International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

A Review on Recent Development in Small Object Detection based on AI

Rishabh Dixit[1], Sushil Kumar Maurya[2]
Computer science and Engineering Department,
Bansal Institute of Engineering and Technology, Lucknow
rishabh.saurabh333@gmail.com

Abstract: The detection of small objects is a difficult computer vision challenge. It has been used extensively in the military, transportation, business, etc. We thoroughly examine the existing small object detection methods based on deep learning from five perspectives, including multi-scale feature learning, data augmentation, training strategy, context-based detection, and GAN-based detection, in order to facilitate in-depth understanding of small object detection. The performance of certain common small object detection techniques is then carefully examined using well-known datasets as MS-COCO and PASCAL-VOC. Last but not least, five perspectives are used to suggest potential future directions for small object detection research: emerging small object detection datasets and benchmarks, multi-task joint learning and optimisation, information transmission, weakly supervised small object detection methods, and framework for small object detection task.

Keywords: Object detection, Deep Convolutional neural networks (CNNs), Recent progress, Computer vision

1. Introduction:

Small object detection is a fundamental computer technology that deals with identifying instances of small objects of a particular class in digital images and videos. It is related to image understanding and computer vision. Many other computer vision tasks, such as object tracking [1, instance segmentation [2,3], image captioning [4, action recognition [5, scene understanding [6, etc.], are based on small object detection, which is an essential and challenging problem. Small object detection has advanced to a research highlight as a result of the compelling success of deep learning techniques in recent years, which has brought in fresh talent. Small object detection has been extensively utilized in academic settings as well as applications in the real world, including robot vision, autonomous driving, intelligent transportation, drone scene analysis, and military reconnaissance and surveillance.

The most common definitions of small objects are two. In the real world, smaller objects are referred to as one. In the MS-COCO [7] metric evaluation, a different definition of small objects is mentioned. The term "small objects" refers to objects with areas smaller than or equal to 32 x 32 pixels. The community generally accepts this size threshold for datasets pertaining to common objects. Figure depicts a few examples of small objects, including "baseball," "tennis," and the traffic sign "pg." 1. Even though many object detectors are good at detecting medium and large objects, they are terrible at detecting small ones. This is because small object detection faces three challenges. First, small objects lack the necessary appearance information to differentiate them from backgrounds or other similar categories. The locations of small objects then have significantly more options. That is to say, accurate localization necessitates greater precision. Due to the fact that the majority of previous endeavors were optimized for the large object detection problem, small object detection expertise and experience are severely limited.

A comprehensive and in-depth look at small object detection in the age of deep learning is presented in this paper. Our study means to cover completely five regards of little article discovery calculations, including multi-scale highlight learning, information expansion, preparing procedure, setting based recognition and GAN-based location. We investigate datasets and evaluation metrics for small object detection in addition to taxonomically examining the existing methods for detecting small objects. In the meantime, we present a number of promising directions for future research and conduct a comprehensive analysis of the effectiveness of small object detection methods.

The history of small object detection is relatively brief in comparison to that of other tasks in computer vision. Utilizing hand-engineered features and shallow classifiers in aerial images, vehicle detection has been the primary focus of previous research on small object detection [8, 9]. Color- and shape-based features were also used to solve traffic sign detection issues prior to the advent of deep learning [10]. Some methods for small object detection that are based on deep learning have emerged as a result of the rapid development of convolutional neural networks (CNNs) in deep learning. However, there are relatively few surveys and studies that concentrate solely on the detection of small objects. The majority of the most cutting-edge techniques are based on existing object detection algorithms that have been modified to better detect small objects. As far as we are aware, Chen et al. 11] are maybe quick to present a little item location (Grass) dataset, an assessment metric, and give a benchmark score to investigate little article recognition. Krishna and Jawahar [12] later expand on their concepts and propose an efficient upsampling-based method with superior results for small object detection. Zhang et al.'s approach is distinct from the R-CNN (regions with CNN features) used in [11,12]. 13] For remote sensing image small object detection, utilize deconvolution R-CNN [14]. Two major approaches to object detection are single shot detector (SSD) [16] and faster R-CNN [15]. Some small object detection techniques—[17, 18], [19], [20], and [21]]—are suggested based on Faster R-CNN or SSD. Small object detection also makes use of generative adversarial networks (GAN) [29,30], multi-scale techniques [22,23], data augmentation techniques [24], training strategies [25,26], contextual information [27,28], and multi-scale techniques [22,23]. Table 1 provides a brief chronology.

