MOBILE CENTERNET FOR EMBEDDED DEEP LEARNING OBJECT DETECTION

Jun Yu[1], *Haonian Xie*[1,*] , *Mengyan Li*[1], *Guochen Xie*[1], *Ye Yu*[2], *Chang Wen Chen*[3]

[1]Department of Automation, University of Science and Technology of China
[2]School of Computer and Information, Hefei University of Technology
[3]Department of Computer Science and Engineering, The State University of New York at Buffalo
harryjun@ustc.edu.cn; {xie233, limmy, xiegc}@mail.ustc.edu.cn; yuye@hfut.edu.cn;
chencw@buffalo.edu

ABSTRACT

Object detection is a fundamental task in computer vision with wide application prospect. And recent years, many novel methods are proposed to tackle this task. However, most algorithms suffer from high computation cost and long inference time, which makes them impossible to be deployed on embedded devices in real industrial application scenarios. In this paper, we propose the Mobile CenterNet to solve this problem. Our method is based on CenterNet but with some key improvements. To enhance detection performance, we adopt HRNet as a powerful backbone and introduce a category-balanced focal loss to deal with category imbalance problem. Moreover, to compress the model size as well as reduce inference time, knowledge distillation is employed to transfer knowledge from cumbersome model to a compact one. We conduct experiments on a large traffic detection dataset BDD100K and validate the effectiveness of all the modifications. Finally, our method achieves the 1st place in the Embedded Deep Learning Object Detection Model Compression Competition held in ICME 2020.

Index Terms— Anchor-free detector, Knowledge distillation, Lightweight detector

1. INTRODUCTION

Object detection is a fundamental technology in the computer vision and is also a crucial component for many high-level artificial intelligence tasks, e.g., object tracking [1] and autonomous driving [2]. Benefited from the tremendous progresses have been made in deep learning, the accuracy of object detection has been greatly improved. Meanwhile, state-of-the-art object detectors also become increasingly more expensive. The high computational complexity seriously hinders their deployment in many real-world applications such as robotics and self-driving cars where model size and latency are highly constrained.

There have been many previous works aiming to the study of how to make trade-off between detection accuracy and computation complexity. Thanks to the powerful parallel processing ability of GPUs, many researchers claimed they have achieved real-time detection on server-class GPUs. However, those methods are hard to achieve real-time inference in mobile scenarios, which are more common in daily life. Consequently, research into fast object detection method on computationally constrained devices is extremely urgent.

To speed up the inference, researchers strive to simplify the detection head and post-processing while retaining accuracy [3, 4]. In a recent study named CenterNet [5], the inference time is further shortened almost the same as the time consumed by the backbone network. CenterNet does not rely on complicated decoding strategies or heavy head designs, which can outperform popular real-time detectors while having faster inference speed. Inspired by the CenterNet, this paper proposes a lightweight and effective object detection method named Mobile CenterNet. CenterNet transforms object detection to the standard keypoint estimation problem, and HRNet [6] is a very strong backbone for tasks related to keypoint estimation. Thus, we choose the smallest HRNet, HRNet-W16, as Mobile CenterNet's backbone. CenterNet's keypoint localization loss, which is a variant of focal loss, only focuses on solving the imbalance problem between positive and negative samples. However, the imbalance problem among different classes in training set also causes performance degradation. Thus, we propose a category-balanced focal loss, which can handle both category imbalance and positive and negative samples imbalance in the training set.

More recently, knowledge distillation (KD) has attracted much attention for its simplicity and efficiency. It can improve the performance of lightweight model without adding any extra computational cost during inference. We introduce knowledge distillation techniques to help the training of our Mobile CenterNet network. Inspired by [7, 8], we apply the structured knowledge distillation to our framework. With the help of structured knowledge distillation, the student network is trained to make its similarity matrix similar to that

$*$ Corresponding author. This work was supported by the National Natural Science Foundation of China (NO.U1736123, NO.61572450) and USTC Research Funds of the Double First-Class Initiative (YD2350002001).

978-1-7281-1485-9/20/$31.00 ©2020 IEEE

of the teacher network. Together with these improvements, our Mobile CenterNet can achieve 44.6 mAP (IOU0.5)/ 7FPS/ 6.04MB on the ivslab dataset with running on the NVIDIA TX2. These performances are very competitive compared to any other state-of-the-art real-time object detectors. Main contributions of this paper are as follows:

(1) Based on the CenterNet, we propose an effective and lightweight Mobile CenterNet for Object detection. The proposed Mobile CenterNet achieves comparable results with other methods, while the number of parameter is 1.62 M.

