAD) systems can support physicians in classifying different kinds of cancer. In detail, after an appropriate acquisition of the images, the fundamental steps carried out by a CAD framework can be identified as image segmentation, feature extraction, classification and object detection. Due to significant individual differences, data scarcity and privacy, there usually exists data distribution difference between source domain and target domain. A domain adaptation framework [182] is needed for medical image detection.

Li *et al.* [77] incorporate the attention mechanism in CNN for glaucoma detection and establish a large-scale attention-based glaucoma dataset. Liu *et al.* [183] design a bidirectional recurrent neural network (RNN) with long short-term memory (LSTM) to detect DNA modifications called DeepMod. Schubert *et al.* [184] propose cellular morphology neural networks (CMNs) for automated neuron reconstruction and automated detection of synapses. Codella *et al.* [185] organize a challenge of skin lesion analysis toward melanoma detection. Please refer to [186] for more details.

#### 5) Life field

In life field, intelligent home, commodity detection, event detection, pattern detection, rain/shadow detection etc. are the most representative applications.

On densely packed scenes like retail shelf displays, Eran Goldman *et al.* [187] propose a novel precise object detector and a new SKU-110K dataset to meet this challenge.

Event detection aims to discover real-world events from the Internet such as festivals, talks, protests, natural disasters, elections etc. with the popularity of social media and its new characters, the data type of which are more diverse than before. Multi-domain event detection (MED) provides comprehensive descriptions of events. Yang *et al.* [188] present an event detection framework to dispose multi-domain data. Wang *et al.* [189] incorporate online social interaction features by constructing affinity graphs for event detection tasks. Schinas *et al.* [190] propose a multimodal graph-based system to detect events from 100 million photos/videos. Please refer to a survey [191] for more details.

Pattern detection always meet some challenges such as, scene occlusion, pose variation, varying illumination and sensor noise. To better address repeated pattern or periodic structure detection, researches propose strong baseline in both 2D images [192] [193] and 3D point clouds [194] [195] [196] [197] [198] [199] [200] [201] [202] [203] [204] [205].

Yang *et al.* [206] present a novel rain model accompany with a deep learning architecture to address rain detection in a single image. Hu *et al.* [207] analyze the spatial image context in a direction-aware manner and design a novel deep neural network to detect shadow.

### B. OBJECT DETECTION BRANCHES

Object detection has a wide range of application scenarios. The research of this domain contains a large variety of branches. We describe some representative branches in this part.

#### 1) Weakly supervised object detection

Weakly supervised object detection (WSOD) aims at utilizing a few fully annotated images (supervision) to detect the large amount of non-fully annotated ones. Traditionally models are learned from images labelled only with the object class and not the object bounding box. Annotating a bounding box for each object in large datasets is expensive, laborious and impractical. Weakly supervised learning relies on incomplete annotated training data to learn detection models. Weakly supervised deep detection networks [208] is a representative work for weakly supervised object detection. Context information [209], instance classifier refinement [210] and image segmentation [211] [212] are adopted to tackle hardly optimized problems. Yang *et al.* [213] show that the action depicted in the image could provide strong cues about the location of the associated object. Wan *et al.* [214] propose a min-entropy latent model optimized with a recurrent learning algorithm for weakly supervised object detection. Tang *et al.* [215] utilize an iterative procedure to generate proposal clusters and learn refined instance classifiers, which makes the network concentrate on the whole object rather than parts of the object. Cao *et al.* [216] propose a novel feedback convolutional neural network for weakly supervised object localization. Wan *et al.* [217] propose continuation multiple instance learning to alleviate the non-convexity problem in WSOD.

### 2) Salient object detection

Salient object detection utilizes deep neural network to predict saliency scores of image regions and obtain accurate saliency maps, as shown in Fig. 10. Salient object detection networks usually need to aggregate multi-level features of backbone network. For fast speed without accuracy dropping, Wu *et al.* [218] present that discarding the shallower layer features can achieve fast speed and the deeper layer features are sufficient to obtain precisely salient map. Liu *et al.* [219] expand the role of pooling in convolutional neural networks. Wang *et al.* [220] utilize fixation prediction to detect salient objects. Wang *et al.* [221] utilize recurrent fully convolutional networks and incorporate saliency prior knowledge for accurate salient object detection. Feng *et al.* [222] propose an attentive feedback module to better explore the structure of objects. Video salient object detection datasets [223] [224] [225] [226] [227] [228] [229] provide benchmarks for video salient object detection, and existing good algorithms [230] [231] [224] [232] [227] [233] [234] [235] [236] [237] [238] [239] [240] [241] devote in this domain.

### 3) Highlight detection

Highlight detection is to retrieve a moment in a short video clip that captures a user's primary attention or interest, which can accelerate browsing many videos, enhance social video sharing and facilitate video recommendation. Typically the highlight detectors [242] [243] [244] [245] [246] [247] are domain-specific for they are tailored to a category of videos. All object detection tasks require a large amount of manual annotation data, highlight detection is no exception. Xiong *et al.* [248] propose a weakly supervised method on shorter user-generated videos to address this issue.

### 4) Edge detection

Edge detection aims at extracting object boundaries and perceptually salient edges from images, which is important to a series of higher level vision tasks like segmentation, object detection and recognition. Edge detection meets some challenges. First, a variety of scale of edges in an image which needs both object-level boundaries and useful local region details. Second, Conv layers of different levels are specialized to predict different parts of the final detection, thus each layer in CNN shall be trained by proper layer-specific supervision. To address these issues, He *et al.* [249] propose a Bi-Directional Cascade Network to let one layer supervised by labeled edges while adopt dilated convolution to generate multi-scale features. Liu *et al.* [250] propose an accurate edge detector which utilizes richer convolutional features.

### 5) Text detection

Text detection aims at identifying text regions of given images or videos which is also an important prerequisite for many computer vision tasks, such as classification, video analysis etc. There have been many successful commercial optical character recognition (OCR) systems for internet content and documentary texts recognition. The detection of text in natural scenes is still a challenge due to complex situations such as blurring, uneven lighting, perspective distortion, various orientation, etc. Some typical works [251] [252] [253] focus on horizontal or nearly horizontal text detection. Recently, researchers find that arbitrary-oriented text detection [254] [255] [256] [257] [258] is a problem that needs to be solved. In general, deep learning based scene text detection methods can be classified into two categories. The first category takes scene text as a type of general object, following the general object detection paradigm and locating scene text by text box regression. These methods have difficulties to deal with the large aspect ratios and arbitrary-orientation of scene text. The second one directly segments text regions, but mostly requires complicated post-processing step. Usually, some methods in this category mainly involve two steps, segmentation (generate text prediction maps) and geometric approaches (for inclined proposals), which is time-consuming. In addition, in order to obtain the desired orientation of text boxes, some methods require complex post-processing step, so it's not as efficient as those architectures that are directly based on detection networks.

Lyu *et al.* [257] combine the ideas of the two categories above while avoiding their shortcomings by localizing corner points of text bounding boxes and segmenting text regions in relative positions to detect scene text, which can handle long oriented text and only need a simple NMS post-processing step. Ma *et al.* [258] develop a novel rotation-based approach and an end-to-end text detection system in which Rotation Region Proposal Networks (RRPN) for generating inclined proposals with text orientation angle information.

### 6) Multi-domain object detection

Domain-specific detectors always achieve high detection performance on the specified dataset. So as to get a universal detector which is capable of working on various image domains, recently many works are focus on training a multi-domain detector while don't require prior knowledge of the newly domain of interest. Wang *et al.* [259] propose a universal detector which utilizes a new domain-attention mechanism working on a variety of image domains (human faces, traffic signs and medical CT images) without prior knowledge of the domain of interest. Wang *et al.* [259] propose a newly established universal object detection bench-