We select significant or influential papers from prestigious conferences and journals with great care. The major advances in small object detection over the past three to five years are the primary focus of this review. However, some additional related works are also included for completeness and easier reading. It is significant that we limit this audit to picture level little article discovery techniques. We won't talk about any other work on small object detection, like video small object detection and 3D small object detection.

2. Related Work:

Deep learning has got a lot of attention since AlexNet [45] won first place in the challenge of ImageNet [14] in 2012. Great improvements have been achieved both in the accuracy [38] and speed [37] of image classification. In this part, we briefly introduce some of the advanced classification architectures that have been widely applied in object detection as backbone, which are utilized to extract features. The development of backbone can be divided into several stages which are represented by some classic network design principles (see Fig. 3). The first is repeat which stacks structure with the same topology and makes the entire network becomes a modular structure. This technique starts from AlexNet and VGG [79] (Fig. 3a) and is adopted by almost all the later works. The second is multi-path which first appears in Inception [82] module (Fig. 3b). The input from the previous layer is divided into different paths to transform by filters with different kernel sizes, and finally, the output is concatenated by a 1×1 convolutional layer. The last is the skip-connection which starts from Highway Network [81] and becomes a standard principle from ResNet [30]. It constructs the connection between high-level and low-level feature information which changes the original single linear structure.

AlexNet: AlexNet [45] consists of five convolutional layers and three fully connected layers. It is a milestone study of deep learning and computer vision for introducing some advanced techniques like training the network with graphics processing unit (GPU) for speeding up the operation of convolution parallelly and using the dropout to prevent from overfitting.

VGGNet: VGGNet [79] won second place in the classification task and the first place in location task in the competition of ILSVRC 2014. The small receptive field is utilized in the whole network for fewer parameters. It has two versions: VGG-16 and VGG-19. VGG-16 has been widely used because of its simple architecture, which has 13 convolutional layers, five pooling layers, and three fully connected layers.

GoogLeNet: To solve the overfitting and computing problem arising with the increasing size of the network, Inception module was introduced in GoogLeNet [82]. Using different kernel sizes of filters in the same layer helps preserve the spatial information and reduce the parameters. It has 22 layers, which is almost three times deeper than AlexNet, but it has 12 times fewer parameters than AlexNet.

ResNet/ResNeXt: ResNet [30] is one of the most successful CNNs and has been exploited in many applications including the very famous AlphaGo [78]. The idea of the ResNet is simple yet effective, which each layer should not learn unreferenced functions but learn residual functions with references to the layer's inputs. This kind of learning makes it easier to train much deeper networks efficiently. ResNet has different architectures: ResNet- 50, ResNet-101 and ResNet-152. ResNeXt [93] is the upgraded version of ResNet. It is constructed by repeating a building block that aggregates a set of transformations with the same topology. It demonstrates that it's more effective to increase the size of the set of transformations (cardinality) than to increase the depth and width. Moreover, a 101-layer ResNeXt can achieve better accuracy than ResNet-200 but with only 50% complexity.

DenseNet: Inspired by the shortcut connection of ResNet, DenseNet [38] connects each layer in the network with every other layer in a feed-forward fashion with L(L+2)/2 direct connections. In addition to the original features (alleviating the vanishing-gradient problem and reducing the number of parameters) of the shortcut connection, this design has new features that strengthen feature propagation and encourage feature reuse. Besides, with the help of the bottleneck layer, translation layer, and small growth rate, the network becomes narrow which can prevent from overfitting. The models above mainly focus on the accuracy improvement of the classification by increasing the depth and width of the network. On the other hand, some architectures are putting their attention on the model size while maintaining considerable accuracy so that they can be utilized on the devices with memory and computation speed constraints.

