TRANSACTIONS OF THE INSTITUTE OF MEASUREMENT & CONTROL

Article

Transactions of the Institute of Measurement and Control
2021, Vol. 43(13) 2888–2898

DOI: 10.1177/01423312211022877
journals.sagepub.com/home/tim
SAGE

Revisiting knowledge distillation for light-weight visual object detection

Tianze Gao[1], Yunfeng Gao[1,2], Yu Li[1] and Peiyuan Qin[2]

Abstract

An essential element for intelligent perception in mechatronic and robotic systems (M&RS) is the visual object detection algorithm. With the ever-increasing advance of artificial neural networks (ANN), researchers have proposed numerous ANN-based visual object detection methods that have proven to be effective. However, networks with cumbersome structures do not befit the real-time scenarios in M&RS, necessitating the techniques of model compression. In the paper, a novel approach to training light-weight visual object detection networks is developed by revisiting knowledge distillation. Traditional knowledge distillation methods are oriented towards image classification is not compatible with object detection. Therefore, a variant of knowledge distillation is developed and adapted to a state-of-the-art keypoint-based visual detection method. Two strategies named as positive sample retaining and early distribution softening are employed to yield a natural adaption. The mutual consistency between teacher model and student model is further promoted through a hint-based distillation. By extensive controlled experiments, the proposed method is testified to be effective in enhancing the light-weight network's performance by a large margin.

Keywords

Visual object detection, artificial neural network, knowledge distillation

Introduction

Intelligent perception systems are crucial for industrial applications that involve autonomous operations (Pan and Sun, 2018; Pan et al., 2017; Sahu and Subudhi, 2017; Zhang et al., 2020). Given an input image captured by an industrial camera, a robotic system is usually asked to accurately perceive the categories and locations of the objects in sight. Such issues can be addressed by visual object detection algorithms.

With the fast advance of artificial intelligence (AI) theories and hardware platforms, current visual object detection methods are dominated by paradigms based on convolutional neural networks (CNN). A milestone of modern object detection methods is the two-stage faster R-CNN network proposed by Ren et al. (2015), who firstly designed a region proposal network (RPN) to reduce the searching space. However, faster R-CNN is computationally expensive and not suitable for applications requiring a high running speed. Single-stage methods, for example SSD (2D object detection method) proposed by Liu et al. (2016) and YOLOv3 proposed by Redmon and Farhadi (2018), run much faster yet have inferior performance. Lately, Zhou et al. (2019) presented a keypoint-based detection framework named as CenterNet, which achieved a decent balance between the accuracy and runtime. In the paper, the CenterNet is used as the baseline. Also, our purpose is to generate a light-weight counterpart of it while minimizing the performance drop as much as possible. Such light-weight models are essential for the deployment in mechatronic and robotic systems (M&RS) since an industrial computer is usually answerable to multiple tasks, with visual object detection merely being one of the many modules.

Researchers have proposed various approaches to model compression, including model pruning, depthwise separable convolution, quantization, and knowledge distillation (Hinton et al., 2015). As a propitious research topic, knowledge distillation is known for its capability in delivering expertise of a teacher model to a light-weight student model. To be specific, a well trained network can be compressed into a light-weight one, for example by replacing its backbone with fewer layers. A direct consequence is a severe drop in performance. The ideology of knowledge distillation is training the light-weight network (student network) under the guidance of the soft labels output by the cumbersome network (teacher network), leading to improved performance compared with training barely using the ground truth labels.

Knowledge distillation has been extensively applied to the scope of image classification. However, this paradigm has been insufficiently investigated for visual object detection, mainly due to the challenge posed by an imbalanced distribution between positive and negative data. Chen et al. (2017)

[1]Harbin Intitute of Technology, China
[2]HIT-Wuhu Robot Technology Research Institute, China

Corresponding author:
Yunfeng Gao, School of Mechatronics Engineering, Harbin Institute of Technology, No. 92, Xi Dazhi Street, Harbin, 150001, China.
Email: gyf@hit.edu.cn

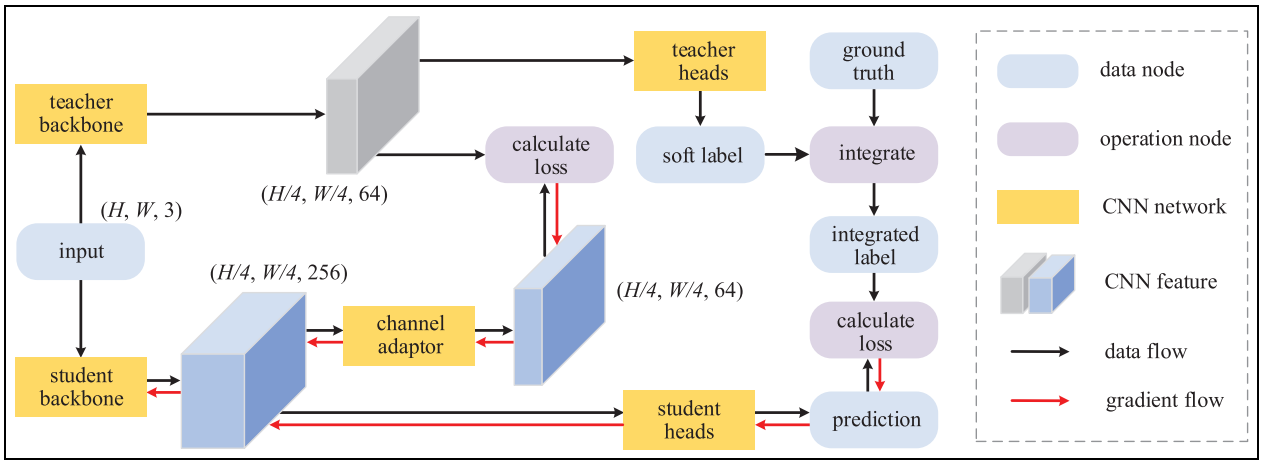

Figure 1. General pipeline of the presented framework.
An SoTA method, CenterNet, is employed to be the teacher network. The student is derived by replacing teacher's backbone with a light-weight one and also reducing the output channels of teacher's heads. The student's training procedure receives guidance from both the integrated labels and the last layer of teacher's backbone (please refer to the 'Method' section).

proposed to tackle this issue by manually assigning different weights to background and foreground, yet this method is limited to faster R-CNN and incompatible with the focal loss (Lin et al. (2017b)), which is employed by most single-stage detection frameworks.