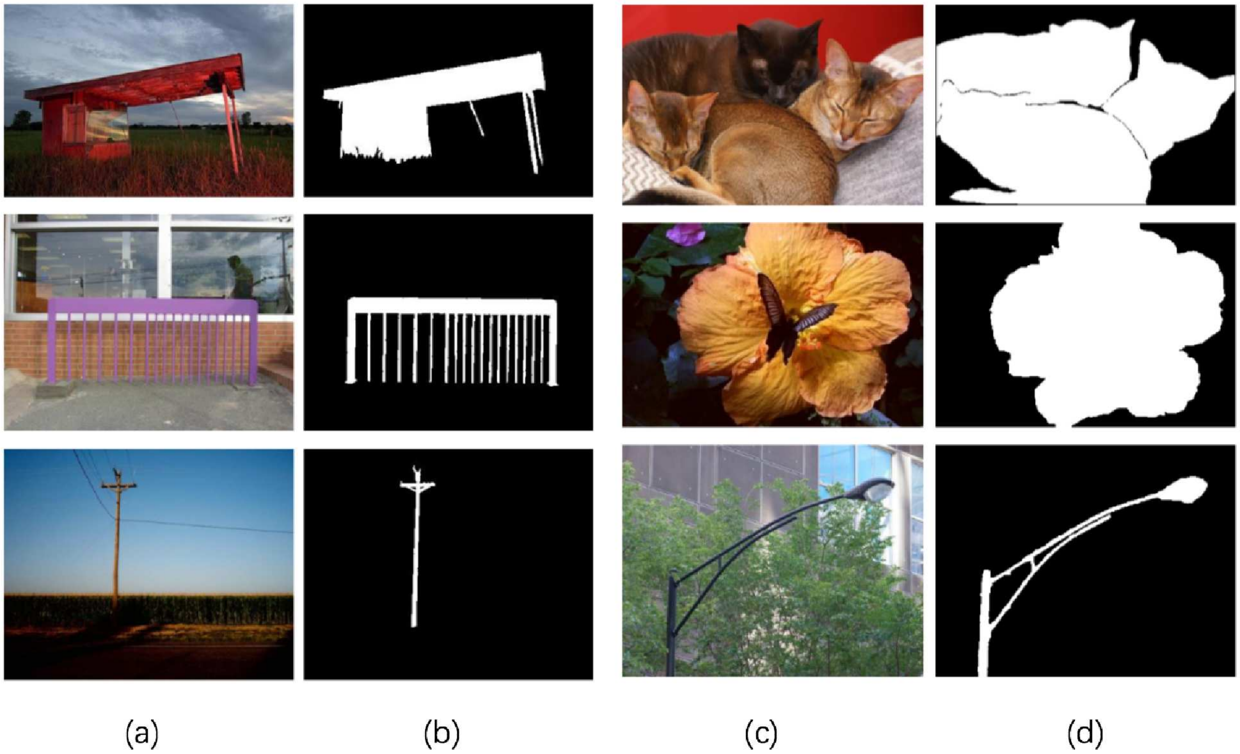


FIGURE 10: Some examples from the salient object detection datasets. (a), (c) are images, (b), (d) are ground truth.

mark consisting of 11 diverse datasets to better meet the challenges of generalization in different domains.

For learning a universal representation for vision, Bilen *et al.* [260] add domain-specific BN (batch normalization) layers to a multi-domain shared network. Rebuffi *et al.* [261] propose adapter residual modules which achieves a high degree of parameter sharing while maintaining or even improving the accuracy of domain-specific representations. Rebuffi *et al.* [261] introduce the Visual Decathlon Challenge, a benchmark contains ten very different visual domains. Inspired by the transfer learning, Rebuffi *et al.* [262] empirically study efficient parameterizations and outperform traditional fine-tuning techniques.

Another requirement for multi-domain object detection is reducing annotation costs. Object detection datasets need heavily annotation works which is time consuming and mechanical. Transferring pre-trained models from label-rich domains to label-poor datasets can solve label-poor detection works. One way is utilizing unsupervised domain adaptation methods to tackle the dataset bias problems. Recently researchers use adversarial learning to align the source and target distributions of samples. Chen *et al.* [263] utilize Faster R-CNN with a domain classifier trained to distinguish source and target samples, like adversarial learning, while the feature extractor learns to deceive the domain classifier. Saito *et al.* [264] propose a weak alignment model to focus on similarity between different images from domains with large discrepancy rather than aligning images that are globally dissimilar. When only in the source domain manual annotations are available, Unsupervised Domain Adaptation methods is to address this issue. Haupmann *et al.* [265] propose a Unsupervised Domain Adaptation method which models both the intra-class and the inter-class domain discrepancy.

### 7) Object detection in videos

Object detection in videos aims at detecting objects in videos, which brings additional challenges due to degraded image qualities such as motion blur and video defocus, leading to unstable classifications for the same object across video. Video detectors [266] [267] [268] [269] [270] [271] [272] [273] [274] [275] exploit temporal contexts to meet this challenge. Some static detectors [266] [267] [268] [269] first detect objects in each frame and then check them by linking detections of the same object in neighbor frames. Due to object motion, the same object in neighbor frames may not have a large overlap. On the other hand, the predicted object movements not accurate enough to link neighbor frames. Tang *et al.* [276] propose an architecture which links objects in the same frame instead of neighboring frames.

#### 8) Point clouds 3D object detection

Compared to image based detection, LiDAR point cloud provides reliable depth information that can be used to accurately localize objects and characterize their shapes. In autonomous navigation, autonomous driving, housekeeping robots and augmented/virtual reality applications, LiDAR point cloud based 3D object detection plays an important role. Point cloud based 3D object detection meets some challenges, for LiDAR point clouds are sparse, highly variable point density, non-uniform sampling of the 3D space, effective range of the sensors, occlusion, and the relative pose variation. Engelcke *et al.* [277] first propose sparse convolutional layers and L1 regularization for efficient large-scale processing of 3D data. Qi *et al.* [278] propose an end-to-end deep neural network called PointNet, which learns point-wise features directly from point clouds. Qi *et al.* [279] improve PointNet which learns local structures at different scales. Zhou *et al.* [280] close the gap between RPN and point set feature learning for 3D detection task. Zhou *et al.* [280] present a generic end-to-end 3D detection framework called VoxelNet, which learns a discriminative feature representation from point clouds and predicts accurate 3D bounding boxes simultaneously.

In autonomous driving application, Chen *et al.* [281] perform 3D object detection from a single monocular image. Chen *et al.* [282] take both LiDAR point cloud and RGB images as input and then predict oriented 3D bounding boxes for high-accuracy 3D object detection. Example 3D detection result is shown in Fig. 11.

#### 9) 2D, 3D pose detection

Human pose detection aims at estimating the 2D or 3D pose location of the body joints and defining pose classes then returning the average pose of the top scoring class, as shown in Fig. 12. Typical 2D human pose estimation methods [284] [285] [286] [287] [288] [289] [290] utilize deep CNN architectures. Rogez *et al.* [291] propose an end-to-end architecture for joint 2D and 3D human pose estimation in natural images which predicts 2D and 3D poses of multiple people simultaneously. Benefit by full-body 3D pose, it can recover body part locations in cases of occlusion between different targets. Human pose estimation approaches can be divided into two categories, one-stage and multi-stage methods. The best performing methods [292] [9] [293] [294] typically base on one-stage backbone networks. The most representative multi-stage methods are convolutional pose machine [295], Hourglass network [286] and MSPN [296].

## VII. CONCLUSIONS AND TRENDS

### A. CONCLUSIONS

Deep learning based object detection has been fast developed with the emergence of powerful computational devices. For deploying on more accurate applications, the need of high accuracy real-time system is more and more urgent. For achieving high accuracy and efficiency detectors is the ultimate goal of this task, researchers have developed a series of directions such as, constructing new architecture, extracting rich features, exploiting good representations, improving processing speed, training from scratch, anchor-free methods, solving sophisticated scene issues (small objects, occluded objects), combining one-stage and two-stage detectors to make good results, improving post-processing NMS method, solving negatives-positives imbalance issue, increasing localization accuracy, enhancing classification confidence and so on. With the increasingly powerful object detectors in security field, military field, transportation field, medical field and life field, the application of object detection is gradually extensive. In addition, a variety of branches in detection domain arise. Although the achievement of this domain has been effective recently, there is still much room for further development.

### B. TRENDS

#### 1) Combining one-stage and two-stage detectors

On the one hand, the two-stage detectors have a densely tailing process to obtain as many as reference boxes, which is time consuming and inefficient. To address this issue, researchers are required to eliminate so much redundancy while maintaining high accuracy. On the other hand, the one-stage detectors achieve fast processing speed which have been used successfully in real-time applications. Although fast, the lower accuracy is still a bottleneck for high precision requirements. How to combine the advantages of both one-stage and two-stage detectors is still a big challenge.

#### 2) Video object detection

In video object detection, motion blur, video defocus, motion target ambiguity, intense target movements, small targets, occlusion and truncation etc. bring this mission hard to achieve good performance in both the actual living scenes and remote sensing scenes. Delving into sports goals and more complex data such as video is one of the key points for future research.

#### 3) Efficient post-processing methods

In the three (for one-stage detectors) or four (for two-stage detectors) stage detection procedure, post-processing is an initial step for the final results. On most of the detection metrics, only the highest prediction result of one object can be send to the metric program to calculate accuracy score. The post-processing methods like NMS and its improvements may eliminate well localized but high classification confidence objects, which is detrimental to the accuracy of the measurement. More efficient and accurate post-processing method is another direction for object detection domain.