MobileNets: MobileNet [35] is a lightweight deep neural network proposed by Google for embedded devices such as mobile phones. The core of network designs, separable convolution, can effectively reduce the number of parameters and computation at the expense of lesser performance. Separable convolution replaces traditional convolution operations with two-step convolution operations: depth-wise convolution and point-wise convolution. Subsequent MobileNet-v2 [72] mainly adds residual structure, and adds a layer of pointwise convolution before depth-wise convolution, which optimizes the bandwidth usage and further improves the performance on embedded devices.

Xception: Xception [8] is an improvement to Inception v3 [83], mainly using Depthwise Separable Convolution to replace the original Inception v3 convolution operation, in the premise of little increase in network complexity to improve the effectiveness of the model. Xception separates the tasks related to learning space from the tasks related to learning channels by adding groups to the convolution layer, which dramatically reduces the theoretical computation complexity and the size of the model. SqueezeNet: Based on three architecture design strategies: (1) replace 3 × 3 filters with 1 × 1 filters; (2) decrease the number of input channels to 3 × 3 filters; (3) downsample late in the network so that convolution layers have large activation maps, SqueezeNet [40] is a small CNN architecture. Fire module which consists of squeeze convolution layer and expand layer is used to reduce the parameter number. With further compression, the model size of SqueezeNet can be compressed to less than 0.5 MB which is 510× smaller than AlexNet [45] while it can achieve AlexNet-level accuracy with 50× fewer parameters.

ShuffleNet: ShuffleNet [100] utilizes two new mechanisms, point-wise group convolution, and channel shuffle, to reduce computation cost while maintaining accuracy. Experiments show that it is an extremely computation-efficient CNN architecture with comparable accuracy. Channel split is introduced in the upgrade version, ShuffleNet v2 [61], to speed up the network. Some practical guidelines for efficient network design are proposed in this work, and ShuffleNet v2 achieves a trade-off between speed and accuracy. In addition to these models mentioned above, there are also some noticeable architectures [37, 97]. ZFNet [97] presents a method of deconvolution for visualization of convolution network, which can analyze the effect of convolution network and guide the improvement of the network. Based on AlexNet network, ZFNet obtains a better result. SE block in SENet [37] is designed by explicitly modeling the interdependence between channels and adaptively recalibrating the channel response. The core of the SENet is squeezing and excitation operation.

OverFeat: Overfeat [73], one of the first advances in using deep learning for object detection, integrates three tasks of image classification, location, and detection into a framework to boost the accuracy and won the first place in the ILSVRC2013 localization competition. OverFeat is based on the multi-scale sliding-window algorithm, which is an intuitive search method of object detection.

R-CNNs: R-CNN [24], one of the most famous region-base convolutional neural networks, is the first to use deep CNN to extract feature for object detection. Firstly, it generates about 2k object candidates named region proposals through Selective Search [85]. Then these proposals are resized to the fixed size to fit the input size of the CNN like AlexNet. A fixed length of feature vectors is generated by the CNN and finally classified using class-specific linear support vector machines (SVMs). This simple yet effective pipeline has reached state-of-the-art performance on the benchmark datasets with momentous performance boost over all previous models, which are mainly based on DPM [23] while the computation for every region proposal is very time-consuming. The whole detection pipeline of R-CNN. To solve the computation and the limited image input size problem, SPPnet [29] introduces spatial pyramid pooling to relax the constraint of the fixed input size due to the fully connected layers. More importantly, SPPnet extracts the feature maps from the entire image independent of the region proposal stage. Then it matches the proposals through spatial pyramid pooling (SPP) and generates a fixed-length vector regardless of the input size. Finally, the fixed-length representation is input into the last two fully connected layers and then classified by category-specific linear SVMs. SPPnet speeds up the R-CNN method 24-102× faster with better or comparable accuracy. Fast R-CNN [25] inherits the spatial pyramid pooling from SPPnet but modifies it as Region of Interest (ROI) Pooling which can be seen as a single-level SPP. It uses the bounding-box regressor instead of linear SVMs and utilizes a multi-task loss which makes the network can be trained in a single stage and no extra storage is required for feature caching during the training. This method can train a very deep detection network with a backbone VGG16 [79], testing 9× faster than R-CNN [24] and 3× faster than SPPnet [29]. At test time, the detection network processes one image in 0.3s (excluding object proposal time). Faster R-CNN [71] replaces the Selective Search [85] in the region proposal stage with the region proposal network (RPN) (the right corner of Fig. 4) which is built by several convolutional layers, which makes the network completely trainable end-to-end. With the RPN, Faster R-CNN can process an image in 0.2 seconds (including region proposal), which is 250× faster than R-CNN and 10× than Fast R-CNN, almost toward real-time. It is noticeable that the backbone of R-CNN is AlexNet [45], and SPPnet is based on ZF-5 [97] while Fast R-CNN and Faster R-CNN adopt VGG-16 [79].

