Application of an Improved Focal Loss in Vehicle Detection*

Xuanlin He[1], Jie Yang[1(✉)], and Nikola Kasabov[2]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
jieyang@sjtu.edu.cn
[2] Auckland University of Technology, New Zealand

Abstract. Object detection is an important and fundamental task in computer vision. Recently, the emergence of deep neural network has made considerable progress in object detection. Deep neural network object detectors can be grouped in two broad categories: the two-stage detector and the one-stage detector. One-stage detectors are faster than two-stage detectors. However, they suffer from a severe foreground-background class imbalance during training that causes a low accuracy performance. RetinaNet is a one-stage detector with a novel loss function named Focal Loss which can reduce the class imbalance effect. Thereby RetinaNet outperforms all the two-stage and one-stage detectors in term of accuracy. The main idea of focal loss is to add a modulating factor to rectify the cross-entropy loss, which down-weights the loss of easy examples during training and thus focuses on the hard examples. However, cross-entropy loss only focuses on the loss of the ground-truth classes and thus it can't gain the loss feedback from the false classes. Thereby cross-entropy loss does not achieve the best convergence. In this paper, we proposed a new loss function named Dual Cross-Entropy Focal Loss, which improves on the focal loss. Dual cross-entropy focal loss adds a modulating factor to rectify the dual cross-entropy loss towards focusing on the hard samples. Dual cross-entropy loss is an improved variant of cross-entropy loss, which gains the loss feedback from both the ground-truth classes and the false classes. We changed the loss function of RetinaNet from focal loss to our dual cross-entropy focal loss and performed some experiments on a small vehicle dataset. The experimental results show that our new loss function improves the vehicle detection performance.

Keywords: Focal Loss · Class Imbalance · Cross-Entropy Loss · RetinaNet · Vehicle Detection · Object Detection · Deep Neural Network.

1 Introduction

Object detection is one of the most fundamental tasks in computer vision, which has received considerable attention for several decades. The emergence of deep

* Corresponding author: Jie Yang (jieyang@sjtu.edu.cn);This research was partly supported by NSFC, China (No:61876107,U1803261).

convolutional neural networks, including CNNs [1–3], has provided a significant improvement in object detection [4–6]. The CNN-based object detection methods are mainly divided into two categories: the two-stage method and the one-stage method.

The R-CNN-like two-stage detectors generate the object candidate regions in the first stage and then classify each candidate region as one of the foreground classes or as background in the second stage. Generation of the object candidate regions in the first stage greatly improves the detection accuracy; however it reduces the detection speed. The representatives of the two-stage method are the region proposal based detectors, such as RCNN [7], Fast RCNN [8], Faster RCNN [9] and RFCN [10].

The one-stage detectors skip the process of generating the oject candidate regions. In order to cover the space of possible image boxes, the one-stage detectors use a dense set of fixed sampling grids, such as multiple 'anchors' [9], at each spatial position, and thus they must process a much larger set of regions sampled across an image. As compared to the two-stage detectors, the one-stage detectors improve the detection speed but reduce the detection accuracy. The representatives of the one-stage method are YOLO [11], YOLO9000 [12], YOLOv3 [13], SSD [14] and DSSD [15].

In the two-stage method, the positive and negative samples are relatively balanced (e.g., 1:3). Because in the first stage, a large set (e.g., 1-2k) of object candidate regions are selected and most of the background regions (the negative samples) are discarded. In the one-stage method, the positive and negative samples are extremely unbalanced (e.g., 1:1000). Because a dense sampling of regions (e.g., 100k) which cover various locations, scales, and aspect ratios need to be classified, and the majority of the regions are background regions (the negative samples). Each sampled region can be treated as an training sample. In the one-stage detector, when the convolutional neural network trains the large set of sampled regions, the majority of the loss function consists of the easily classified negatives (background exam-ples) and they dominate the gradient. Thus, extreme foreground-background class imbalance during training is one of the main reasons that causes the two-stage detectors perform more accurate than one-stage detectors.