#### 4) Weakly supervised object detection methods

Utilizing high proportion labelled only with the object class but not the object bounding box images to replace a large amount of fully annotated images for training is high efficient and easy to get. Weakly supervised object detection (WSOD) aims at utilizing a few fully annotated images (supervision) to detect the large amount of non-fully annotated ones. Thus

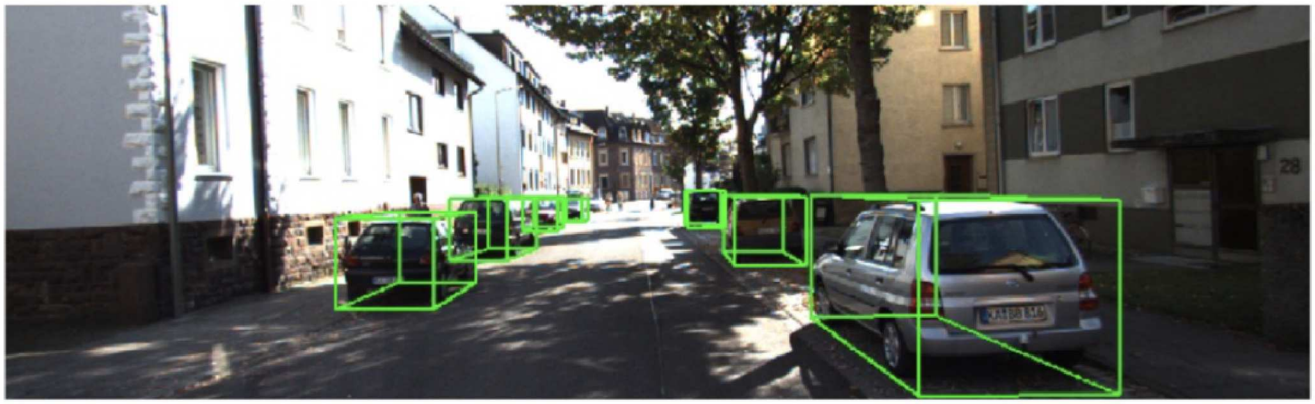


FIGURE 11: Example 3D detection result from the KITTI validation set projected onto an image. Image from Vishwanath A. Sindagi *et al.* [283].

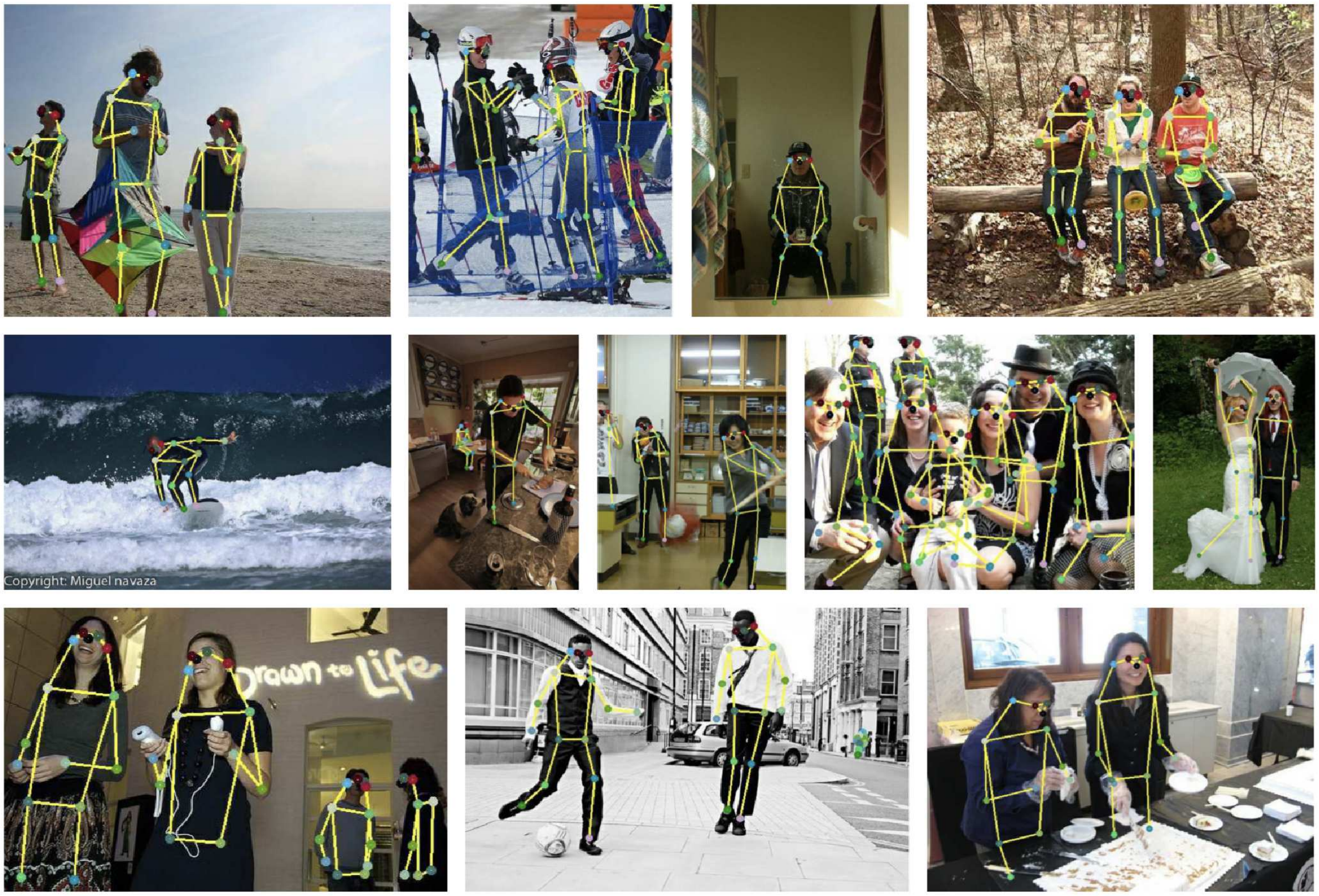


FIGURE 12: Some examples of multi-person pose estimation. Image from Chen *et al.* [292].

developing WSOD methods is a significant problem for further study.

### 5) Multi-domain object detection

Domain-specific detectors always achieve high detection performance on the specified dataset. So as to get a universal detector which is capable of working on various image domains, a multi-domain detector don't require prior knowledge of the newly domain of interest can address this problem. Domain transfer is a challenging mission for further study.

### 6) 3D object detection

With the advent of 3D sensors and diverse applications of 3D understanding, 3D object detection gradually becomes a hot research direction. Compared to 2D image based detection, LiDAR point cloud provides reliable depth information that can be used to accurately localize objects and characterize their shapes. LiDAR enables accurate localization of objects in the 3D space. Object detection techniques based on LiDAR data often outperform the 2D counterparts as well.

### 7) Salient object detection

Salient object detection (SOD) aims at highlighting salient object regions in images. Video object detection is to classify and locate objects of interest in a continuous scene. SOD is driven by and applied to a widely spectrum of object-level applications in various areas. Given salient object regions of interest in each frame can assist accurate object detecting in videos. Thus salient object detection is a vital previous process for high-level recognition tasks and challenging detection missions.

### 8) Unsupervised object detection

Supervised methods are time consuming and inefficient in training process, which need well annotated dataset used for supervision information. Annotating a bounding box for each object in large datasets is expensive, laborious and impractical. Developing automatic annotation technology to release human annotation work is a promising trend for unsupervised object detection. Unsupervised object detection is a future research direction for intelligent detection mission.

### 9) Multi-task learning

Aggregating multi-level features of backbone network is a significant way to enhance detection performance. Furthermore, performing multiple computer vision tasks simultaneously such as object detection, semantic segmentation, instance segmentation, edge detection, highlight detection etc. can enhance separate task performance by a large margin because of richer information. Adopting multi-task learning is a good way to aggregate multiple tasks in a network, and it presents great challenges to researchers to maintain processing speed and improve accuracy as well.

### 10) Multi-source information assistance

Due to the popularity of social media and the development of big data technology, multi-source information becomes easy to access. Many social media information can provide both pictures and descriptions of them in textual form, which can help detection task. Fusing multi-source information is an emerging research direction with the progress of various technologies.

### 11) Constructing terminal object detection system

From the cloud to the terminal, the terminalization of artificial intelligence can help people deal with mass information and solve problems better and faster. With the emergence of lightweight networks, terminalized detectors are developed into more efficient and reliable devices with broad application scenarios. The chip detection network based on FPGA will make real-time application possible.

### 12) Medical imaging and diagnosis

FDA (U.S. Food and Drug Administration) is promoting "AI is medical devices" and firstly approved AI software, IDx-DR, which is a diabetic retinopathy detector achieves higher than 87.4% precision, in April 2018. For customers, the combination of image recognition systems and mobile devices can make cell phone a powerful family diagnostic tool. This direction is full of challenges and expectations.

### 13) Advanced medical biometrics

Utilizing deep neural network, researchers began to study and measure atypical risk factors that had previously been difficult to quantify. Using neural networks to analyze retinal images and speech patterns may help identify the risk of heart disease. In the near future, medical biometrics will be used for passive monitoring.

### 14) Remote sensing airborne and real-time detection

Both military and agricultural fields require accurate analysis of remote sensing images. Automated detection software and integrated hardware will bring unprecedented development to these fields. Loading the deep learning based object detection system to the SoC (System on Chip) realizes the real-time high-altitude detection.