(2) We propose a category-balanced focal loss, which can handle both category imbalance and positive and negative samples imbalance in the training set.

(3) Without adding any extra computational cost during inference, knowledge distillation is introduced and integrated with the proposed Mobile CenterNet, leading to further improvement in detection accuracy.

2. RELATED WORK

2.1. Object Detection

Current CNN-based object detection consists of anchor-based and anchor-free detectors.

Anchor-based Detector: Anchor-based detectors inherit the ideas from traditional sliding-window. The emergence of Faster R-CNN [9] establishes the dominant position of two-stage anchor based detectors. After that, many works extend Faster R-CNN architecture in many aspects to achieve better performance. For example, R-FCN [10] proposed to use region-based fully convolution network to replace the original fully connected network. FPN [11] proposed a top-down architecture with lateral connections for building high-level semantic feature maps for variant scales. Mask R-CNN [12] extended Faster R-CNN by adding a branch for predicting a pixel-wise object mask in parallel with the original bounding box recognition branch.

With the advent of SSD [3], one-stage anchor-based detectors have attracted much attention because of their high computational efficiency.Thereafter, a large number of researches have paid more attention to bridging the detection accuracy gap between two-stage and one-stage detectors. F-SSD [13] and DSOD [14] exploit different feature fusion methods to ameliorate the weak representation ability of low-level feature. RefineDet [15] introduces an extra loss refinement stage to considerably improve small-object detection accuracy. However, anchor boxes result in excessively many hyper-parameter. These hyper-parameters have shown a great impact on the final accuracy, which typically need to be carefully tuned in order to achieve good performance.

Anchor-free Detector: The most popular anchor-free detector might be YOLOv1 [16]. Instead of using anchor boxes, YOLOv1 predicts bounding boxes at points near the center of objects. CornerNet [17] detects an object bounding box as a pair of keypoints (top-left corner and bottom-right corner). However, CornerNet requires much more complicated post-processing to group the pairs of corners belonging to the same instance. Apart from corner-based anchor-free design, many anchor-free detectors relying on FPN are proposed. F-COS [18] regards all the locations inside the object bounding box as positives with four distances. CenterNet [5] use keypoint estimation to find center point of objects and regress to other properties. It does not rely on complicated decoding strategies or heavy head designs, which can outperform popular real-time detectors while having faster inference speed.

2.2. Knowledge Distillation

The concept of knowledge distillation is introduced by Hinton et al. [19] based on a teacher-student framework. Recently, it has attracted much attention due to its capability of transferring rich information from a large and complex teacher network to a small and compact student network. Originally, KD is used in the task of image classification [19], where a compact model can learn from the output of a large model, namely soft target. So the student is supervised by both softened labels and hard labels simultaneously.

Following [19], some subsequent works have tried to transfer intermediate representations of the teacher network to that of the student network. Ba et al. [20] trained the student network to mimic the teacher via regressing logits before the Softmax layer. Romero et al. [21] proposed FitNet, which extracted the feature maps of the intermediate layer as well as the final output to teach the student network. Zagoruyko et al. [22] defined Attention Transfer (AT) based on attention maps to improve the performance of the student network. Liu et al. [8] propose to distill structured knowledge from large networks to small ones, taking into account the fact that dense prediction is a structured prediction problem, and demonstrate the effectiveness on multiple dense prediction tasks.

3. APPROACH

In this section, we start with an introduction of Mobile CenterNet, and then we take a look at our proposed category-balanced focal loss and knowledge distillation scheme.

3.1. Model Architecture

The architecture of our model is shown in Fig. 1. Our network is based on CenterNet. We choose HRNet as the backbone, which is a strong network for keypoint estimation tasks.