RetinaNet[16] is a one-stage detector that has a superior performance for dense sampling of object locations in an input image. The network structure of RetinaNet draws on a variety of recent ideas, such as the concept of anchor in RPN [9], the feature pyramids in SSD [14] and FPN [17]. However, Tsung-Yi Lin et al. [16] emphasized, "We emphasize that our simple detector achieves top results not based on innovations in network design but due to our novel loss." Tsung-Yi Lin et al. [16] proposed a new loss function named Focal Loss to address the extreme class imbalance. They greatly reduced the weight of easy negatives in the loss function by adding a modulating factor to the standard cross-entropy loss. As a one-stage detector, "RetinaNet is able to match the speed of previous one-stage detectors while surpas-sing the accuracy of all existing state-of-the-art two-stage detectors"[16].

Focal loss is based on the cross-entropy loss. However, cross-entropy loss only focuses on the loss of the ground-truth classes and thus it can't gain the loss feedback from the false classes. Xiaoxu Li et al. [20] proposed an improved variant of cross-entropy loss named Dual Cross-Entropy Loss to gain the loss feedback from both the ground-truth classes and the false classes. In this paper, we combined the idea of focal loss[16] to focus more on hard examples with dual cross-entropy loss and proposed a new loss function named Dual Cross-Entropy Focal Loss. We substituted the loss of RetinaNet[16] with our proposed dual cross-entropy focal loss and applied it to a small vehicle dataset. The experimental results show that our new loss function improves the vehicle detection performance.

2 Related Work

In this section, we will introduce cross-entropy loss and focal loss first. We will analyse why focal loss can reduce the class imbalance effect. Then we will point out the shortage of the cross-entropy loss and introduce the dual cross-entropy loss. We will analyse the advantages of the dual cross-entropy loss compared with the cross-entropy loss. Finally, we will integrate the dual cross-entropy loss and focal loss to create a new loss function named Dual Cross-Entropy Focal Loss.

2.1 Cross-Entropy Loss and Focal Loss

Suppose that $D = \{(x_1, \mathbf{Y}_1), ..., (x_k, \mathbf{Y}_k), ..., (x_M, \mathbf{Y}_M)\}$ is a training dataset of M samples. We assume that all the samples have C categories: background and $C-1$ types of objects. $\mathbf{Y}_k$ is the ground-truth label of the $kth\,(k \in \{1, 2, ..., M\})$ sample x_k and $\mathbf{Y}_k$ is a C-dimensional one-hot vector. Only one component in $\mathbf{Y}_k$ is 1, and the other components are equal to zero. ${y_k}^{(i)}$ denotes the $ith\,(i \in \{1, 2, ..., C\})$ component of the vector $\mathbf{Y}_k$, then ${y_k}^{(i)}$ is defined as follows:

$$y_k^{(i)} = \begin{cases} 1 & \text{if } x_k \text{ belongs to the } ith\,(i \in \{1, 2, ..., C\}) \text{ class} \\ 0 & \text{if } x_k \text{ does not belong to the } ith\,(i \in \{1, 2, ..., C\}) \text{ class} \end{cases} \tag{1}$$

$\mathbf{P}_k$ is the probability distribution of the $kth\,(k \in \{1, 2, ..., M\})$ sample x_k predicted by the detector and $\mathbf{P}_k$ is also a C-dimensional vector. $p_k^{(i)}$ denotes the $ith\,(i \in \{1, 2, ..., C\})$ component of the vector $\mathbf{P}_k$. $p_k^{(i)}$ is the probability that the detector predicts the sample x_k belonging to the $ith\,(i \in \{1, 2, ..., C\})$ class.

Because of the softmax function, we have $\forall k\,(k \in \{1, 2, ..., M\}),\ \sum_{i=1}^{C} p_k^{(i)} = 1$.

For the sake of brevity, we use $t_k\,(t_k \in \{1, 2, ..., C\})$ to represent the ground-truth class of the $kth\,(k \in \{1, 2, ..., M\})$ sample x_k, and then the cross-entropy loss of the sample x_k is defined as follows:

$$CE\,(x_k, \mathbf{Y}_k) = -\mathbf{Y}_k^T \cdot \log\,(\mathbf{P}_k) = -\log\left(p_k^{(t_k)}\right) \tag{2}$$

The total cross-entropy loss of M samples is defined as follows:

$$L_{CE} = \sum_{k=1}^{M} CE\left(x_k, \mathbf{Y}_k\right) = -\sum_{k=1}^{M} \left(\mathbf{Y}_k^T \cdot \log\left(\mathbf{P}_k\right)\right) = -\sum_{k=1}^{M} \log\left({p_k}^{(t_k)}\right) \tag{3}$$

According to equation (3), the cross-entropy loss of each training sample is accumulated by an equal weight. It means that, the easy examples ($p_k^{(t_k)} \gg 0.5$) and the hard examples have the same weight. Even easy examples have a loss with non-trivial magnitude. Most of the training examples of the one-stage detectors are easy negatives. Tsung-Yi Lin [16] said, "When summed over a large number of easy examples, these small loss values can overwhelm the rare class." Therefore, easy negatives generally lead to degenerate models.