### 15) GAN based detector

Deep learning based systems always require large amounts of data for training, whereas Generative Adversarial Network is a powerful structure for generating fake images. How much you need, how much it can produce. Mixing the real world scene and simulated data generated by GAN trains object detector to make the detector grow more robust and obtain stronger generalization ability.

The research of object detection still needs further study. We hope that deep learning methods will make breakthroughs in the near future.

## ACKNOWLEDGMENT

Thanks to the scholars involved in this paper. This paper quotes the research literature of several scholars. Without the help and inspiration of the research results of all scholars, it would be difficult for me to complete the writing of this paper.

Thanks to my tutor and family for their guidance and support in the process of writing this article.

## Reference

[1] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, pp. 743–761, April 2012.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361, June 2012.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, pp. 211–252, Dec 2015.

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

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in Computer Vision – ECCV 2014 (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), (Cham), pp. 740–755, Springer International Publishing, 2014.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, pp. 1137–1149, June 2017.

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

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in Computer Vision – ECCV 2016 (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 21–37, Springer International Publishing, 2016.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988, Oct 2017.

[10] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," arXiv preprint arXiv:1901.06032, 2019.

[11] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," CoRR, vol. abs/1905.05055, 2019.

[12] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," arXiv preprint arXiv:1809.02165, 2018.

[13] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944, July 2017.

[14] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Detnet: A backbone network for object detection," arXiv preprint arXiv:1804.06215, 2018.

[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500, 2017.

[16] G. Ghiasi, T.-Y. Lin, R. Pang, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," arXiv preprint arXiv:1904.07392, 2019.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[18] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856, 2018.

[19] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258, 2017.

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520, 2018.

[22] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," in Advances in Neural Information Processing Systems, pp. 1963–1972, 2018.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, June 2016.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[25] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," Neural computation, vol. 29, no. 9, pp. 2352–2449, 2017.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, June 2014.

[27] R. Girshick, "Fast r-cnn," in 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448, Dec 2015.

[28] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517–6525, July 2017.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[30] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," CoRR, vol. abs/1804.02767, 2018.

[31] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007, Oct 2017.

[32] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," arXiv preprint arXiv:1701.06659, 2017.

[33] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4203–4212, 2018.

[34] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3588–3597, 2018.

[35] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in Proceedings of the IEEE international conference on computer vision, pp. 764–773, 2017.

[36] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," arXiv preprint arXiv:1811.11168, 2018.

[37] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2det: A single-shot object detector based on multi-level feature pyramid network," arXiv preprint arXiv:1811.04533, 2018.

[38] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 761–769, 2016.

[39] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2874–2883, 2016.

[40] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in Advances in neural information processing systems, pp. 379–387, 2016.

[41] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, and H. Lu, "Couplenet: Coupling global structure with local parts for object detection," in Proceedings of the IEEE International Conference on Computer Vision, pp. 4126–4134, 2017.

[42] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7310–7311, 2017.

[43] A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta, "Beyond skip connections: Top-down modulation for object detection," arXiv preprint arXiv:1612.06851, 2016.

[44] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in Proceedings of the IEEE International Conference on Computer Vision, pp. 5561–5569, 2017.

[45] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162, 2018.

[46] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3578–3587, 2018.

[47] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness nms and bounded iou loss," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6877–6885, 2018.
[48] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "Ron: Reverse connection with objectness prior networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5936–5944, 2017.
[49] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750, 2018.
[50] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," arXiv preprint arXiv:1804.07437, 2018.
[51] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," arXiv:1811.00982, 2018.
[52] M. Braun, S. Krebs, F. Flohr, and D. Gavrila, "Eurocity persons: A novel benchmark for person detection in traffic scenes," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019.
[53] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3221, 2017.
[54] X. Li, F. Flohr, Y. Yang, H. Xiong, M. Braun, S. Pan, K. Li, and D. M. Gavrila, "A new benchmark for vision-based cyclist detection," in 2016 IEEE Intelligent Vehicles Symposium (IV), pp. 1028–1033, IEEE, 2016.
[55] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in international Conference on computer vision & Pattern Recognition (CVPR'05), vol. 1, pp. 886–893, IEEE Computer Society, 2005.
[56] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8, IEEE, 2007.
[57] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 794–801, IEEE, 2009.
[58] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 12, pp. 2179–2195, 2008.
[59] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, pp. 2672–2680, 2014.
[60] B. Zoph, E. D. Cubuk, G. Ghiasi, T. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," CoRR, vol. abs/1906.11172, 2019.
[61] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," arXiv preprint arXiv:1904.01355, 2019.
[62] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Foveabox: Beyond anchor-based object detector," arXiv preprint arXiv:1904.03797, 2019.
[63] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," arXiv preprint arXiv:1904.02701, 2019.
[64] S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, and S.-J. Ko, "Parallel feature pyramid network for object detection," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 234–250, 2018.
[65] Z. Li and F. Zhou, "Fssd: feature fusion single shot multibox detector," arXiv preprint arXiv:1712.00960, 2017.
[66] Y. Chen, J. Li, B. Zhou, J. Feng, and S. Yan, "Weaving multi-scale context for single shot detector," arXiv preprint arXiv:1712.03149, 2017.
[67] L. Zheng, C. Fu, and Y. Zhao, "Extend the shallow part of single shot multibox detector via convolutional neural network," in Tenth International Conference on Digital Image Processing (ICDIP 2018), vol. 10806, p. 1080613, International Society for Optics and Photonics, 2018.
[68] S.-H. Bae, "Object detection based on region decomposition and assembly," arXiv preprint arXiv:1901.08225, 2019.
[69] E. Barnea and O. Ben-Shahar, "On the utility of context (or the lack thereof) for object detection," CoRR, vol. abs/1711.05471, 2017.
[70] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: Object detection using scene-level context and instance-level relationships," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6985–6994, 2018.
[71] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," in Advances in Neural Information Processing Systems, pp. 9310–9320, 2018.
[72] K. Liang, H. Chang, B. Ma, S. Shan, and X. Chen, "Unifying visual attribute learning with object recognition in a multiplicative framework," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, pp. 1747–1760, July 2019.
[73] C. Zhang and J. Kim, "Object detection with location-aware deformable convolution and backward attention filtering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9452–9461, 2019.
[74] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, and I. So Kweon, "Attentionnet: Aggregating weak directions for accurate object detection," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2659–2667, 2015.
[75] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," arXiv preprint arXiv:1502.03044, 2015.
[76] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," arXiv preprint arXiv:1412.7755, 2014.
[77] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, "Attention based glaucoma detection: A large-scale database and cnn model," arXiv preprint arXiv:1903.10831, 2019.
[78] K. Hara, M.-Y. Liu, O. Tuzel, and A.-m. Farahmand, "Attentional network for visual object detection," arXiv preprint arXiv:1702.01478, 2017.
[79] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, "An attentive survey of attention models," CoRR, vol. abs/1904.02874, 2019.
[80] T. Kong, F. Sun, C. Tan, H. Liu, and W. Huang, "Deep feature pyramid reconfiguration for object detection," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 169–185, 2018.
[81] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in Proceedings of the 24th ACM international conference on Multimedia, pp. 516–520, ACM, 2016.
[82] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," arXiv preprint arXiv:1902.09630, 2019.
[83] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2888–2897, 2019.
[84] Y. He, X. Zhang, M. Savvides, and K. Kitani, "Softer-nms: Rethinking bounding box regression for accurate object detection," arXiv preprint arXiv:1809.08545, 2018.
[85] C. Cabriel, N. Bourg, P. Jouchet, G. Dupuis, C. Leterrier, A. Baron, M.-A. Badet-Denisot, B. Vauzeilles, E. Fort, and S. Lévêque-Fort, "Combining 3d single molecule localization strategies for reproducible bioimaging," Nature Communications, vol. 10, no. 1, p. 1980, 2019.
[86] M. Bucher, S. Herbin, and F. Jurie, "Hard negative mining for metric learning based zero-shot classification," in European Conference on Computer Vision, pp. 524–531, Springer, 2016.
[87] H. Yu, Z. Zhang, Z. Qin, H. Wu, D. Li, J. Zhao, and X. Lu, "Loss rank mining: A general hard example mining method for real-time detectors," in 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, IEEE, 2018.
[88] K. Chen, J. Li, W. Lin, J. See, J. Wang, L. Duan, Z. Chen, C. He, and J. Zou, "Towards accurate one-stage object detection with ap-loss," arXiv preprint arXiv:1904.06373, 2019.
[89] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 784–799, 2018.
[90] S. Liu, D. Huang, and Y. Wang, "Adaptive nms: Refining pedestrian detection in a crowd," arXiv preprint arXiv:1904.03629, 2019.
[91] J. Hosang, R. Benenson, and B. Schiele, "A convnet for non-maximum suppression," in German Conference on Pattern Recognition, pp. 192–204, Springer, 2016.
[92] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4507–4515, 2017.
[93] J. Jeong, H. Park, and N. Kwak, "Enhancement of ssd by concatenating feature maps for object detection," arXiv preprint arXiv:1705.09587, 2017.
[94] W. Xiang, D.-Q. Zhang, H. Yu, and V. Athitsos, "Context-aware single-shot detector," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1784–1793, IEEE, 2018.
[95] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, and J. Wu, "Feature-fused ssd: fast detection for small objects," in Ninth International Conference

on Graphic and Image Processing (ICGIP 2017), vol. 10615, p. 106151E, International Society for Optics and Photonics, 2018.