In this paper, the formulation of knowledge distillation under the context of visual detection is firstly developed. A heuristic algorithm is developed for the purpose of adapting the obtained soft labels to focal loss, wherein two strategies named as positive sample retaining and early distribution softening are adopted. Then, the last layer of teacher network's backbone is employed as a hint to guide the student model. The proposed hint-based distillation differs from that of Romero et al. (2014), who utilized an intermediate layer as the hint layer. The reason for this designing difference is further explained in the 'Method' section. Finally, a dynamic training scheme is devised, whereby the influence of teacher network gradually fades out as the training process goes. This training scheme has the advantage that the student network is unlikely to be excessively guided by the teacher network. To corroborate the rationality of the presented algorithm, a proper dataset has to be selected. Common datasets like ImageNet, MS COCO and PASCAL VOC are not appropriate since the proposed method is targeted at applying to robotic systems (Pan et al., 2018). Thus, we conduct extensive experiments using KITTI (Geiger et al., 2012), which is the most prestigious dataset for self-driving cars (also known as intelligent wheeled robots). The general pipeline of the presented framework is depicted in Figure 1.

The remainder of paper is arranged as follows. In 'Related work', previous works that have direct relation to our work are briefly reviewed from two aspects, namely visual object detection and knowledge distillation. The formalism and strategies adopted in our framework are elaborated on in the 'Method' section. In the 'Experiment' section, the evaluation benchmark, implementation details, and experimental results are described. At the end of the paper, a summary of the contributions and limitations of this work is made in the 'Conclusion' section.

Related work

In this section, we first make a review and analysis on modern visual object detection methods. Then related works in the scope of knowledge distillation are introduced, where the relationship and difference between previous works and ours are also explained.

Visual object detection

Traditional visual object detection algorithms resort to feature descriptors such as Histogram of Oriented Gradients (HOG), Haar-like, and Scale-Invariant Feature Transform (SIFT). Accompanying the remarkable achievement of CNN, modern visual object detection is occupied with CNN-based methods, which are categorized as single-stage methods and two-stage ones according to the detection pipeline. Girshick et al. (2014) firstly introduced CNN features into the field of visual object detection. By leveraging the selective search algorithm to generate potential object candidates and an SVM for object classification, the presented R-CNN network achieved a 30% improvement over detectors based on traditional feature descriptors. Later, Ren et al. (2015) presented faster R-CNN that boosted the performance and efficiency of R-CNN through a RPN. Since then, faster R-CNN has been a baseline for many two-stage detection frameworks. For example, Lin et al. (2017a) embedded a feature pyramid network (FPN) into faster R-CNN to better detect objects with various scales.

Although two-stage detection methods generally yield high detection accuracy, they have the drawbacks of large model sizes and slow running speeds. In this regard, Redmon et al. (2016) firstly designed a single-stage framework which phrased the object detection into a regression problem by

treating input image as an evenly arranged mesh. Liu et al. (2016) further proposed to predict boxes of different scales from different layers of the backbone feature. Despite of the high efficiency, these single-stage methods yielded much worse results compared with two-stage counterparts. Lin et al. (2017b) pointed out the poor performance was attributed to the imbalanced distribution of positive and negative samples and thus proposed the focal loss to address this issue. The focal loss has proven to be quite effective so that it is adopted by most single-stage successors. A major contribution of our work is adapting knowledge distillation to a variant of focal loss.

Most of the advanced object detectors, whether being two-stage or single-stage, have leveraged the power of anchor points (i.e., predefined object centers) and anchor boxes (i.e., candidates of predefined sizes and aspect ratios). This kind of manual design increases the training time and brings about extra hyperparameters, hence motivating the emergence of more recent anchor-free detectors. Law and Deng (2018) firstly proposed to regress the top-left and bottom-right corners of target boxes. An embedding vector was also predicted for corner grouping. Tian et al. (2019) designed a fully-convolutional network that estimated a dense map indicating the distance from each point to the corresponding bounding box's edges. The state-of-the-art method CenterNet (Zhou et al., 2019) is an exception lying between the anchor-based and anchor-free paradigms. It is a keypoint-based framework that simultaneously estimated the center points (can be viewed as implicit anchor points) via a keypoint head and the sizes and discretization offsets via regression heads. In this work, the CenterNet is chosen as the baseline due to its balanced performance in accuracy and runtime.

Besides, some researchers have also exploited other novel approaches to visual object detection. For example, Carion et al. (2020) combined CNN features with an encoder-decoder transformer (Vaswani et al., 2017). We do not deep dive into these methods since they are not directly related to our work.

Knowledge distillation

Knowledge distillation was pioneered by Hinton et al. (2015) and originally used as a model compression method for image classification problems. It was found that by integrating soft labels generated by a pretrained network during the optimizing procedure, the performance of a light-weight student network could be enhanced. Based on such philosophy, Romero et al. (2014) further proposed the FitNet, which introduced extra supervision by comparing the intermediate layers of teacher's and student's backbone networks. A similar approach is adopted in this work but we distill the last layer's output of backbone network due to the distinctive architectures of the student and teacher.

Similar to FitNet, variants of knowledge distillation aiming at transferring various types of information were developed by later researchers. Yim et al. (2017) offered to distill the solution procedure flow via calculating Gram matrices across different feature layers. Vaswani et al. (2017) proposed to transfer attention maps to the student network and compared three different ways in calculating the attention maps. Heo et al. (2019) pointed out that the activation states of neurons instead of their specific values could be more efficient in transferring knowledge. Park et al. (2019) designed a relational potential function that facilitated transferring the mutual relations of teacher's output to the student. With a similar notion, Lassance et al. (2020) built graphs for both the student and teacher. Latent representation geometry was then transferred by measuring the discrepancy between corresponding adjacency matrices.