CenterNet treats object detection as consisting of two part: center localization and size regression. For localization, it adopts the Gaussian kernel to produce heatmaps, which enables the network to produce higher activations near the object center. For regression, it defines the pixel at the object center as a training sample and directly predicts the height

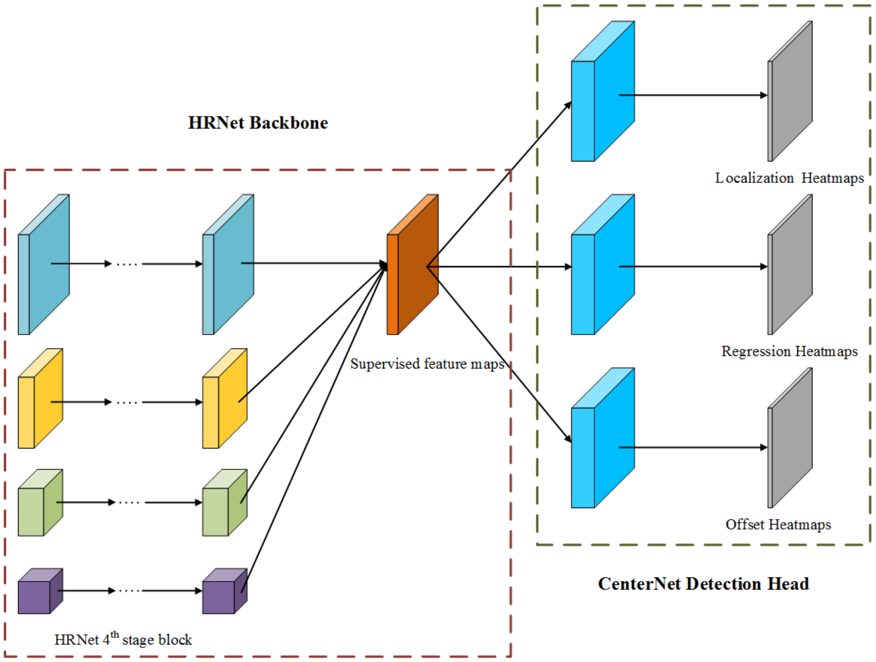

Fig. 1. The architecture of our proposed Mobile CenterNet. We input HRNet's 4^{th} stage highest resolution feature map to the head convolutional layer. Head convolutional layer will produce heatmaps for localization, regression and offset.

and width of the object. It also predicts the offset to recover the discretization error caused by the output stride.

Let $I \in R^{W\times H\times 3}$ be an input image of width W and height H and $(x_1^k, y_1^k, x_2^k, y_2^k)$ be the bounding box of object k with category c. The keypoint $p^k \in \mathbb{R}^2$ represents the center of an object, which is defined as:

$$p^k = \left(\frac{x_1^k + x_2^k}{2}, \frac{x_1^k + x_2^k}{2}\right) \tag{1}$$

A keypoint heatmap $\mathbf{H}_{xyc} \in W' \times H' \times C$ is produced by 2D Gaussian kernel, formulated as:

$$\mathbf{H}_{xyc} = \exp\left(-\frac{(x-\tilde{p}_x)^2 + (y-\tilde{p}_y)^2}{2\sigma_p^2}\right) \tag{2}$$

where $W' = \frac{W}{R}$, $H' = \frac{H}{R}$, $\tilde{p} = \lfloor \frac{p}{R} \rfloor$, R is the output stride, C is the number of object classes, and σ_p is an object size-adaptive standard deviation [17].

Category-Balanced Focal Loss: CenterNet's keypoint localization loss, which is a variant of focal loss [23], can effectively solving the imbalance problem between positive and negative samples. However, it can't deal with the imbalance among different classes in training data, which will cause detection accuracy degradation in rare categories. To this end, we propose a category-balanced focal loss for keypoints localization, which can handle both category imbalance and positive and negative samples imbalance in the training set. The category-balanced focal loss is defined as:

$$L_k = -\frac{1}{N}\sum_{xyc} w_c \begin{cases} (1-\hat{\mathbf{H}}_{xyc})^\alpha \log(\hat{\mathbf{H}}_{xyc}), \text{if } \mathbf{H}_{xyc} = 1 \\ (1-\mathbf{H}_{xyc})^\beta (\hat{\mathbf{H}}_{xyc})^\alpha \log(1-\hat{\mathbf{H}}_{xyc}), \text{otherwise} \end{cases} \tag{3}$$

where α and β are hyper-parameters of the focal loss. N is the number of keypoints in image I. $\hat{\mathbf{H}}_{xyc}$ is the output of the keypoint localization branch. We employ the same method as [24] to obtain the class weight w_c, defined as:

$$w_c = \frac{W-1}{r_{\max}^\gamma - 1}(r_c^\gamma - 1) + 1 \tag{4}$$

where M_c represents the number of labeled boxes for class c. The maximal and minimal number are M_{max} and M_{min}. $r_c = \frac{M_{\max}}{M_c}$ and $r_{\max} = \frac{M_{\max}}{M_{\min}}$. W and γ are two hyper-parameters. The class weight for the most frequent class is 1 and the most rare class is W.