[96] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, “Perceptual generative adversarial networks for small object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1222–1230, 2017.

[97] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, “Learning efficient single-stage pedestrian detectors by asymptotic localization fitting,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 618–634, 2018.

[98] P. Hu and D. Ramanan, “Finding tiny faces,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 951–959, 2017.

[99] M. Xu, L. Cui, P. Lv, X. Jiang, J. Niu, B. Zhou, and M. Wang, “Mdssd: Multi-scale deconvolutional single shot detector for small objects,” arXiv preprint arXiv:1805.07009, 2018.

[100] J. Wang, Y. Yuan, and G. Yu, “Face attention network: an effective face detector for the occluded faces,” arXiv preprint arXiv:1711.07246, 2017.

[101] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, “Repulsion loss: Detecting pedestrians in a crowd,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7774–7783, 2018.

[102] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, “Occlusion-aware r-cnn: detecting pedestrians in a crowd,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 637–653, 2018.

[103] P. Baqué, F. Fleuret, and P. Fua, “Deep occlusion reasoning for multi-camera multi-target detection,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 271–279, 2017.

[104] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.

[105] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” arXiv preprint arXiv:1312.6229, 2013.

[106] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9, pp. 1627–1645, 2009.

[107] H. Law, Y. Teng, O. Russakovsky, and J. Deng, “Cornernet-lite: Efficient keypoint based object detection,” arXiv preprint arXiv:1904.08900, 2019.

[108] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “Centernet: Keypoint triplets for object detection,” arXiv preprint arXiv:1904.08189, 2019.

[109] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, “Region proposal by guided anchoring,” arXiv preprint arXiv:1901.03278, 2019.

[110] X. Zhou, J. Zhuo, and P. Krähenbühl, “Bottom-up object detection by grouping extreme and center points,” arXiv preprint arXiv:1901.08043, 2019.

[111] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” CoRR, vol. abs/1904.07850, 2019.

[112] S. Chen, J. Li, C. Yao, W. Hou, S. Qin, W. Jin, and X. Tang, “Dubox: No-prior box objection detection via residual dual scale detectors,” arXiv preprint arXiv:1904.06883, 2019.

[113] C. Zhu, Y. He, and M. Savvides, “Feature selective anchor-free module for single-shot object detection,” arXiv preprint arXiv:1903.00621, 2019.

[114] R. Zhu, S. Zhang, X. Wang, L. Wen, H. Shi, L. Bo, and T. Mei, “Scratchdet: Training single-shot object detectors from scratch,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019.

[115] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, “Dsod: Learning deeply supervised object detectors from scratch,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 1919–1927, 2017.

[116] Z. Shen, H. Shi, R. Feris, L. Cao, S. Yan, D. Liu, X. Wang, X. Xue, and T. S. Huang, “Learning object detectors from scratch with gated recurrent feature pyramids,” arXiv preprint arXiv:1712.00886, 2017.

[117] Y. Li, J. Li, W. Lin, and J. Li, “Tiny-dsod: Lightweight object detection for resource-restricted usages,” arXiv preprint arXiv:1807.11013, 2018.

[118] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, “Object detection from scratch with deep supervision,” arXiv preprint arXiv:1809.09294, 2018.

[119] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, “Light-head r-cnn: In defense of two-stage object detector,” arXiv preprint arXiv:1711.07264, 2017.

[120] A. Womg, M. J. Shafiee, F. Li, and B. Chwyl, “Tiny ssd: A tiny single-shot detection deep convolutional neural network for real-time embedded object detection,” in 2018 15th Conference on Computer and Robot Vision (CRV), pp. 95–101, IEEE, 2018.

[121] L. Tychsen-Smith and L. Petersson, “Denet: Scalable real-time object detection with directed sparse sampling,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 428–436, 2017.

[122] S. Tripathi, G. Dane, B. Kang, V. Bhaskaran, and T. Nguyen, “Lcdet: Low-complexity fully-convolutional neural networks for object detection in embedded systems,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 94–103, 2017.

[123] Y. Lee, H. Kim, E. Park, X. Cui, and H. Kim, “Wide-residual-inception networks for real-time object detection,” in 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 758–764, IEEE, 2017.

[124] Q. Li, S. Jin, and J. Yan, “Mimicking very efficient network for object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6356–6364, 2017.

[125] H.-Y. Zhou, B.-B. Gao, and J. Wu, “Adaptive feeding: Achieving fast and accurate detections by adaptively combining object detectors,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 3505–3513, 2017.

[126] S. Liu, D. Huang, et al., “Receptive field block net for accurate and fast object detection,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 385–400, 2018.

[127] R. Ranjan, V. M. Patel, and R. Chellappa, “Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, pp. 121–135, Jan 2019.

[128] R. He, X. Wu, Z. Sun, and T. Tan, “Wasserstein cnn: Learning invariant features for nir-vis face recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, pp. 1761–1773, July 2019.

[129] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, “Adacos: Adaptively scaling cosine logits for effectively learning deep face representations,” arXiv preprint arXiv:1905.00292, 2019.

[130] Y. Liu, H. Li, and X. Wang, “Rethinking feature discrimination and polymerization for large-scale recognition,” arXiv preprint arXiv:1710.00870, 2017.

[131] R. Ranjan, C. D. Castillo, and R. Chellappa, “L2-constrained softmax loss for discriminative face verification,” arXiv preprint arXiv:1703.09507, 2017.

[132] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, “Normface: l 2 hypersphere embedding for face verification,” in Proceedings of the 25th ACM international conference on Multimedia, pp. 1041–1049, ACM, 2017.

[133] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” arXiv preprint arXiv:1801.07698, 2018.

[134] M. Wang and W. Deng, “Deep face recognition: A survey,” arXiv preprint arXiv:1804.06655, 2018.

[135] Z. Cai, M. J. Saberian, and N. Vasconcelos, “Learning complexity-aware cascades for pedestrian detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019.

[136] M. J. Saberian and N. Vasconcelos, “Learning optimal embedded cascades,” IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 10, pp. 2005–2018, 2012.

[137] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast feature pyramids for object detection,” IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 8, pp. 1532–1545, 2014.

[138] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, “Computer vision and deep learning techniques for pedestrian detection and tracking: A survey,” Neurocomputing, vol. 300, pp. 17–33, 2018.

[139] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, “Change-point detection in time-series data by relative density-ratio estimation,” Neural Networks, vol. 43, pp. 72–83, 2013.

[140] P. Senin, J. Lin, X. Wang, T. Oates, S. Gandhi, A. P. Boedihardjo, C. Chen, and S. Frankenstein, “Grammarviz 3.0: Interactive discovery of variable-length time series patterns,” ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 12, no. 1, p. 10, 2018.

[141] M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos, “A general suspiciousness metric for dense blocks in multimodal data,” in 2015 IEEE International Conference on Data Mining, pp. 781–786, IEEE, 2015.

[142] E. Wu, W. Liu, and S. Chawla, "Spatio-temporal outlier detection in precipitation data," in International Workshop on Knowledge Discovery from Sensor Data, pp. 115–133, Springer, 2008.

[143] B. Barz, E. Rodner, Y. G. Garcia, and J. Denzler, "Detecting regions of maximal divergence for spatio-temporal anomaly detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, pp. 1088–1101, May 2019.

[144] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 12, pp. 7405–7415, 2016.

[145] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," IEEE Transactions on Geoscience and Remote Sensing, 2019.

[146] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X. X. Zhu, "R$^3$-net: A deep network for multioriented vehicle detection in aerial images and videos," IEEE Transactions on Geoscience and Remote Sensing, 2019.

[147] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 8, pp. 3652–3664, 2017.

[148] N. Audebert, B. Le Saux, and S. Lefèvre, "Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images," Remote Sensing, vol. 9, no. 4, p. 368, 2017.

[149] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "Hsf-net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," IEEE Transactions on Geoscience and Remote Sensing, no. 99, pp. 1–15, 2018.

[150] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "R$^2$-cnn: Fast tiny object detection in large-scale remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, 2019.

[151] J. Pei, Y. Huang, W. Huo, Y. Zhang, J. Yang, and T.-S. Yeo, "Sar automatic target recognition based on multiview deep learning framework," IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 4, pp. 2196–2210, 2017.