YOLOs: Focusing on real-time object detection, YOLO [68] borrows ideas from the design of the architecture of GoogLeNet [82]. The input image is divided into S × S grid and the grid where the center of the object lies in charge of the prediction of the object. Each grid cell outputs B bounding boxes and confidence scores for those boxes, as well as C class probabilities. The unified framework runs at 45 frames per second with the performance outperforming DPM [23] and R-CNN [24]. The architecture of YOLO can be seen in the above part of Fig. 5. To improve the precision and recall of object localization, YOLOv2 [69] adopts some advanced methods to make the detection better, stronger and faster. Briefly, the idea of anchor box is introduced from Faster R-CNN [71] and the network architecture is altered to fit the modification where the fully connected layer of the output layer is replaced by a convolutional layer. UsingWordTree and joint training method, the authors train YOLOv2 simultaneously on the MS COCO [52] detection dataset and the ImageNet classification dataset. YOLOv2 gets 78.6 mAP at the speed of 40 frames per second, outperforming state-of-the-art methods like Faster R-CNN with ResNet and SSD while still running significantly faster. Based on Darknet-53 [67], which is as accurate as ResNet- 101 or ResNet-152 [30] but much faster, YOLOv3 [70] makes an incremental improvement not only on the accuracy perspective but also speed. Multi-scale prediction employed to get more meaningful semantic information from the upsampled features and finer-grained information from the earlier feature map. At the image size of 320 × 320, YOLOv3 runs as accurate as SSD [56] but three times faster. It achieves similar performance but 3.8× faster compared to RetinaNet [54].

3. Conclusion:

The detection of small objects is one of the most difficult computer vision problems. This work compares and analyses the existing classic small object identification algorithms on certain well-known object detection datasets, such as PASCAL-VOC, MS-COCO, KITTI, and TT100K, and thoroughly evaluates tiny object recognition approaches based on deep learning from five dimensions.

References:

[1]. Bansal A, Sikka K, Sharma G, Chellappa R, Divakaran A (2018) Zero-shot object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 384–400

[2]. Bell S, Lawrence Zitnick C, Bala K, Girshick R (2016) 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

[3]. Bodla N, Singh B, Chellappa R, Davis LS (2017) Soft-nms—improving object detection with one line of code. In: 2017 IEEE international conference on Computer vision (ICCV). IEEE, pp 5562–5570

[4]. Byeon YH, Pan SB, Moh SM, Kwak KC (2016) A surveillance system using cnn for face recognition with object, human and face detection. In: Information science and applications (ICISA) 2016. Springer, pp 975–984

[5]. Cai Z, Fan Q, Feris RS, Vasconcelos N (2016) A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In: European conference on computer vision. Springer, pp 354–370

[6]. Cai Z, Vasconcelos N (2017) Cascade r-cnn: Delving into high quality object detection. arXiv:1712.00726

[7]. Cao G, Xie X, Yang W, Liao Q, Shi G, Wu J (2018) Feature-fused ssd: fast detection for small objects. In: Ninth international conference on graphic and image processing (ICGIP 2017). International Society for Optics and Photonics, vol 10615, pp 106151e