To recover the discretization error caused by the output stride, offset prediction branch is introduced to adjust the center position slightly before remapping the center position to the input resolution and all classes share the same offset prediction branch. The offset is trained with an l_1 loss:

$$L_{off} = \frac{1}{N}\sum_p \left|\hat{O}_{\tilde{p}} - \left(\frac{p}{R} - \tilde{p}\right)\right| \tag{5}$$

where $\hat{O}_{\tilde{p}} \in W' \times H' \times 2$ is the offset prediction for each center point.

To regress the object size $s_k = \left(x_2^k - x_1^k, y_2^k - y_1^k\right)$ for objects, we use a size prediction branch to obtain $\hat{S} \in$

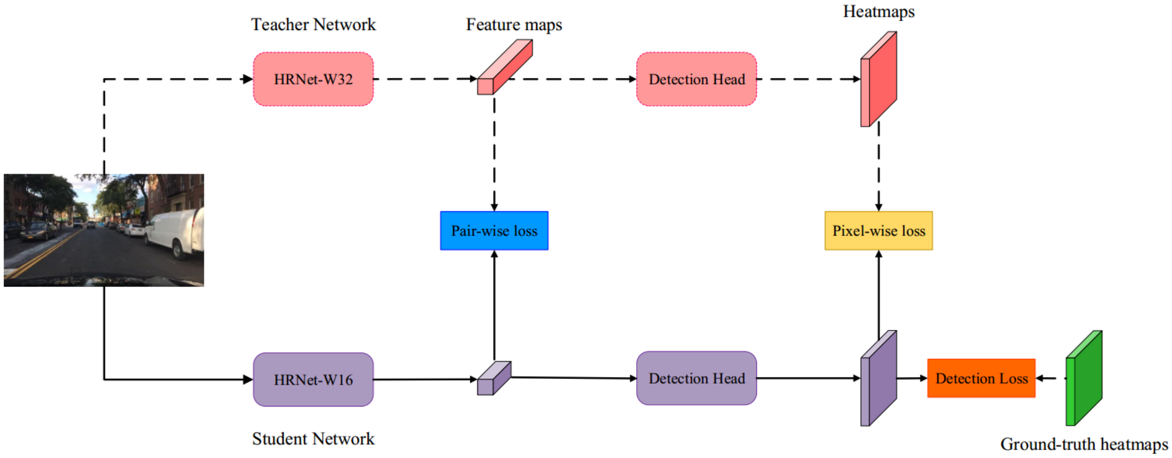

Fig. 2. An overview of our knowledge distillation framework.

$\mathbb{R}^{W'\times H'\times 2}$ for all object categories. The size prediction branch is trained with an l_1 loss at the center point:

$$L_{size} = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{S}_{p_k} - s_k\right| \tag{6}$$

The overall training objective for the detection model is:

$$L_{\text{det}} = L_k + \lambda_{size}L_{size} + \lambda_{off}L_{off} \tag{7}$$

where λ_{size} and λ_{off} are tunable parameters to balance different loss functions.

3.2. Objector Knowledge Distillation

To further improvements in detection accuracy, we introduce and integrate knowledge distillation with our model. Knowledge distillation can transfer useful structure information and semantic information from the teacher network to the student network. An illustration of our knowledge distillation framework is shown in Fig. 2.

CenterNet structure is adopted by the teacher network and the student network. The backbone of teacher network and student network are HRNet-W32 and HRNet-W16 separately. The head convolutional layer in the student network has 32 filters with 3 × 3 kernel, and 64 filters with 3 × 3 kernel in the teacher network. Similar to [8], we also distill the pixel-wise information and pair-wise information from the teacher network to the student network.

The pixel-wise distillation employs heatmaps produced from the cumbersome network as soft targets for training the compact network. The loss function is given as follows:

$$L_{pi} = \frac{\sum\limits_{i\in\Re} KL\left(\mathbf{H}_i^s||\mathbf{H}_i^t\right)}{W'\times H'}, \Re = \{1, 2, \ldots, W'\times H'\} \tag{8}$$

where $\mathbf{H}_i^s$ represents the response of the $i-th$ pixel produced from the student network S, $\mathbf{H}_i^t$ represents the response of the $i-th$ pixel produced from the teacher network T, $KL(\cdot)$ is the Kullback-Leibler divergence between two heatmaps.