[152] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 5, pp. 2486–2498, 2017.

[153] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. X. Zhu, "Buildings detection in vhr sar images using fully convolution neural networks," IEEE transactions on geoscience and remote sensing, vol. 57, no. 2, pp. 1100–1116, 2018.

[154] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 9, pp. 5553–5563, 2016.

[155] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 6, pp. 3325–3337, 2014.

[156] Q. Li, Y. Wang, Q. Liu, and W. Wang, "Hough transform guided deep feature extraction for dense building detection in remote sensing images," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1872–1876, IEEE, 2018.

[157] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," IEEE Transactions on Geoscience and Remote Sensing, no. 99, pp. 1–13, 2018.

[158] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," IEEE Geoscience and remote sensing letters, vol. 11, no. 10, pp. 1797–1801, 2014.

[159] N. Ammour, H. Alhichri, Y. Bazi, B. Benjdira, N. Alajlan, and M. Zuair, "Deep learning approach for car detection in uav imagery," Remote Sensing, vol. 9, no. 4, p. 312, 2017.

[160] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 98, pp. 119–132, 2014.

[161] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3974–3983, 2018.

[162] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images," IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 9, pp. 1938–1942, 2015.

[163] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," Journal of Visual Communication and Image Representation, vol. 34, pp. 187–203, 2016.

[164] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 117, pp. 11–28, 2016.

[165] P. Shivakumara, D. Tang, M. Asadzadehkaljahi, T. Lu, U. Pal, and M. H. Anisi, "Cnn-rnn based method for license plate recognition," CAAI Transactions on Intelligence Technology, vol. 3, no. 3, pp. 169–175, 2018.

[166] M. Sarfraz and M. J. Ahmed, "An approach to license plate recognition system using neural network," in Exploring Critical Approaches of Evolutionary Computation, pp. 20–36, IGI Global, 2019.

[167] H. Li, P. Wang, and C. Shen, "Toward end-to-end car license plate detection and recognition with deep neural networks," IEEE Transactions on Intelligent Transportation Systems, no. 99, pp. 1–11, 2018.

[168] J. Qian and B. Qu, "Fast license plate recognition method based on competitive neural network," in 2018 3rd International Conference on Communications, Information Management and Network Security (CIMNS 2018), Atlantis Press, 2018.

[169] R. Laroca, E. Severo, L. A. Zanlorensi, L. S. Oliveira, G. R. Gonçalves, W. R. Schwartz, and D. Menotti, "A robust real-time automatic license plate recognition based on the yolo detector," in 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–10, IEEE, 2018.

[170] A. S. Nair, S. Raju, K. Harikrishnan, and A. Mathew, "A survey of techniques for license plate detection and recognition," i-manager's Journal on Image Processing, vol. 5, no. 1, p. 25, 2018.

[171] W. Lu, Y. Zhou, G. Wan, S. Hou, S. Song, and B. A. D. B. U. ADU, "L3-net: Towards learning based lidar localization for autonomous driving," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6389–6398, 2019.

[172] X. Song, P. Wang, D. Zhou, R. Zhu, C. Guan, Y. Dai, H. Su, H. Li, and R. Yang, "Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving," arXiv preprint arXiv:1811.12222, 2018.

[173] K. Banerjee, D. Notz, J. Windelen, S. Gavarraju, and M. He, "Online camera lidar fusion and object detection on hybrid data for autonomous driving," in 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1632–1638, IEEE, 2018.

[174] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," IEEE Transactions on Intelligent Transportation Systems, 2019.

[175] J. Li and Z. Wang, "Real-time traffic sign recognition based on efficient cnns in the wild," IEEE Transactions on Intelligent Transportation Systems, no. 99, pp. 1–10, 2018.

[176] T. Moritani, Y. Otsubo, and T. Arinaga, "Traffic sign recognition system," Jan. 9 2018. US Patent 9,865,165.

[177] S. Khalid, N. Muhammad, and M. Sharif, "Automatic measurement of the traffic sign with digital segmentation and recognition," IET Intelligent Transport Systems, vol. 13, no. 2, pp. 269–279, 2018.

[178] Á. Arcos-García, J. A. Álvarez-García, and L. M. Soria-Morillo, "Deep neural network for traffic sign recognition systems: An analysis of spatial transformers and stochastic optimisation methods," Neural Networks, vol. 99, pp. 158–165, 2018.

[179] D. Li, D. Zhao, Y. Chen, and Q. Zhang, "Deepsign: Deep learning based traffic sign recognition," in 2018 international joint conference on neural networks (IJCNN), pp. 1–6, IEEE, 2018.

[180] B.-X. Wu, P.-Y. Wang, Y.-T. Yang, and J.-I. Guo, "Traffic sign recognition with light convolutional networks," in 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), pp. 1–2, IEEE, 2018.

[181] S. Zhou, W. Liang, J. Li, and J.-U. Kim, "Improved vgg model for road traffic sign recognition," Computers, Materials and Continua, vol. 57, pp. 11–24, 2018.

[182] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, and Z. Zeng, "Clu-cnns: Object detection for medical images," Neurocomputing, vol. 350, pp. 53–59, 2019.

[183] Q. Liu, L. Fang, G. Yu, D. Wang, C.-L. Xiao, and K. Wang, "Detection of dna base modifications by deep recurrent neural network on oxford

nanopore sequencing data," Nature Communications, vol. 10, no. 1, p. 2449, 2019.

[184] P. J. Schubert, S. Dorkenwald, M. Januszewski, V. Jain, and J. Kornfeld, "Learning cellular morphology with neural networks," Nature Communications, vol. 10, no. 1, p. 2736, 2019.

[185] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172, IEEE, 2018.

[186] S. Naji, H. A. Jalab, and S. A. Kareem, "A survey on skin detection in colored images," Artificial Intelligence Review, pp. 1–47, 2018.

[187] E. Goldman, R. Herzig, A. Eisenschtat, O. Ratzon, I. Levi, J. Goldberger, and T. Hassner, "Precise detection in densely packed scenes," CoRR, vol. abs/1904.00853, 2019.

[188] Z. Yang, Q. Li, L. Wenyin, and J. Lv, "Shared multi-view data representation for multi-domain event detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019.

[189] Y. Wang, H. Sundaram, and L. Xie, "Social event detection with interaction graph modeling," in Proceedings of the 20th ACM international conference on Multimedia, pp. 865–868, ACM, 2012.

[190] M. Schinas, S. Papadopoulos, G. Petkos, Y. Kompatsiaris, and P. A. Mitkas, "Multimodal graph-based event detection and summarization in social media streams," in Proceedings of the 23rd ACM international conference on Multimedia, pp. 189–192, ACM, 2015.

[191] M. Hasan, M. A. Orgun, and R. Schwitter, "A survey on real-time event detection from the twitter data stream," Journal of Information Science, vol. 44, no. 4, pp. 443–463, 2018.

[192] O. Teboul, I. Kokkinos, L. Simon, P. Koutsourakis, and N. Paragios, "Shape grammar parsing via reinforcement learning," in CVPR 2011, pp. 2273–2280, IEEE, 2011.

[193] P. Zhao, T. Fang, J. Xiao, H. Zhang, Q. Zhao, and L. Quan, "Rectilinear parsing of architecture in urban environment," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 342–349, IEEE, 2010.

[194] S. Friedman and I. Stamos, "Online detection of repeated structures in point clouds of urban scenes for compression and registration," International journal of computer vision, vol. 102, no. 1-3, pp. 112–128, 2013.

[195] C.-H. Shen, S.-S. Huang, H. Fu, and S.-M. Hu, "Adaptive partitioning of urban facades," in ACM Transactions on Graphics (TOG), vol. 30, p. 184, ACM, 2011.

[196] G. Schindler, P. Krishnamurthy, R. Lublinerman, Y. Liu, and F. Dellaert, "Detecting and matching repeated patterns for automatic geo-tagging in urban environments," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7, IEEE, 2008.

[197] C. Wu, J.-M. Frahm, and M. Pollefeys, "Detecting large repetitive structures with salient boundaries," in European conference on computer vision, pp. 142–155, Springer, 2010.

[198] P. Müller, G. Zeng, P. Wonka, and L. Van Gool, "Image-based procedural modeling of facades," in ACM Transactions on Graphics (TOG), vol. 26, p. 85, ACM, 2007.

[199] O. Barinova, V. Lempitsky, E. Tretiak, and P. Kohli, "Geometric image parsing in man-made environments," in European conference on computer vision, pp. 57–70, Springer, 2010.

[200] M. Kozinski, R. Gadde, S. Zagoruyko, G. Obozinski, and R. Marlet, "A mrf shape prior for facade parsing with occlusions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2820–2828, 2015.

[201] A. Cohen, A. G. Schwing, and M. Pollefeys, "Efficient structured parsing of facades using dynamic programming," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3206–3213, 2014.

[202] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," Inverse Problems, vol. 27, no. 2, p. 025010, 2011.