Focal loss [16] multiplies the cross-entropy loss of the sample x_k by a modulating factor $\left(1-p_k^{(t_k)}\right)^{\gamma}$. $\gamma \geqslant 0$ is a constant variable that is suggested to be γ=2 in [16]. The focal loss of the sample x_k is defined as follows:

$$\begin{aligned} FL\left(x_k, \mathbf{Y}_k\right) =& \left(1-p_k^{(t_k)}\right)^{\gamma} \cdot CE\left(x_k, \mathbf{Y}_k\right) \\ =& -\left(1-p_k^{(t_k)}\right)^{\gamma} \cdot \mathbf{Y}_k^T \cdot \log\left(\mathbf{P}_k\right) \\ =& -\left(1-p_k^{(t_k)}\right)^{\gamma} \cdot \log\left(p_k^{(t_k)}\right) \end{aligned} \tag{4}$$

The total focal loss of M samples is defined as follows:

$$\begin{aligned} L_{FL} =& \sum_{k=1}^{M} FL\left(x_k, \mathbf{Y}_k\right) \\ =& -\sum_{k=1}^{M} \left(\left(1 - p_k^{(t_k)}\right)^{\gamma} \cdot \mathbf{Y}_k^T \cdot \log\left(\mathbf{P}_k\right)\right) \\ =& -\sum_{k=1}^{M} \left(1 - p_k^{(t_k)}\right)^{\gamma} \cdot \log\left({p_k}^{(t_k)}\right) \end{aligned} \tag{5}$$

$\gamma \geqslant 0$. When $\gamma = 0$, the focal loss is equivalent to the cross-entropy loss. When an example is easy to classify and ${p_k}^{(t_k)}$ is near 1, the modulating factor $\left(1 - {p_k}^{(t_k)}\right)^{\gamma}$ is near 0 and the loss is down-weighted. When an example is hard to classify and ${p_k}^{(t_k)}$ is small, the modulating factor $\left(1 - {p_k}^{(t_k)}\right)^{\gamma}$ is near 1 and the loss is unaffected. Indeed, focal loss significantly down-weights the loss of the easy examples, and thus focuses on the hard examples. Therefore, focal loss can reduce the class imbalance effect.

2.2 Dual Cross-Entropy Loss

More recently, dual cross-entropy loss [20] is proposed to apply for the vehicle image classification, in which the accuracy of the model improves. .

During training, the cross-entropy loss (equation (2)) only focuses on increasing the probability that a sample is classified to its corresponding ground-truth class. Although due to the effect of the softmax function, the probability that a sample is classified to a class other than its ground-truth class correspondingly reduces. The cross-entropy loss does not achieve the best convergence because it can't gain the loss feedback from the false classes.

The dual cross-entropy loss not only increases the probability that a sample is correctly classified but also decreases the probability that a sample is classified to a class other than its ground-truth class. The dual cross-entropy loss of the sample x_k is defined as follows:

$$\begin{aligned} DCE\left(x_k, \mathbf{Y}_k\right) =& CE\left(x_k, \mathbf{Y}_k\right) + \beta \cdot Reg\left(x_k, \mathbf{Y}_k\right) \\ =& -\mathbf{Y}_k^T \cdot \log\left(\mathbf{P}_k\right) + \beta \cdot \left(1 - \mathbf{Y}_k^T\right) \cdot \log\left(\alpha + \mathbf{P}_k\right) \\ =& -\log\left(p_k^{(t_k)}\right) + \beta \cdot \sum_{\substack{i=1 \\ i \neq t_k}}^{C} \log\left(\alpha + p_k^{(i)}\right) \end{aligned} \tag{6}$$