[8]. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp 1800–1807. https://doi.org/10.1109/CVPR.2017.195

[9]. Cui L (2018) Mdssd: Multi-scale deconvolutional single shot detector for small objects. arXiv:1805.07009

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

[11]. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable Convolutional Networks, pp 764–773

[12]. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005. CVPR 2005. IEEE computer society conference on Computer vision and pattern recognition, vol 1. IEEE, pp 886– 893

[13]. Demirel B, Cinbis RG, Ikizler-Cinbis N (2018) Zero-shot object detection by hybrid region embedding. arXiv:1805.06157

[14]. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009. CVPR 2009. IEEE conference on Computer vision and pattern recognition. IEEE, pp 248–255

[15]. Doll´ar P, Wojek C, Schiele B, Perona P (2009) Pedestrian detection: a benchmark. In: 2009. CVPR 2009. IEEE conference on Computer vision and pattern recognition. IEEE, pp 304–311

[16]. Dollar P, Wojek C, Schiele B, Perona P (2012) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34(4):743–761

[17]. Erhan D, Szegedy C, Toshev A, Anguelov D (2014) Scalable object detection using deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2147–2154

[18]. Everingham M, Van Gool L,Williams CK,Winn J, Zisserman A (2008) The pascal visual object classes challenge 2007 (voc 2007) results (2007). http://host.robots.ox.ac.uk/pascal/VOC/voc2007/workshop/ index.html

[19]. Everingham M, Van gool L,Williams CKI,Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338

[20]. Fu C, Liu W, Ranga A, Tyagi A, Berg AC (2017) Dssd: Deconvolutional single shot detector. arXiv:1701.06659

[21]. Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on computer vision and pattern recognition (CVPR)

[22]. Gidaris S, Komodakis N (2015) Object detection via a multi-region and semantic segmentation-aware cnn model. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1134–1142