The pair-wise distillation transfers structural information from the teacher network to the student network by comparing their similarity matrix. Here, the similarity between two pixels is implemented by cosine similarity. Given a feature map that has dimensions of $W'\times H'\times C$, where C is the total number of channels and $W'\times H'$ is the size of a feature map, $\mathbf{f}_i\in\mathbb{R}^C$ denotes a feature vector extracted from the i spatial location of this feature map. The cosine similarity α_{ij} between $\mathbf{f}_i$ and $\mathbf{f}_j$ is calculated as:

$$\alpha_{ij} = \frac{\mathbf{f}_i^T\cdot\mathbf{f}_j}{\|\mathbf{f}_i\|_2\cdot\|\mathbf{f}_j\|_2} \tag{9}$$

Let α_{ij}^t denote the similarity matrix from the teacher network and α_{ij}^s denote the similarity matrix from the student network. The pair-wise similarity distillation loss is then formulated as:

$$L_{pa} = \frac{\sum\limits_{i\in\Re}\sum\limits_{j\in\Re}\left(\alpha_{ij}^s - \alpha_{ij}^t\right)^2}{\left(W'\times H'\right)^2}, \Re = \{1, 2, \ldots, W'\times H'\} \tag{10}$$

As shown in Fig. 1, we add the pair-wise supervision on the HRNet's 4th stage highest resolution feature map. With knowledge distillation, the student network is trained to optimize the following objective function:

$$L_{KD} = L_{\text{det}} + \lambda_{pa}L_{pa} + \lambda_{pi}L_{pi} \tag{11}$$

where λ_{pa} and λ_{pi} are tunable parameters to balance different loss functions.

4. EXPERIMENTS

In this section, we evaluate the effectiveness of our model on the BDD100K [25] and ivslab benchmark. The experiment is conducted on 2 NVIDIA GTX 2080 Ti with CUDA 10.0 and cuDNN v7.

Table 2. Ablation study of our model based on BDD100K dataset.

Backbone	mAP(IOU@0.5)	PARAMS(M)	GFLOPs	FPS@2080Ti
ResNet-18	41.6	14.4	10.4	71
ResNet-50	49.0	30.7	18.9	45
DLA-34	54.7	19.8	14.4	24
HRNet-W32(Teacher)	55.8	28.6	29.5	16
HRNet-W16(Student)	36.6	1.6	5.7	67
HRNet-W16 + Category-Balanced Focal Loss	38.5	1.6	5.7	67
HRNet-W16 + Category-Balanced Focal Loss + KD	40.2	1.6	5.7	67

Table 1. Final evaluation result of Embedded Deep Learning Object Detection Model Compression Competition.

Team	mAP	Model Size	Complexity	Speed@TX2
	(IOU@0.5)	(MB)	(GOPS/frame)	(ms/frame)
USTC-NELSLIP	44.60	6.04	11.16	141.8
BUPT_MCPRL	49.20	7.35	12.89	401.21
DD_VISION	25.60	0.86	0.52	56.18
Deep Learner	49.00	45.20	56.64	1560.43
IBDO-AIOT	38.70	187.27	285.7	287.11
ACVLab	41.10	87.46	31.82	696.72

4.1. Dataset and Implementation details

The BDD100K dataset [25] is a large-scale autonomous driving dataset. It consists of ten classes: bike, bus, car, motor, person, rider, traffic light, traffic sign, train, and truck. The ratio of training and validation set is 7:1 and image resolution is 1280 × 720. The ivslab dataset is the dataset for Embedded Deep Learning Object Detection Model Compression Competition. It provide 89,002 unannotated 1920 × 1080 images for training and 2,700 annotated images for testing. The objects in the dataset are annotated with four categories: pedestrian, vehicle, scooter and bicycle. In the competition, since training images of ivslab dataset are extracted from videos, most of them are similar due to temporal correlation. Therefore, we choose one out of ten pictures as our training set. We pre-train our model on the BDD100K, and then fine-tune it on the ivslab dataset. Specifically, we convert BDD100K classes to be consistent with ivslab. We divide bus, car and truck into unified vehicle category, and ignore the category of traffic light, traffic sign and rider. In our experiment, the IOU threshold of the Mean Average Precision (mAP) is set to 0.5, which is also adopted by the competition.