Despite that previous works have attained remarkable accomplishments in the area of image classification, only a few of them are oriented towards visual object detection. The reason for this phenomenon mainly lies in the challenge posed by the unevenly distributed numbers of foreground and background labels. Wang et al. (2019) circumvented such an issue by encouraging the student network to imitate the fine grained features of teacher network, which could be viewed as a reformed version of FitNet. The work most relevant to ours was presented by Chen et al. (2017), who proposed to cope with the label imbalance by manually assigning different weights to foreground and background losses. However, this method is limited to the two-stage faster R-CNN and incompatible with the focal loss, which is commonly employed by single-stage detection frameworks. By contrast, we propose to remould the traditional knowledge distillation method so that it adapts to the focal loss, by which the solution to label imbalance is readily available.

Method

In this section, the baseline network CenterNet is firstly introduced. Then the formulation is developed for integrating traditional knowledge distillation into the CenterNet framework. Two strategies regarding the adaption to focal loss and the adaption of temperature are put forward. Finally, it is elaborated on how we use teacher networks' convolutional feature as a hint to guide the student network together with the dynamic training scheme.

Recap of baseline method

Overview. The overall architecture of the baseline network (Zhou et al., 2019) is illustrated in Figure 2. In general, it is composed of a backbone network followed by multiple head networks. In terms of two-dimensional (2D) visual object detection, a keypoint branch is designed to predict the discretized locations of object centers, and a regression branch containing two regression heads are responsible for predicting the sizes (width & height) and discretization offsets.

Backbone. The backbone consists of a DLA-34 (Yu et al., 2018)) and an upsampling network. The stride of the backbone is 4, that is, given an input image in size $H \times W \times 3$, a $H/4 \times W/4 \times 64$ volume will be produced, with 64 being the amount of output channels. The resulting backbone feature is then processed by the head networks for further prediction.

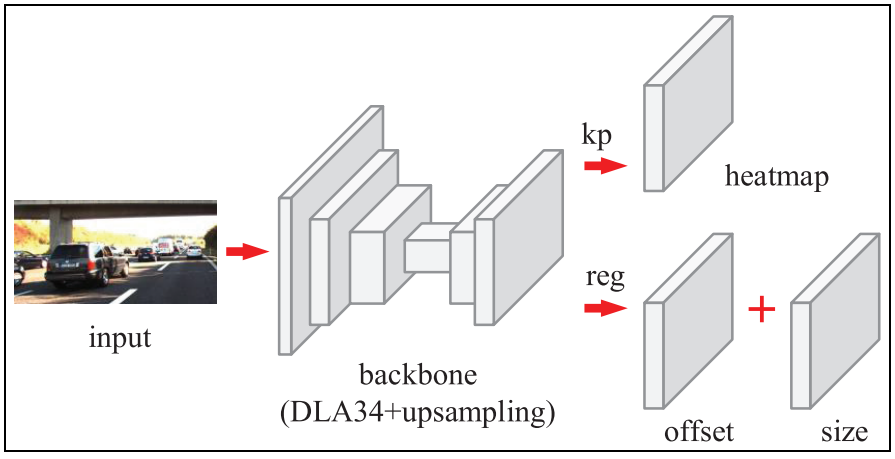

Figure 2. Overall architecture of CenterNet. *kp* and *reg* stand for *keypoint branch* and *regression branch*, respectively.

Keypoint branch. The keypoint branch is a two-layer head network, whose output is a $H/4 \times W/4 \times C$ heatmap, with C being the amount of object categories (in KITTI dataset, $C = 3$ for *car*, *pedestrian* and *cyclist*). The value of each element in the heatmap indicates the possibility of the existence of a class-specific object center.

Regression branch

The regression branch consists of two two-layer head networks. Since the resolution of the predicted heatmap is four times smaller than the input image, discretization errors will occur if the locations of object centers are solely determined by the keypoint branch. Out of such motivation, a regression head is designed to predict these discretization offsets. The function of the other regression head is predicting the width and height of objects at corresponding positions.

Adapting knowledge distillation to visual object detection

Task description. We intend to design a light-weight counterpart of CenterNet that befits the real-time scenarios. To this end, the backbone structure DLA-34 is replaced with ResNet-18 (He et al., 2016) and the output channels of head networks are reduced from 256 to 64, leading to around 2.5 times improvement for the overall running speed. An emerging problem is that the light-weight network suffers from severe performance drop if trained solely using ground truth data. Thus, we propose to mitigate this problem through knowledge distillation. Specifically, the original CenterNet is leveraged as the teacher, while the light-weight network (the one with ResNet-18 as backbone) is treated as the student to be guided in the training stage.

Distillation with soft labels. In the preceding subsection, it has been introduced that the baseline network locates object centers by a heatmap. In our formulation, the task of generating a heatmap is treated as a pixel-wise classification problem. Let $\mathcal{Y} = \{p_1, p_2, \ldots, p_N\}$ be the ensemble of pixelwise labels p_i in the ground truth heatmap with N denoting the cardinality, and accordingly $\hat{\mathcal{Y}} = \{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_N\}$ be the prediction of the student network. For a single pixel that is either an object center with the label $y_i = 1$ or not an object center with the label $y_i = 0$, the problem naturally falls into a 0/1 classification problem. From this perspective, the cross entropy loss is conventionally employed as a performance indicator. Without loss of generality, the ensemble loss $L(\mathcal{Y}, \hat{\mathcal{Y}})$ is henceforth represented by pixelwise loss $L(p, \hat{p})$ for notational simplicity

$$L(p, \hat{p}) = -[p\log(\hat{p}) + (1-p)\log(1-\hat{p})]. \tag{1}$$

In view of knowledge distillation, the cross entropy loss turns into L_{KD} by incorporating an additional loss term $L_{\mathrm{TE}}(p_{\mathrm{TE}}, \hat{p})$ as follows

$$L_{\mathrm{KD}}(p, \hat{p}, p_{\mathrm{TE}}) = \gamma L(p, \hat{p}) + (1-\gamma)L_{\mathrm{TE}}(p_{\mathrm{TE}}, \hat{p}), \tag{2}$$

where p_{TE} is the soft label output by the teacher model, and γ is a weighting factor.