[23]. Girshick R, Felzenszwalb PF, Mcallester D (2012) Discriminatively trained deformable part models release 5
[24]. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
[25]. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
[26]. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
[27]. Goyal P, Doll´ar P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, Tulloch A, Jia Y, He K (2017) Accurate, large minibatch sgd: training imagenet in 1 hour. arXiv:1706.02677
[28]. Han J, Zhang D, Cheng G, Liu N, Xu D (2018) Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Proc Mag 35(1):84–100
[29]. He K, Zhang X, Ren S, Sun J (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European conference on computer vision. Springer, pp 346–361
[30]. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 770–778. https://doi.org/ 10.1109/CVPR.2016.90
[31]. He K, Gkioxari G, Doll´ar P, Girshick R (2017) Mask r-cnn. In: 2017 IEEE international conference on Computer vision (ICCV). IEEE, pp 2980–2988
[32]. He K, Girshick R, Doll´ar P (2018) Rethinking imagenet pre-training. arXiv:1811.08883
[33]. Hong S, Roh B, Kim KH, Cheon Y, Park M (2016) Pvanet: lightweight deep neural networks for real-time object detection. arXiv:1611.08588
[34]. Hosang J, Benenson R, Schiele B (2017) Learning non-maximum suppression. In: The IEEE conference on computer vision and pattern recognition (CVPR), vol 2
[35]. Howard A, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861
[36]. Hu H, Gu J, Zhang Z, Dai J, Wei Y (2017) Relation networks for object detection. arXiv:1711.11575.8 37. Hu J, Shen L, Sun G (2017) Squeeze-and-Excitation Networks. arXiv:1709.01507, pp 1–11. https://doi.org/10.1109/CVPR.2018.00745
[38]. Huang G, Liu Z, Van Der Maaten L,Weinberger KQ (2017) Densely connected convolutional networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp 2261–2269. https://doi.org/10.1109/CVPR.2017.243
[39]. Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S et al (2017) Speed/accuracy trade-offs for modern convolutional object detectors. In: IEEE CVPR, vol 4
[40]. Iandola F, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2016) SqueezeNet: AlexNetlevel accuracy with 50x fewer parameters and <,0.5MB model size. arXiv:1602.07360, pp 1–13. https://doi.org/10.1007/978-3-[319-24553-9
[41]. Jeong J, Park H, Kwak N (2017) Enhancement of SSD by concatenating feature maps for object detection. arXiv:1705.09587
[42]. Jian M, Qi Q, Dong J, Sun X, Sun Y, Lam KM (2018) Saliency detection using quaternionic distance based weber local descriptor and level priors. Multimedia tools and applications, pp 1–18
[43]. Kong T, Yao A, Chen Y, Sun F (2016) Hypernet: Towards accurate region proposal generation and joint object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 845–853
[44]. Kong T, Sun F, Yao A, Liu H, Lu M, Chen Y (2017) Ron: Reverse connection with objectness prior networks for object detection. In: IEEE Conference on computer vision and pattern recognition, vol 1, pp 2
[45]. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. Advances In Neural Information Processing Systems, pp 1–9. https://doi.org/10.1016/ j.protcy.2014.09.007
[46]. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
[47]. Lee H, Eum S, Kwon H (2017) Me r-cnn: Multi-expert r-cnn for object detection. arXiv:1704.01069
[48]. Li Z, Peng C, Yu G, Zhang X, Deng Y, Sun J (2017) Light-head r-cnn: In defense of two-stage object detector. arXiv:1711.07264
[49]. Li Z, Zhou F (2017) Fssd: Feature fusion single shot multibox detector. arXiv:1712.00960
[50]. Li J, Liang X, Li J, Wei Y, Xu T, Feng J, Yan S (2018) Multistage Object Detection With Group Recursive Learning. IEEE Trans Multimed 20(7):1645–1655
[51]. Li Z, Peng C, Yu G, Zhang X, Deng Y, Sun J (2018) DetNet: A Backbone network for Object Detection. arXiv:1804.06215, 1(2), 3
[52]. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Doll˙ar P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, pp 740–755
[53]. Lin TY, Doll´ar P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: CVPR, vol 1, pp 4
[54]. Lin TY, Goyal P, Girshick R, He K, Doll´ar P (2017) Focal loss for dense object detection. arXiv:1708.02002
[55]. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV)

[56]. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, Berg AC (2016) Ssd: Single shot multibox detector. In: European conference on computer vision. Springer, pp 21–37
[57]. Liu Y, Wang R, Shan S, Chen X (2018) 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
[58]. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440 59. Lowe DG (2004) Distinctive image features from scale invariant keypoints. Int J Comput Vis 60:91– 11020,042. https://doi.org/10.1023/B:VISI.0000029664.99615.94
[60]. Lu J, Sibai H, Fabry E (2017) Adversarial examples that fool detectors. arXiv:1712.02494
[61]. Ma N, Zhang X, Zheng HT, Sun J (2018) ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. arXiv:1807.11164, pp 1–19
[62]. Mehta R, Ozturk C (2018) Object detection at 200 frames per second. arXiv:1805.06361
[63]. Oksuz K, Cam BC, Akbas E, Kalkan S (2018) Localization recall precision (lrp): A new performance metric for object detection. arXiv:1807.01696
[64]. Ouyang W, Wang X, Zeng X, Qiu S, Luo P, Tian Y, Li H, Yang S, Wang Z, Loy CC et al (2015) Deepid-net: Deformable deep convolutional neural networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2403–2412
[65]. Peng C, Xiao T, Li Z, Jiang Y, Zhang X, Jia K, Yu G, Sun J (2018) Megdet: a large mini-batch object detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6181–6189
[66]. QiongYan J, LiXu Y (2017) Accurate single stage detector using recurrent rolling convolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
[67]. Redmon J (2013) Darknet: Open source neural networks in c. http://pjreddie.com/darknet 2016
[68]. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
[69]. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger arXiv preprint
[70]. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv:1804.02767
[71]. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
[72]. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L (2018) Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. arXiv:1801.04381
[73]. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229
[74]. Shen Z, Liu Z, Li J, Jiang YG, Chen Y, Xue X (2017) Dsod: Learning deeply supervised object detectors from scratch. In: The IEEE international conference on computer vision (ICCV), vol 3, pp 7
[75]. Shrivastava A, Gupta A (2016) Contextual priming and feedback for faster r-cnn. In: European conference on computer vision. Springer, pp 330–348
[76]. Shrivastava A, Gupta A, Girshick R (2016) 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
[77]. Shrivastava A, Sukthankar R, Malik J, Gupta A (2016) Beyond skip connections: Top-down modulation for object detection. arXiv:1612.06851
[78]. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A et al (2017) Mastering the game of go without human knowledge. Nature 550(7676):354
[79]. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
[80]. Singh B, Li H, Sharma A, Davis LS (2018) R-fcn-3000 at 30fps: Decoupling detection and classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1081–1090
[81]. Srivastava RK, Greff K, Schmidhuber J (2015) Highway networks. arXiv:1505.00387
[82]. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
[83]. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
[84]. Tang Y, Wang J, Wang X, Gao B, Dellandr´ea E, Gaizauskas R, Chen L (2017) Visual and semantic knowledge transfer for large scale semi-supervised object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence
[85]. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW(2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171. https://doi.org/10.1007/s11263-013-0620-5
[86]. Wang J, Fu W, Liu J, Lu H et al (2014) Spatiotemporal group context for pedestrian counting. IEEE Trans Circ Syst Video Technol 24(9):1620–1630
[87]. Wang X, Shrivastava A, Gupta A (2017) A-fast-rcnn: Hard positive generation via adversary for object detection. In: IEEE Conference on computer vision and pattern recognition
[88]. Wang RJ, Li X, Ao S, Ling CX (2018) Pelee: A real-time object detection system on mobile devices. arXiv:1804.06882