[203] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," Journal of the ACM (JACM), vol. 58, no. 3, p. 11, 2011.

[204] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pp. 208–220, 2012.

[205] J. Liu, E. Psarakis, Y. Feng, and I. Stamos, "A kronecker product model for repeated pattern detection on 2d urban images," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2018.

[206] W. Yang, R. T. Tan, J. Feng, J. Liu, S. Yan, and Z. Guo, "Joint rain detection and removal from a single image with contextualized deep networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019.

[207] X. Hu, C. Fu, L. Zhu, J. Qin, and P. Heng, "Direction-aware spatial context features for shadow detection and removal," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019.

[208] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2846–2854, 2016.

[209] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "Contextlocnet: Context-aware deep network models for weakly supervised localization," in European Conference on Computer Vision, pp. 350–365, Springer, 2016.

[210] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2843–2851, 2017.

[211] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 914–922, 2017.

[212] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Image co-localization by mimicking a good detector's confidence score distribution," in European Conference on Computer Vision, pp. 19–34, Springer, 2016.

[213] Z. Yang, D. Mahajan, D. Ghadiyaram, R. Nevatia, and V. Ramanathan, "Activity driven weakly supervised object detection," arXiv preprint arXiv:1904.01665, 2019.

[214] F. Wan, P. Wei, Z. Han, J. Jiao, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019.

[215] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. L. Yuille, "Pcl: Proposal cluster learning for weakly supervised object detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2018.

[216] C. Cao, Y. Huang, Y. Yang, L. Wang, Z. Wang, and T. Tan, "Feedback convolutional neural network for visual localization and segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, pp. 1627–1640, July 2019.

[217] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, "C-mil: Continuation multiple instance learning for weakly supervised object detection," arXiv preprint arXiv:1904.05647, 2019.

[218] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," arXiv preprint arXiv:1904.08739, 2019.

[219] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," arXiv preprint arXiv:1904.09569, 2019.

[220] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019.

[221] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, pp. 1734–1746, July 2019.

[222] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1623–1632, 2019.

[223] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8554–8564, 2019.

[224] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," IEEE Transactions on Image Processing, vol. 24, no. 8, pp. 2552–2564, 2015.

[225] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2192–2199, 2013.

[226] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," IEEE Transactions on Image Processing, vol. 27, no. 1, pp. 349–364, 2017.

[227] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal

propagation," IEEE transactions on circuits and systems for video technology, vol. 27, no. 12, pp. 2527–2542, 2016.
[228] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 6, pp. 1187–1200, 2013.
[229] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 4185–4196, 2015.
[230] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," IEEE Transactions on Image Processing, vol. 26, no. 7, pp. 3156–3170, 2017.
[231] Y. Chen, W. Zou, Y. Tang, X. Li, C. Xu, and N. Komodakis, "Scom: Spatiotemporal constrained optimization for salient object detection," IEEE Transactions on Image Processing, vol. 27, no. 7, pp. 3345–3357, 2018.
[232] S. Li, B. Seybold, A. Vorobyov, X. Lei, and C.-C. Jay Kuo, "Unsupervised video object segmentation with motion-based bilateral networks," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 207–223, 2018.
[233] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," IEEE transactions on circuits and systems for video technology, vol. 24, no. 9, pp. 1522–1540, 2014.
[234] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 715–731, 2018.
[235] Y. Tang, W. Zou, Z. Jin, Y. Chen, Y. Hua, and X. Li, "Weakly supervised salient object detection with spatiotemporal cascade neural networks," IEEE Transactions on Circuits and Systems for Video Technology, 2018.
[236] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," IEEE Transactions on Image Processing, vol. 27, no. 1, pp. 38–49, 2017.
[237] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2334–2342, 2016.
[238] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3395–3402, 2015.
[239] T. Xi, W. Zhao, H. Wang, and W. Lin, "Salient object detection with spatiotemporal background priors for video," IEEE Transactions on Image Processing, vol. 26, no. 7, pp. 3425–3436, 2016.
[240] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in Proceedings of the IEEE international conference on computer vision, pp. 1404–1412, 2015.
[241] F. Zhou, S. Bing Kang, and M. F. Cohen, "Time-mapping using space-time saliency," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3358–3365, 2014.
[242] M. Sun, A. Farhadi, and S. Seitz, "Ranking domain-specific highlights by analyzing edited videos," in European conference on computer vision, pp. 787–802, Springer, 2014.
[243] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 982–990, 2016.
[244] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo, "Unsupervised extraction of video highlights via robust recurrent auto-encoders," in Proceedings of the IEEE international conference on computer vision, pp. 4633–4641, 2015.
[245] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3707–3715, 2015.
[246] R. Panda, A. Das, Z. Wu, J. Ernst, and A. K. Roy-Chowdhury, "Weakly supervised summarization of web videos," in Proceedings of the IEEE International Conference on Computer Vision, pp. 3657–3666, 2017.
[247] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in European conference on computer vision, pp. 540–555, Springer, 2014.
[248] B. Xiong, Y. Kalantidis, D. Ghadiyaram, and K. Grauman, "Less is more: Learning highlight detection from video duration," arXiv preprint arXiv:1903.00859, 2019.
[249] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," arXiv preprint arXiv:1902.10903, 2019.
[250] Y. Liu, M. Cheng, X. Hu, J. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2018.
[251] X. Ren, Y. Zhou, J. He, K. Chen, X. Yang, and J. Sun, "A convolutional neural network-based chinese text detection algorithm via text structure modeling," IEEE Transactions on Multimedia, vol. 19, no. 3, pp. 506–518, 2016.
[252] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[253] D. Bazazian, R. Gomez, A. Nicolaou, L. Gomez, D. Karatzas, and A. D. Bagdanov, "Improving text proposals for scene images with fully convolutional networks," arXiv preprint arXiv:1702.05089, 2017.
[254] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4159–4167, 2016.
[255] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene text detection via holistic, multi-channel prediction," arXiv preprint arXiv:1606.09002, 2016.
[256] T. He, W. Huang, Y. Qiao, and J. Yao, "Accurate text localization in natural image with cascaded convolutional text network," arXiv preprint arXiv:1603.09423, 2016.
[257] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7553–7563, 2018.
[258] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," IEEE Transactions on Multimedia, vol. 20, no. 11, pp. 3111–3122, 2018.
[259] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, "Towards universal object detection by domain attention," arXiv preprint arXiv:1904.04402, 2019.
[260] H. Bilen and A. Vedaldi, "Universal representations: The missing link between faces, text, planktons, and cat breeds," arXiv preprint arXiv:1701.07275, 2017.
[261] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in Advances in Neural Information Processing Systems, pp. 506–516, 2017.
[262] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Efficient parametrization of multi-domain deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8119–8127, 2018.
[263] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3339–3348, 2018.
[264] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," arXiv preprint arXiv:1812.04798, 2018.
[265] A. Haupmann, G. Kang, L. Jiang, and Y. Yang, "Contrastive adaptation network for unsupervised domain adaptation," 2019.
[266] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, "Seq-nms for video object detection," arXiv preprint arXiv:1602.08465, 2016.
[267] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in Proceedings of the IEEE International Conference on Computer Vision, pp. 3038–3046, 2017.
[268] K. Kang, W. Ouyang, H. Li, and X. Wang, "Object detection from video tubelets with convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 817–825, 2016.
[269] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, et al., "T-cnn: Tubelets with convolutional neural networks for object detection from videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896–2907, 2017.
[270] K. Kang, H. Li, T. Xiao, W. Ouyang, J. Yan, X. Liu, and X. Wang, "Object detection in videos with tubelet proposal networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 727–735, 2017.
[271] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2349–2358, 2017.

[272] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in Proceedings of the IEEE International Conference on Computer Vision, pp. 408–417, 2017.

[273] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in Proceedings of the IEEE international conference on computer vision, pp. 3119–3127, 2015.

[274] G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 331–346, 2018.

[275] F. Xiao and Y. Jae Lee, "Video object detection with an aligned spatial-temporal memory," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 485–501, 2018.

[276] P. Tang, C. Wang, X. Wang, W. Liu, W. Zeng, and J. Wang, "Object detection in videos by high quality object linking," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019.

[277] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks," in 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 1355–1361, IEEE, 2017.

[278] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660, 2017.

[279] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in Advances in Neural Information Processing Systems, pp. 5099–5108, 2017.

[280] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4490–4499, 2018.

[281] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2147–2156, 2016.

[282] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1907–1915, 2017.

[283] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," arXiv preprint arXiv:1904.01649, 2019.

[284] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299, 2017.

[285] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in European Conference on Computer Vision, pp. 717–732, Springer, 2016.

[286] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in European Conference on Computer Vision, pp. 483–499, Springer, 2016.

[287] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in Advances in neural information processing systems, pp. 1736–1744, 2014.

[288] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1653–1660, 2014.

[289] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1347–1355, 2015.