Let $P_{\mathrm{TE}}(x)$ and $P(x)$, respectively, denote the class distribution of the pixel x with the probabilities of p_{TE} and $\hat{p}$. The $L_{\mathrm{TE}}(p_{\mathrm{TE}}, \hat{p})$ is then instantiated as the Kullback-Leibler divergence for $P_{\mathrm{TE}}(x)$ and $P(x)$, namely

$$\begin{aligned} L_{\mathrm{TE}}(p_{\mathrm{TE}}, \hat{p}) &= \sum_{\{0,1\}} P_{\mathrm{TE}}(x)\log\frac{P_{\mathrm{TE}}(x)}{P(x)} \\ &= (1-p_{\mathrm{TE}})\log\frac{1-p_{\mathrm{TE}}}{1-\hat{p}} + p_{\mathrm{TE}}\log\frac{p_{\mathrm{TE}}}{\hat{p}} \\ &= -[p_{\mathrm{TE}}\log(\hat{p}) + (1-p_{\mathrm{TE}})\log(1-\hat{p})] + C, \end{aligned} \tag{3}$$

where $C = p_{\mathrm{TE}}\log(p_{\mathrm{TE}}) + (1-p_{\mathrm{TE}})\log(1-p_{\mathrm{TE}})$. Since C provides no gradient information for the student model, it can be safely discarded from the equation.

By substituting equation (1) and equation (3) into equation (2), we have

$$\begin{aligned} L_{\mathrm{KD}}(p, \hat{p}, p_{\mathrm{TE}}) =& -[\gamma p + (1-\gamma)p_{TE}]\log(\hat{p}) \\ & - \{1 - [\gamma p + (1-\gamma)p_{TE}]\}\log(1-\hat{p}). \end{aligned} \tag{4}$$

To have a clearer observation, define $\tilde{p} = \gamma p + (1-\gamma)p_{\text{TE}}$. Then, we have

$$L_{\text{KD}}(p,\hat{p},p_{\text{TE}}) = L_{\text{KD}}(\tilde{p},\hat{p}) = -[\tilde{p}\log(\hat{p}) + (1-\tilde{p})\log(1-\hat{p})]. \tag{5}$$

By such formulation, it is observed that without considering *temperature* (will be discussed later), the intrinsic nature of knowledge distillation can be understood as training the student network with a linear combination of the ground truth and the soft labels from teacher network. This conclusion is important since it provides a natural pathway for adapting knowledge distillation to the focal loss.

Adaption to focal loss. Focal loss was pioneered by Lin et al. (2017b). It has been broadly adopted in single-stage detectors for alleviating the imbalance in positive and negative training samples. A more dedicated variant of focal loss proposed by Law and Deng (2018) is formulated as

$$L^{\text{focal}}(p,\hat{p}) = \begin{cases} -(1-\hat{p})^{\alpha}\log(\hat{p}) & \text{if } p = 1, \\ -\hat{p}^{\alpha}(1-p)^{\beta}\log(1-\hat{p}) & \text{otherwise.} \end{cases} \tag{6}$$

where α and β are hyperparameters. In all our experiments, α and β are set as $\alpha = 2$, $\beta = 4$ following the baseline. By the developed formulation of equation (5), the knowledge distillation can be adapted to focal loss by simply replacing the ground truth label p with the integrated label $\tilde{p}$. However, this simple adaption strategy does not work as it appears. Because the soft labels from the teacher network are obtained by a sigmoid function (which is equivalent to a softmax function in our case), the resulting values are always smaller than 1, causing the integrated labels $\tilde{p}$ to be always smaller than 1 as well. Therefore, all the samples, including the originally positive ones, will turn to negative samples after the label integration.

The problem is addressed by a simple strategy named as positive sample retaining. As the name implies, we retain the positive samples in ground truth by keeping their original labels as 1 without undergoing the linear combination. The adapted loss function is then formulated as

$$L_{\text{KD}}^{\text{focal}}(\tilde{p},\hat{p}) = \begin{cases} -(1-\hat{p})^{\alpha}\log(\hat{p}) & \text{if } \tilde{p} = 1, \\ -\hat{p}^{\alpha}(1-\tilde{p})^{\beta}\log(1-\hat{p}) & \text{otherwise,} \end{cases} \tag{7}$$

with

$$\tilde{p} = \begin{cases} p & \text{if } p = 1, \\ \gamma p + (1-\gamma)p_{\text{TE}} & \text{otherwise.} \end{cases} \tag{8}$$

Adaption of temperature. Another important factor to be adapted is the temperature. Hinton et al. (2015) proposed to adjust the extent of *softness* of the probability distribution by a variable T called temperature

$$p_i = \frac{\exp(l_i/T)}{\sum_j \exp(l_j/T)}, \tag{9}$$

where p_i is the predicted probability of category i, and l_i is the corresponding logit. Here *softness* can be intuitively interpreted as how sure the network is for its own classification results. Keeping constant the predicted logits, if adopting a larger T, the network will be less sure about the predicted probabilities and the predictions will get softer. In many cases, softer predictions are able to convey more about the untold secrets in a neural network. As is depicted in Figure 3, more hidden information (sometimes referred to as *dark knowledge*) emerges as the temperature T increases. The *dark knowledge* is informative and valuable for network training, because it forms directly from network's inference procedure and contains implicit information like which part of the background looks more like foreground compared with other parts.