We implement the model based on PyTorch. The data augmentation and optimizer are kept the same as [5]. The initial learning rate is set to 0.001, and is divided by 10 at epoch 45 and 55 respectively. The whole training takes 60 epochs. All input images are resized to 576 × 320 pixels. The hyper-parameters of α, β, γ, W, λ_{size}, λ_{off}, λ_{pa} and λ_{pi} are set to 2, 4, 0.75, 10, 0.1, 1, 10 and 0.1 respectively.

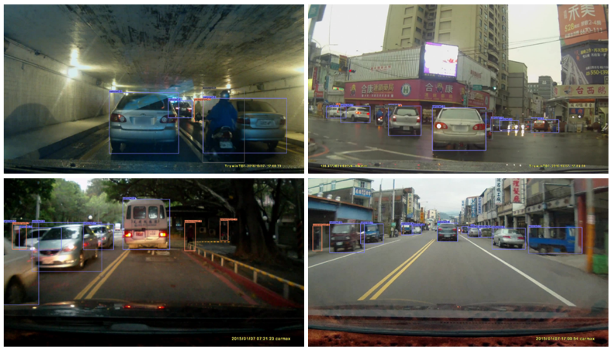

Fig. 3. Detection results on the ivslab dataset.

4.2. Results on the Competition Dataset

The officially announced competition results are shown in Table 1. It includes the evaluation of accuracy, model size, computation complexity and speed on NVIDIA Jetson TX2, respectively. Our team USTC-NELSLIP finally achieves the 1st place by our accuracy-speed trade-off method. Qualitative results of our model are shown in Fig. 3.

4.3. Ablation Studies on BDD100K

BDD100K validation set is utilized for the performance evaluation. The results are exhibited in Table 2. The basic HRNet-W16 based CenterNet achieves similar performance with ResNet-18 based CenterNet, while HRNet-W16 based model outperforms it with 45% of the FLOPs and 89% of the model size. However, the inference speed of ResNet-18 based model is 6% faster. This may be ResNet has been efficiently optimized by the current deep learning framework, and HRNet has only recently been proposed, it is not well optimized. However, lower computational complexity and smaller model size make HRNet based model more potential than ResNet based model. Without adding any extra size and computation cost, Category-Balanced Focal Loss and knowledge distillation can facilitate the model to get better performance.

5. CONCLUSION

In this paper, we propose the Mobile CenterNet for embedded systems. Our method is based on CenterNet. To enhance detection performance, we adopt HRNet as a powerful backbone and improve CenterNet's keypoint localization loss as category-balanced focal loss, which can deal with both imbalance between positive and negative samples and imbalance among different classes in the training data. What's more, we employ knowledge distillation to transfer knowledge from large network to small network. By this way, the performance of Mobile CenterNet is further improved without adding any extra computational cost during inference. As a result, the size of Mobile CenterNet is 6 MByte, with the inference speed of 7 FPS on NVIDIA Jetson TX2. The accuracy of our model achieves 44.6 mAP on the ICME2020 Competition final testing dataset.

6. REFERENCES

[1] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan, "Poi: Multiple object tracking with high performance detection and appearance feature," in *European Conference on Computer Vision*. Springer, 2016, pp. 36–42.

[2] Bichen Wu, Forrest Iandola, Peter H Jin, and Kurt Keutzer, "Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 129–137.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[4] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[5] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[6] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[7] Quanquan Li, Shengying Jin, and Junjie Yan, "Mimicking very efficient network for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6356–6364.

[8] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2604–2613.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[10] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

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

[13] Zuoxin Li and Fuqiang Zhou, "Fssd: feature fusion single shot multibox detector," *arXiv preprint arXiv:1712.00960*, 2017.

[14] Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue, "Dsod: Learning deeply supervised object detectors from scratch," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1919–1927.

[15] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4203–4212.

[16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[17] Hei Law and Jia Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[18] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9627–9636.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[20] Jimmy Ba and Rich Caruana, "Do deep nets really need to be deep?," in *Advances in neural information processing systems*, 2014, pp. 2654–2662.

[21] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[22] Sergey Zagoruyko and Nikos Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[24] Chen Chen, Mengyuan Liu, Xiandong Meng, Wanpeng Xiao, and Qi Ju, "Refinedetlite: A lightweight one-stage object detection framework for cpu-only devices," *arXiv preprint arXiv:1911.08855*, 2019.

[25] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," .