[290] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2329–2336, 2014.

[291] G. Rogez, P. Weinzaepfel, and C. Schmid, "Lcr-net++: Multi-person 2d and 3d pose detection in natural images," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019.

[292] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112, 2018.

[293] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4903–4911, 2017.

[294] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 466–481, 2018.

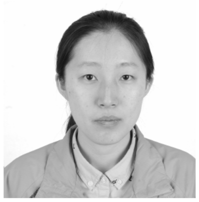
[295] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732, 2016.

[296] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," arXiv preprint arXiv:1901.00148, 2019.

LICHENG JIAO (SM'89–F'18) received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982, the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

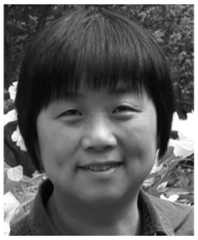
From 1990 to 1991, he was a Postdoctoral Fellow with the National Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, China. Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China, Xidian University, Xi'an, International Research Center of Intelligent Perception and Computation. He has led approximately 40 important scientific research projects. He has authored over 10 monographs, 100 papers in international journals and conferences, and three books: Theory of Neural Network Systems (Xidian University Press, 1990), Theory and Application on Nonlinear Transformation Functions (Xidian University Press, 1992), and Applications and Implementations of Neural Networks (Xidian University Press, 1996). He has authored or co-authored over 150 scientific papers. His current research interests include intelligent information processing, image processing, machine learning, and pattern recognition.

Dr. Jiao is the President of the Computational Intelligence Chapter, the IEEE Xi'an Section, and the IET Xi'an Network, the Chairman of the Awards and Recognition Committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, a Councilor of the Chinese Institute of Electronics, a Committee Member of the Chinese Committee of Neural Networks, a member of the IEEE Xi'an Section Execution Committee, and an Expert of the Academic Degrees Committee of the State Council. He was a recipient of the Second Prize of the National Natural ScienceAward in 2013.



FAN ZHANG received the B.S. degree in electronic information engineering from Shenyang Agricultural University, Shenyang, China, in 2016, where she is currently pursuing the Ph.D. degree with the the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China, Xidian University, Xi'an, China.

Her current research interests include deep learning, object detection, and image understanding.



FANG LIIU (SM'07) received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.S. degree from Xidian University, Xi'an, in 1995, both in computer science and technology.

She is currently a Professor with Xidian University. She has authored or co-authored five books and over 80 papers in journals and conferences. Her current research interests include image perception and pattern recognition, machine learning, and data mining.

Prof. Liu was a recipient of the Second Prize of the National Natural Science Award in 2013.

SHUYUAN YANG (SM'14) received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in circuit and system from Xidian University, Xi'an, China, in 2000, 2003, and 2005, respectively.

She is currently a Professor with Xidian University. Her research interests include intelligent signal processing, machine learning, and image processing.

LINGLING LI (M'18) received the B.S. andPh.D. degrees from Xidian University, Xi'an, China, in 2011 and 2017, respectively.

She is currently a lecturer with the School of Artificial Intelligence, Xidian University. Between 2013 and 2014, she was an exchange Ph.D. student with the Intelligent Systems Group,Department of Computer Science and Artificial Intelligence, University of theBasque Country UPV/EHU, Spain. Her research interests include image processing, deep learning and pattern recognition.

RONG QU (SM'12) received her BSc in Computer Science and Its Applications from XiDian University, Xi'an, China, in 1996, and PhD in Computer Science from the University of Nottingham, Nottingham, UK, in 2002.

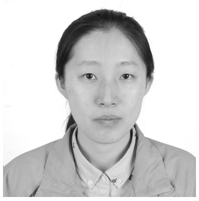
She is currently an Associated Professor at the School of Computer Science, University of Nottingham, Nottingham, UK. has published more than 60 peer-refereed papers at international journals since 2000. Among these several have been awarded the Top Cited Paper at leading Operational Research journals (i.e. 5 year top cited paper at EJOR, top 0.1% or top 1% cited papers by ISI Essential Science Indicators). She has co-authored the book "Hyper-heuristics: theory and applications". Her main research interests include the modelling and optimisation algorithms for scheduling and optimisation algorithms in transport scheduling in logistics, personnel scheduling, telecommunication network routing, portfolio optimisation, and timetabling problems, etc. by using evolutionary algorithms, mathematical programming, constraint programming in operational research and artificial intelligence, and hybridisations of these techniques.

Dr. Qu has been the program chair of several symposium and special sessions on automatic algorithm design at IEEE flagship events. She is the vice-chair of Task Committee of Intelligence Systems and Applications, and Task Force of Hyper-heuristics at IEEE Computational Intelligence Society. She is an associated editor at IEEE Computational Intelligence Magazine since 2016. Dr. Qu was elected by the "China 1000 Elites Plan" in 2013, and appointed as a honored professor at Xidian University 2013-2018.

LICHENG JIAO (SM'89–F'18) received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982, the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

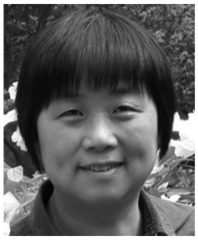
From 1990 to 1991, he was a Postdoctoral Fellow with the National Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, China. Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China, Xidian University, Xi'an, International Research Center of Intelligent Perception and Computation. He has led approximately 40 important scientific research projects. He has authored over 10 monographs, 100 papers in international journals and conferences, and three books: Theory of Neural Network Systems (Xidian University Press, 1990), Theory and Application on Nonlinear Transformation Functions (Xidian University Press, 1992), and Applications and Implementations of Neural Networks (Xidian University Press, 1996). He has authored or co-authored over 150 scientific papers. His current research interests include intelligent information processing, image processing, machine learning, and pattern recognition.

Dr. Jiao is the President of the Computational Intelligence Chapter, the IEEE Xi'an Section, and the IET Xi'an Network, the Chairman of the Awards and Recognition Committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, a Councilor of the Chinese Institute of Electronics, a Committee Member of the Chinese Committee of Neural Networks, a member of the IEEE Xi'an Section Execution Committee, and an Expert of the Academic Degrees Committee of the State Council. He was a recipient of the Second Prize of the National Natural ScienceAward in 2013.



FAN ZHANG received the B.S. degree in electronic information engineering from Shenyang Agricultural University, Shenyang, China, in 2016, where she is currently pursuing the Ph.D. degree with the the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China, Xidian University, Xi'an, China.

Her current research interests include deep learning, object detection, and image understanding.



FANG LIIU (SM'07) received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.S. degree from Xidian University, Xi'an, in 1995, both in computer science and technology.

She is currently a Professor with Xidian University. She has authored or co-authored five books and over 80 papers in journals and conferences. Her current research interests include image perception and pattern recognition, machine learning, and data mining.

Prof. Liu was a recipient of the Second Prize of the National Natural Science Award in 2013.

SHUYUAN YANG (SM'14) received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in circuit and system from Xidian University, Xi'an, China, in 2000, 2003, and 2005, respectively.

She is currently a Professor with Xidian University. Her research interests include intelligent signal processing, machine learning, and image processing.

LINGLING LI (M'18) received the B.S. andPh.D. degrees from Xidian University, Xi'an, China, in 2011 and 2017, respectively.

She is currently a lecturer with the School of Artificial Intelligence, Xidian University. Between 2013 and 2014, she was an exchange Ph.D. student with the Intelligent Systems Group,Department of Computer Science and Artificial Intelligence, University of theBasque Country UPV/EHU, Spain. Her research interests include image processing, deep learning and pattern recognition.

RONG QU (SM'12) received her BSc in Computer Science and Its Applications from XiDian University, Xi'an, China, in 1996, and PhD in Computer Science from the University of Nottingham, Nottingham, UK, in 2002.

She is currently an Associated Professor at the School of Computer Science, University of Nottingham, Nottingham, UK. has published more than 60 peer-refereed papers at international journals since 2000. Among these several have been awarded the Top Cited Paper at leading Operational Research journals (i.e. 5 year top cited paper at EJOR, top 0.1% or top 1% cited papers by ISI Essential Science Indicators). She has co-authored the book "Hyper-heuristics: theory and applications". Her main research interests include the modelling and optimisation algorithms for scheduling and optimisation algorithms in transport scheduling in logistics, personnel scheduling, telecommunication network routing, portfolio optimisation, and timetabling problems, etc. by using evolutionary algorithms, mathematical programming, constraint programming in operational research and artificial intelligence, and hybridisations of these techniques.

Dr. Qu has been the program chair of several symposium and special sessions on automatic algorithm design at IEEE flagship events. She is the vice-chair of Task Committee of Intelligence Systems and Applications, and Task Force of Hyper-heuristics at IEEE Computational Intelligence Society. She is an associated editor at IEEE Computational Intelligence Magazine since 2016. Dr. Qu was elected by the "China 1000 Elites Plan" in 2013, and appointed as a honored professor at Xidian University 2013-2018.