[89]. Williams RJ, Zipser D (1989) A learning algorithm for continually running fully recurrent neural networks. Neural Comput 1(2):270–280
[90]. Wong A, Shafiee MJ, Li F, Chwyl B (2018) Tiny ssd: A tiny single-shot detection deep convolutional neural network for real-time embedded object detection. arXiv:1802.06488
[91]. Wu B, Iandola F, Jin PH, Keutzer K (2016) Squeezedet: unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In: IEEE Conference on computer vision and pattern recognition workshops, pp 446–454
[92]. Xie C, Wang J, Zhang Z, Zhou Y, Xie L, Yuille A (2017) Adversarial examples for semantic segmentation and object detection. In: Proceedings of International Conference on Computer Vision (ICCV), pp 1378–1387
[93]. Xie S, Girshick R, Doll˙ar P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp 5987–5995. https://doi.org/10.1109/CVPR.2017.634
[94]. Yang B, Yan J, Lei Z, Li SZ (2016) Craft objects from images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6043–6051
[95]. Yang F, Choi W, Lin Y (2016) Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2129–2137
[96]. You Y, Zhang Z, Hsieh C, Demmel J, Keutzer K Imagenet training in minutes
[97]. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer, pp 818–833
[98]. Zeng X, Ouyang W, Yang B, Yan J, Wang X (2016) Gated bi-directional cnn for object detection. In: European conference on computer vision. Springer, pp 354–369
[99]. Zhang S, Wen L, Bian X, Lei Z, Li SZ (2017) Single-shot refinement neural network for object detection. arXiv preprint
[100]. Zhang X, Zhou X, Lin M, Sun J (2017) ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. arXiv:1707.01083. https://doi.org/10.1109/CVPR.2018.00716
[101]. Zhang Y, Bai Y, Ding M, Li Y, Ghanem B (2018) W2f: a weakly-supervised to fully-supervised framework for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 928–936
[102]. Zhang Z, Qiao S, Xie C, Shen W,Wang B, Yuille AL (2018) Single-shot object detection with enriched semantics. Technical report, Center for Brains, Minds and Machines (CBMM)
[103]. Zhang Z, He T, Zhang H, Zhang Z, Xie J, Li M (2019) Bag of freebies for training object detection neural networks. arXiv:1902.04103