For traditional knowledge distillation, the temperature is simultaneously applied to student's output $\hat{p}$ and soft labels p_{TE}, so that equation (4) turns into

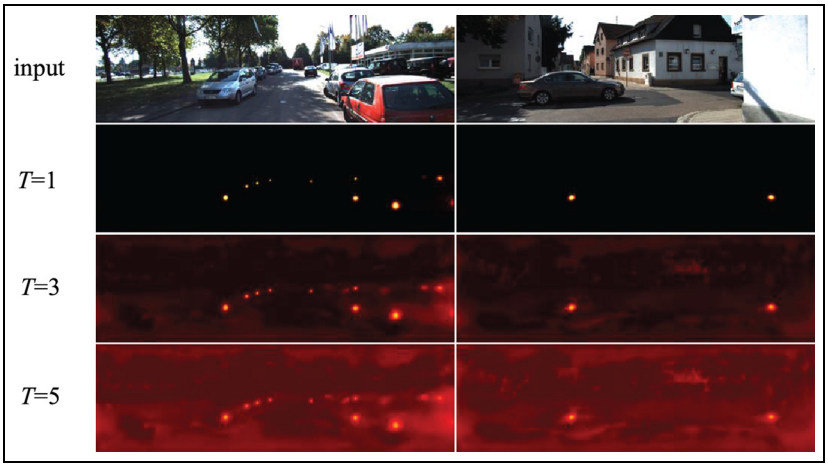

Figure 3. Visualization of heatmaps output by the teacher network under different temperatures.
Brighter areas correspond to elements with larger values and vice versa. The heatmaps are resized to match the input image's resolution for better visualization.

$$L_{\mathrm{KD}}(p,\hat{p},p_{\mathrm{TE}}) = \gamma p\log(\hat{p}) + (1-\gamma p)\log(1-\hat{p}) \\ + (1-\gamma)y_{TE}|_T\log(\hat{p}|_T) + [1-(1-\gamma)y_{TE}|_T]\log(1-\hat{p}|_T). \quad (10)$$

In this way, it is easily seen that the equation (5) no longer holds. Our solution to this problem is that we treat the softened probability $y_{TE}|_T$ as the original output of the teacher network. And the temperature is no more involved into the formulation so that the issue of $\hat{p}|_T$ is naturally bypassed.

Assume

$$y_{TE} = (1+\exp(-l_{TE}))^{-1}, \quad (11)$$

where l_{TE} is the predicted logit for y_{TE}. By considering the temperature T, we have

$$y_{TE}|_T = (1+\exp(-l_{TE}/T))^{-1}. \quad (12)$$

By comparing equation (11) and equation (12), the relationship between y_{TE} and $y_{TE}|_T$ is easily obtained, thereby equation (8) turns into

$$\tilde{p} = \begin{cases} p & \text{if } p = 1, \\ \gamma p + (1-\gamma)[1+(p_{\mathrm{TE}}^{-1}-1)^{1/T}]^{-1} & \text{otherwise.} \end{cases} \quad (13)$$

Contrast to the traditional knowledge distillation which softens the labels when calculating the loss function, we soften the probability distribution of teacher network's output beforehand, hence the strategy is named as early distribution softening. It should be noted that our adaption strategy only makes sense together with the previously developed formulation for label integration, that is, $\tilde{p} = \gamma p + (1-\gamma)p_{\mathrm{TE}}$.

Guiding with CNN feature

Hint-based distillation,. Similar to FitNet (Romero et al., 2014), the teacher network's backbone feature is employed as an extra hint for instructing the student network. Its purpose is to promote the mutual consistency of the two networks by aligning the final layer of the student's backbone (ResNet-18) with the teacher's backbone (DLA-34). In our case, the last layer of the student's backbone has a dimension of $H/4 \times W/4 \times 256$ and that of the teacher's is shaped as $H/4 \times W/4 \times 64$, where H, W correspond to the size of network input. Due to the mismatch in feature dimension, a single CNN layer with its input / output channels being 256 / 64 is leveraged as the channel adaptor.

To measure the consistency between the hint and the guided layer, an element-wise mean square loss function is adopted

$$L_{hint} = \sum_{x,y,c}(z_{TE}^{xyc} - z_{ST}^{xyc})^2, \quad (14)$$

where z_{TE}^{xyc} and z_{ST}^{xyc} denote the element values at location (x,y,c) of the final output of teacher's and student's backbones, respectively.

Difference from FitNet. The major difference between FitNet and the presented method is FitNet transfers knowledge through a middle layer, whereas our method adopts the final layer's output as a hint. This modification is owing to the structural differences between the teacher and student from two aspects:

(i) Compared with the student's backbone, the teacher's backbone adopts extra structural design of hierarchical deep aggregation (HDA) and iterative deep aggregation (IDA).
(ii) Following Zhou et al. (2019), deformable convolutional layers are utilized in the upsampling subnetwork of teacher's backbone. For the student's backbone, we discard the deformable convolutional layers to reduce training parameters.

In view of the prominent differences lying between the teacher's and student's backbones, it is difficult to find a pair of well matched layers from each side. Even worse, forcing a mismatched pair of features to consist with each other could severely hamper the training dynamics, since different layers extract visual information in different abstract levels and also have different sizes of perception fields. Instead, it is more reasonable to view the backbones with distinct architectures as two black boxes, and we only care about the consistency between the final layers' output. To verify the effectiveness of our modification, an ablation study is designed and reported in Table 7.

Joint optimization. The previous work FitNet adopts a stage-wise optimizing strategy, that is, the student network is pre-optimized for some epochs solely with the hint-based loss L_{hint}. Then the hint branch is discarded so that the student is optimized using the knowledge distillation loss L_{KD} alone. Conversely, we jointly optimize the hint-based and soft label-based target functions. It is found by experiment that this joint optimization strategy performs better in our framework (Table 8).

Network training

Loss function. The loss function used for parameter optimization include three parts:

(i) Classification loss. The classification loss in our method is the previously developed $L_{\mathrm{KD}}^{\mathrm{focal}}$ in equation (7), which reflects the error between the predicted 0/1 classification results (or equally, the predicted heatmap) and the integrated labels obtained by equation (13).
(ii) Regression loss. Unlike the classification problem, the regression labels can not be softened via the temperature. Thus, we only supervise the regression heads of the student network with the ground truth labels. Following CenterNet, the L_1 loss is employed to measure the distance between predictions and labels

$$L_{\mathrm{wh}} = \sum_{x,y,c}|w^{xyc} - \hat{w}^{xyc}| + |h^{xyc} - \hat{h}^{xyc}|, \\ L_{\mathrm{off}} = \sum_{x,y,c}|d_u^{xyc} - \hat{d}_u^{xyc}| + |d_v^{xyc} - \hat{d}_v^{xyc}|, \quad (15)$$

wherein (w, h) and $(\hat{w}, \hat{h})$ denote ground truth and estimated object sizes, (d_u, d_v) and $(\hat{d}_u, \hat{d}_v)$ refer to the ground truth and estimated discretization offsets, (x, y, c) is the locational index.