The total dual cross-entropy loss loss of M samples is defined as follows:

$$\begin{aligned} L_{DCE} =& L_{CE} + \beta \cdot L_R \\ =& \sum_{k=1}^{M} CE\left(x_k, \mathbf{Y}_k\right) + \beta \cdot \sum_{k=1}^{M} Reg\left(x_k, \mathbf{Y}_k\right) \\ =& -\sum_{k=1}^{M} \left(\mathbf{Y}_k^T \cdot \log\left(\mathbf{P}_k\right)\right) + \beta \cdot \sum_{k=1}^{M} \left(\left(1 - \mathbf{Y}_k^T\right) \cdot \log\left(\alpha + \mathbf{P}_k\right)\right) \\ =& -\sum_{k=1}^{M} \left(\log\left(p_k^{(t_k)}\right)\right) + \beta \cdot \sum_{k=1}^{M} \sum_{\substack{i=1 \\ i \neq t_k}}^{C} \log\left(\alpha + p_k^{(i)}\right) \end{aligned} \tag{7}$$

L_{CE} is the cross-entropy loss in equation (3) and L_R is a regularization term. $\alpha > 0, \beta \geqslant 0$. When $\beta = 0$, the dual cross-entropy loss has a same value as the cross-entropy loss. We set the $\alpha = 1$ and $\beta = 10$ as suggested in [20]. While training, L_{CE} is increasing the probability that a sample is correctly classified($p_k^{(t_k)}$), and L_R is decreasing the probability that a sample is classified to another class (rather than its ground-truth).

Xiaoxu Li et al. [20] summarized the advantages of the dual cross-entropy loss compared with the cross-entropy loss as follows:

First, dual cross-entropy loss can accelerate the optimization of the neural network.

Second, dual cross-entropy loss works better on small-sample datasets and per-forms well on large-sample datasets.

Third, dual cross-entropy loss can ensure the network or model has a more stable performance compared to the cross-entropy loss.

2.3 Dual Cross-Entropy Focal Los

We take the idea of focusing more on hard examples from focal loss and add a modulating factor to the dual cross-entropy loss to down-weights the loss of the easy examples. We named the new loss Dual Cross-Entropy Focal Loss. We define the dual cross-entropy focal loss of the sample x_k as follows:

$$\begin{aligned} DCFL\left(x_k, \mathbf{Y}_k\right) = & -\left(1 \text{ - } p_k^{(t_k)}\right)^{\gamma_1} \cdot \log\left(p_k^{(t_k)}\right) \\ & +\beta \cdot \sum_{\substack{i=1 \\ i \neq t_k}}^{C} \left(\left(p_k^{(i)}\right)^{\gamma_2} \cdot \log\left(\alpha + p_k^{(i)}\right)\right) \end{aligned} \tag{8}$$

$\gamma_1 \geqslant 0, \gamma_2 \geqslant 0$. When $\gamma_1 = \gamma_2 = 0$, the dual cross-entropy focal loss is the same as the dual cross-entropy loss in equation (6). The dual cross-entropy focal loss consists of two parts. The first part is $-\left(1 \text{ - } p_k^{(t_k)}\right)^{\gamma_1} \cdot \log\left(p_k^{(t_k)}\right)$, which is the same as the focal loss in equation (4). This part increases the probability that a sample is assigned to its ground-truth class ($p_k^{(t_k)}$), and focuses on the hard examples whose $p_k^{(t_k)}$ is small. The second part is $\left(p_k^{(i)}\right)^{\gamma_2} \cdot \log\left(\alpha + p_k^{(i)}\right)$, which decreases the probability that a sample is classified to a class other than its ground-truth class. The second part also focuses on the hard examples. Because some of the probabilities that a hard example is classified to a class other than its ground-truth class are large. For example, for a hard example, if $p_k^{(i)}$ ($i \in \{1, 2, ..., C\}$, $i \neq t_k$) is large, the loss can focus on decreasing $p_k^{(i)}$. Dual cross-entropy focal loss gains the loss feedback from both the ground-truth classes and the false classes through the two parts and focuses on the hard examples.