(iii) Hint loss. The hint loss equation (14) reflects the consistency between the last layers of teacher's and student's backbones.

Combining all three parts, the total loss function is given as

$$L_{\text{total}} = L_{\text{KD}}^{\text{focal}} + \omega_1 L_{\text{wh}} + \omega_2 L_{\text{off}} + \omega_3 L_{\text{hint}}, \tag{16}$$

with $w = [1, \omega_1, \omega_2, \omega_3]^{\text{T}}$ as the weighting vector.

Dynamic training scheme. Concerning that the student network is liable to be excessively instructed by the teacher network, a dynamic training scheme is adopted. Specifically, the variable γ in equation (13) are kept constant for the first N training epochs, then it linearly increases to 1 in the remaining epochs. In other words, the influence imposed by the teacher network is gradually discarded during training. It is found by experiment that the dynamic training scheme slightly increases the final performance of the student network.

Experiment

In this section, the benchmark used for evaluating our method is firstly introduced. Then the implementation details are presented, including training details, parameter settings and hardware setups. After that, the quantitative results of controlled experiments are compared and analyzed. Finally, some visualized detection results are presented for an intuitive exhibition.

Evaluation benchmark

Dataset. Modern public visual object detection datasets are mostly targeted at the tasks of generic object detection. The contents of them are not suitable for industrial applications. In this regard, we choose the famous self-driving dataset KITTI (Geiger et al., 2012) to evaluate the presented approach.

The KITTI dataset comprises 7481 images for training and 7518 images for testing. Objects of three categories are evaluated, including car, pedestrian and cyclist. All objects are classified into *Easy*, *Moderate* and *Hard* regarding detection difficulty levels.

Note that the ground truth labels of KITTI's testing set are inaccessible to its users. The users are allowed to submit the testing results to the official website for performance evaluation for only one time. This policy heavily limits the conduction of controlled experiments. Therefore, we follow Chen et al. (2015) who split the training dataset into two subsets: a training subset with 3712 images and a validation subset with 3769 for images.

Metric. Average precision (AP) is employed as the evaluation metric for the experimental results. As a comprehensive metric that considers both the detection precision and recall, the AP scores well reflect networks' performance. Following Simonelli et al. (2019), the 40-point Interpolated Average Precision ($\text{AP}|_{\text{R}_{40}}$) scores are reported in all our experiments.

Adhering to the convention of KITTI benchmark, the intersection over union (IoU) of 0.7 is set as the threshold of identifying a true positive (TP) for the car class. For the pedestrian and cyclist class, the IoU = 0.5 is used.

Implementation details

Training details. For the pretraining of teacher network, we follow the original training pipeline of CenterNet. The only difference is that the three-dimensional heads for predicting dimension, depth and orientation are removed since we focus on 2D detection. Readers are referred to Zhou et al. (2019) for more training details.

Two versions of student networks are trained for comparison. The first version (denoted as *ST-plain*) is trained with an identical configuration to the teacher network, without using knowledge distillation. For the second version (denoted as *ST-KD*), knowledge distillation is utilized and the hint loss is additionally incorporated for the training procedure. Accordingly, weighting factors in the total loss are adjusted due to the changing of loss items. All the other training settings, including the choice of optimizer, learning rate, training epochs, data augmentation, weight decay, and so forth, are configured following the training stage of the teacher network to yield a fair comparison.

Parameter settings. The hyperparameters in focal loss and weighting factors in equation (16) are listed in Table 1. For all the experiments, we set $\alpha = 2$ and $\beta = 4$ following the teacher network so as to prevent the results from being affected by the hyperparameters. The loss weights ω_1 and ω_2 of *Teacher* and *ST-plain* follow the same as CenterNet. As for *ST-KD*, the loss weights are reassigned because an extra hint loss is involved. The weighting factor γ is not listed here since it varies across different experiments.

Hardware setups. All the training and inference procedures are processed by a personal computer with a single Geforce 2080Ti GPU. The codes are written in the Python language with the PyTorch framework. Under these setups, the inference speeds of the teacher and student networks are reported in Table 2. It is observed that the light-weight student network runs about 2.5 times faster.

Table 1. Parameter settings in the expriments. *n/a* refers to *not applied*.

	α	β	ω_1	ω_2	ω_3
Teacher	2	4	0.1	1	n/a
ST-plain	2	4	0.1	1	n/a
ST-KD	2	4	2	10	1

Table 2. Comparison of structures and inference speeds between teacher and student networks.
The running speeds vary slightly across different runs. Thus, each network is tested for three times, then the mean inference speed is reported.

	Backbone	Head channels	Inference speed (Hz)
Teacher	DLA-34	256	38.2
Student	ResNet-18	64	94.5

Controlled experiments

Effectiveness of the presented approach. A performance comparison between the teacher network, the student network trained in a plain way, and the student network trained with the presented approach is reported in Table 3. For *ST-KD*, the parameters are set as $\gamma = 0.8$, $T = 10$, and the dynamic training scheme is employed. It is observed by applying the presented approach, the student network's performance is prominently enhanced for all object classes and all difficulty levels.

Influence of temperature T. To study the influence of temperature T, controlled experiments are conducted and the resultant comparison is reported in Table 4. It is observed the temperature should be neither too high or too low. $T = 10$ is a decent trade-off.

Influence of weighting factor γ. The weighting factor γ is crucial for balancing the contribution of the softened labels and ground truth labels. We conduct controlled experiments to choose a proper value for it. The results in Table 5 indicate that $\gamma = 0.8$ yields the best performance.

Influence of hint-based distillation. An ablation study on the influence of hint-based distillation is conducted. As is reported in Table 6, without the hint guidance, the student network suffers from a performance drop, but still outperforms the plain version by a large margin.

Selection of hint layer. As is previously demonstrated, the last layer of the teacher's backbone is selected as the hint layer to deliver knowledge to the student network. To provide grounds for this design, we do comparative experiments with the following two configurations:

Table 3. Comparison between with / without the presented approach.
Experiments are conducted on the KITTI dataset. *E*, *M* and *H* are short for *Easy*, *Moderate* and *Hard*, respectively. The reported performance is measured by AP scores in percentage (also for the tables hereafter).

	Car			Pedestrian			Cyclist		
	E	M	H	E	M	H	E	M	H
Teacher	97.75	89.33	79.81	71.28	61.47	51.96	76.04	49.05	45.22
ST-plain	89.65	70.61	61.30	63.62	52.48	45.26	52.10	33.39	31.77
ST-KD	94.53	84.79	74.83	71.04	60.45	50.61	60.88	38.04	36.25

Table 4. Student network's performance under different T.

	Car			Pedestrian			Cyclist		
	E	M	H	E	M	H	E	M	H
$T = 5$	91.32	80.07	72.08	69.20	58.31	48.96	57.97	36.79	35.24
$T = 10$	94.53	84.79	74.83	71.04	60.45	50.61	60.88	38.04	36.25
$T = 15$	93.87	84.17	73.62	70.53	59.34	49.36	59.98	37.67	35.47
$T = 20$	93.03	81.10	72.39	68.65	57.26	48.73	56.56	33.22	32.19

Table 5. Student network's performance under different γ.

	Car			Pedestrian			Cyclist		
	E	M	H	E	M	H	E	M	H
$\gamma = 0.5$	91.30	83.21	72.11	68.87	58.06	47.12	56.20	34.68	32.41
$\gamma = 0.8$	94.53	84.79	74.83	71.04	60.45	50.61	60.88	38.04	36.25
$\gamma = 0.9$	92.64	83.82	73.87	70.42	59.57	48.78	60.54	37.71	36.05
$\gamma = 0.95$	90.86	79.68	69.95	69.60	58.40	48.09	59.46	35.35	34.96

Table 6. Ablation study on the hint-based distillation. *ST-KD_w/o_hint* refers to *ST-KD without hint-based distillation*.

	Car			Pedestrian			Cyclist		
	E	M	H	E	M	H	E	M	H
ST-plain	89.65	70.61	61.30	63.62	52.48	45.26	52.10	33.39	31.76
ST-KD_w/o_hint	91.13	79.63	71.96	69.40	58.71	48.52	58.59	36.45	34.99
ST-KD	94.53	84.79	74.83	71.04	60.45	50.61	60.88	38.04	36.25

Table 7. Comparison between different choices of hint layers.

	Car			Pedestrian			Cyclist		
	E	M	H	E	M	H	E	M	H
ST-KD-conv3	94.17	83.27	73.53	69.01	57.89	48.86	59.30	36.52	35.24
ST-KD-conv5	93.40	82.46	72.65	68.74	57.44	48.73	57.96	35.02	34.28
ST-KD	94.53	84.79	74.83	71.04	60.45	50.61	60.88	38.04	36.25

Table 8. Comparison between joint optimization and stage-wise optimization. The results of *Moderate* level in each object category are reported.

Optimization strategy	Car	Pedestrian	Cyclist
Stage-wise optimization	83.07	58.22	34.39
Joint optimization	84.79 (+ 1.72)	60.45 (+ 2.23)	38.04 (+ 3.65)

Table 9. Comparison between dynamic training scheme and plain training scheme. The results of *Moderate* level in each object category are reported.

	Car	Pedestrian	Cyclist
Plain training	84.49	60.33	37.27
Dynamic training	84.79 (+ 0.30)	60.45 (+ 0.12)	38.04 (+ 0.77)

(i) *ST-KD-conv3*: Hint-based distillation is implemented between the third level's output of DLA-34 and the third convolutional block's output of ResNet-18. Both features have the resolution of 48×160.
(ii) *ST-KD-conv5*: Hint-based distillation is implemented between the fifth level's output of DLA-34 and the fifth convolutional block's output of ResNet-18. Both features have the resolution of 12×40.

The comparison is displayed in Table 7. It is observed that the adopted configuration (*ST-KD*) outperforms the other two alternatives.

Effectiveness of joint optimization. The effectiveness of the joint optimization strategy is validated by a controlled experiment and the resultant comparison is exhibited in Table 8. It is observed the joint optimization strategy is superior to the stage-wise one in our framework.

Effectiveness of dynamic training scheme. A controlled experiment is conducted to compare the proposed dynamic training scheme and the plain training scheme. For the plain training scheme, γ is set as 0.8 for all 70 training epochs. For the dynamic training scheme, γ is fixed as 0.8 for the first 65 epochs, and is then linearly increased to 1.0 in the last 10 epochs. As is reported in Table 9, the student network trained by dynamic training scheme slightly outperforms its counterpart.

Visualization

To give an intuitive exhibition of the resulting student network, some of the detection results are visualized in Figure 4. Two observations are drawn as below:

(i) The resulting light-weight student network is capable of producing accurate detections in most cases, presenting decent robustness under common detection challenges including multiple scales, occlusion, truncation, illumination change, and so forth.
(ii) The main limitation to the resulting network is the phenomenon of overlapped boxes, that is, excessive bounding boxes sometimes co-exist for the same object. This phenomenon is shown in the last row of Figure 4.

Conclusion

In this work, the traditional knowledge distillation has been revisited for light-weight visual object detection. The

Figure 4. Visualization of detection results produced by the resulting light-weight student network.
Due to the space limit, *pedestrian* and *cyclist* are abbreviated as *ped* and *cyc*, respectively, in the captions of bounding boxes. The confidence scores produced by the network are appended following the category names. Due to the space limit, the top-left corners of bounding boxes are prioritized for placing their captions, and the bottom-left corners are used if there is otherwise not enough space.

formulation that converts traditional knowledge distillation into the linear combination of ground truth labels and soft labels has been developed. Based on the formulation, two novel strategies have been further proposed so that the knowledge distillation can be naturally adapted to focal loss. Hint-based distillation has also been employed to promote mutual consistency between teacher's and student's backbones. Through extensive experiments, effectiveness of the presented approach has been corroborated. The main limitation to be tackled is the phenomenon of overlapped boxes, which will be investigated by the future research.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Yunfeng Gao https://orcid.org/0000-0003-3308-3255

References

Carion N, Massa F, Synnaeve G, et al. (2020) End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872* .

Chen G, Choi W, Yu X, et al. (2017) Learning efficient object detection models with knowledge distillation. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (eds I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett), Long Beach, CA, USA, 4–9 December 2017, pp. 742–751. Curran Associates, Inc.

Chen X, Kundu K, Zhu Y, et al. (2015) 3d object proposals for accurate object class detection. In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)* (eds C Cortes, N Lawrence, D Lee, M Sugiyama and R Garnett), Montreal, Canada, 7–12 December 2015, pp. 424–432. Curran Associates, Inc.

Geiger A, Lenz P and Urtasun R (2012) Are we ready for autonomous driving? the kitti vision benchmark suite. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, Providence, RI, USA, 16–21 June 2012, pp. 3354–3361, IEEE.

Girshick R, Donahue J, Darrell T and Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, Columbus, OH, USA, 23–28 June 2014, pp. 580–587, IEEE.

He K, Zhang X, Ren S and Sun J (2016) Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778, IEEE.

Heo B, Lee M, Yun S and Choi JY (2019) Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019)*, Honolulu, HI, USA, 27 January 2019–01 February 2019, pp. 3779–3787, AAAI.

Hinton G, Vinyals O and Dean J (2015) Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* .

Lassance C, Bontonou M, Hacene GB, et al. (2020) Deep geometric knowledge distillation with graphs. In: *2020 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain, 4–8 May 2020, pp. 8484–8488, IEEE.

Law H and Deng J (2018) Cornernet: Detecting objects as paired keypoints. In: *European Conference on Computer Vision (ECCV 2018)* (eds V Ferrari, M Hebert, C Sminchisescu and Y Weiss), Munich, Germany, 8–14 September 2018, pp. 734–750, Cham: Springer International Publishing.

Lin TY, Dollár P, Girshick R, et al. (2017a) Feature pyramid networks for object detection. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, Honolulu, HI, USA, 21–26 July 2017, pp. 2117–2125, IEEE.

Lin TY, Goyal P, Girshick R, et al. (2017b) Focal loss for dense object detection. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, Honolulu, HI, USA, 21–26 July 2017, pp. 2980–2988, IEEE.

Liu W, Anguelov D, Erhan D, et al. (2016) SSD: Single shot multibox detector. In: *European Conference on Computer Vision (ECCV 2016)* (eds B Leibe, J Matas, N Sebe and M Welling), Amsterdam, the Netherlands, 11–14 October 2016, pp. 21–37, Cham: Springer International Publishing.

Pan H and Sun W (2018) Nonlinear output feedback finite-time control for vehicle active suspension systems. *IEEE Transactions on Industrial Informatics* 15(4): 2073–2082.

Pan H, Jing X, Sun W and Gao H (2017) A bioinspired dynamics-based adaptive tracking control for nonlinear suspension systems. *IEEE Transactions on Control Systems Technology* 26(3): 903–914.

Pan H, Jing X, Sun W and Li Z (2018) Analysis and design of a bioinspired vibration sensor system in noisy environment. *IEEE/ASME Transactions on Mechatronics* 23(2): 845–855.

Park W, Kim D, Lu Y and Cho M (2019) Relational knowledge distillation. In: *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, CA, USA, 15–20 June 2019, pp. 3967–3976, IEEE.

Redmon J and Farhadi A (2018) Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Redmon J, Divvala S, Girshick R and Farhadi A (2016) You only look once: Unified, real-time object detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, 27–30 June 2016, pp. 779–788, IEEE.

Ren S, He K, Girshick R and Sun J (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)* (eds C Cortes, N Lawrence, D Lee, M Sugiyama and R Garnett), Montreal, Canada, 7–12 December 2015, pp. 424–432. Curran Associates, Inc.

Romero A, Ballas N, Kahou SE, et al. (2014) Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Sahu BK and Subudhi B (2017) Potential function-based path-following control of an autonomous underwater vehicle in an obstacle-rich environment. *Transactions of the Institute of Measurement and Control* 39(8): 1236–1252.

Simonelli A, Bulo SR, Porzi L, et al. (2019) Disentangling monocular 3d object detection. In: *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, CA, USA, 15–20 June 2019, pp. 1991–1999, IEEE.

Tian Z, Shen C, Chen H and He T (2019) Fcos: Fully convolutional one-stage object detection. In: *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, CA, USA, 15–20 June 2019, pp. 9627–9636, IEEE.

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (eds I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett), Long Beach, CA, USA, 4–9 December 2017, pp. 5998–6008. Curran Associates, Inc.

Wang T, Yuan L, Zhang X and Feng J (2019) Distilling object detectors with fine-grained feature imitation. In: *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, CA, USA, 15–20 June 2019, pp. 4933–4942, IEEE.

Yim J, Joo D, Bae J and Kim J (2017) A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, Honolulu, HI, USA, 21–26 July 2017, pp. 4133–4141, IEEE.

Yu F, Wang D, Shelhamer E and Darrell T (2018) Deep layer aggregation. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, Salt Lake City, UT, USA, 18–23 June 2018, pp. 2403–2412, IEEE.

Zhang X, Zhang L, Pei H and Lewis FL (2020) Part-based multi-task deep network for autonomous indoor drone navigation. *Transactions of the Institute of Measurement and Control* 42(16): 3243–3253.

Zhou X, Wang D and Krähenbühl P (2019) Objects as points. *arXiv preprint arXiv:1904.07850*.