The total dual cross-entropy focal loss of M samples is defined as follows:

$$\begin{aligned} L_{DCFL} = & \sum_{k=1}^{M} DCFL\left(x_k, \mathbf{Y}_k\right) \\ = & -\sum_{k=1}^{M} \left(\left(1 \text{ - } p_k^{(t_k)}\right)^{\gamma_1} \cdot \log\left(p_k^{(t_k)}\right)\right) \\ & +\beta \cdot \sum_{k=1}^{M} \sum_{\substack{i=1 \\ i \neq t_k}}^{C} \left(\left(p_k^{(i)}\right)^{\gamma_2} \cdot \log\left(\alpha + p_k^{(i)}\right)\right) \end{aligned} \tag{9}$$

3 Experimental Results

3.1 UA-DETRAC Dataset

We verified the dual cross-entropy focal loss on an open vehicle dataset named UA-DETRAC. UA-DETRAC is a real-world multi-object detection dataset that

consists of 10 hours of videos. The videos are captured at 24 different locations in Beijing and Tianjin, China. The resolution of each picture is 960×540 pixels. There are more than 140 thousand frames in the UA-DETRAC dataset, and 8250 vehicles have been manually annotated, with a total of 1.21 million labeled bounding boxes of objects. In Fig. 1, we have shown some examples of the dataset. Our GPU resources are limited, to save experimental time, we only selected 2000 pictures from UA-DETRAC to create our dataset. We divided the obtained dataset into three parts. The training set consists of 1200 pictures. The validation set consists of 400 pictures. The test set consists of 400 pictures.

Fig. 1. Examples in the dataset

3.2 Comparison of Focal Loss and Dual Cross-Entropy Focal Loss on Our Dataset

For the experiments and performance evaluation, we trained the same RetinaNet[16] model by minimizing the focal loss or our dual cross-entropy focal loss on the same training set for 100 epochs. In our work, we set focal loss (equation (5)) as $\gamma = 2$ as suggested in [16]. Dual cross-entropy focal loss (equation (8)) was set to $\gamma_1 = \gamma_2 = 2, \alpha = 1, \beta = 10$ as suggested in [20]. The cited reference [16] provides the structural details and description of RetinaNet. We saved the trained models after each epoch, and finally, we tested each saved model on the same test set. The mAP50 index ($IoU = 0.5$) on the same test set for the saved models trained by minimizing two loss functions after each epoch is shown in Fig. 2.

We trained the same RetinaNet model by minimizing the focal loss or our dual cross-entropy focal loss on the same training set 10 times respectively. In each training epoch, the performance of the model was monitored on the validation dataset, and the model that had the best performance was saved. Finally, we tested all the saved models that trained by minimizing two loss functions on the same test set and calculated the mean and standard deviation of the 10 mAP50

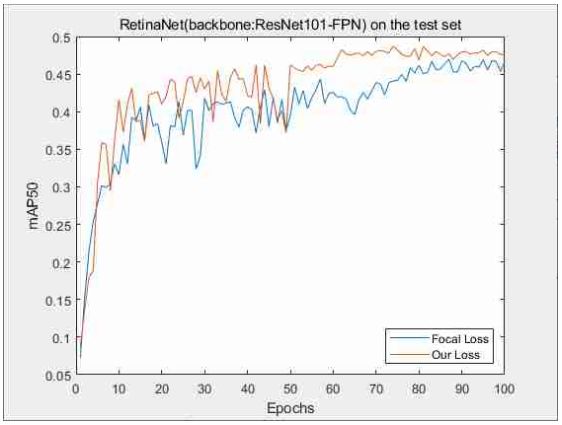

Fig. 2. Curves of the mAP50 index obtained by the RetinaNet network trained by minimizing the focal loss and our dual cross-entropy focal loss on the same dataset

indexes. As shown in Table 1, dual cross-entropy focal loss improves the mAP50 index by 1.6% and has a smaller standard deviation.

Table 1. The means and standard deviations of the 10 mAP50 indexes obtained by the RetinaNet network trained by minimizing the focal loss and our loss on the same dataset

Loss Function	Mean	Std.
Focal Loss	0.471	0.018
Dual Cross-Entropy Focal Loss	0.487	0.013

3.3 Comparison to Traditional Object Detecters on Our Dataset

We evaluated three different one-stage detectors on the bounding box detection task on our dataset. SSD[14] and RetinaNet[16] used depth 101 ResNet[4] as their backbone network. YOLOv3[13] used the 53 depth Darknet [13] as its backbone network. J. Redmon et al. [13] said that Darknet-53 has equal accuracy to ResNet-101. The ResNet-101 and Darknet-53 were pre-trained on the ImageNet dataset. In order to match the default input image size of the detectors, we resized the image. The input images for SSD were resized into 512×512 dimension. The input images for YOLOv3 were resized into 608×608 dimension. The shorter side of the input images for RetinaNet was resized into 608 dimension. For the data augmentation, horizontal flipping was only used. The corresponding cited references provide the structural details and description of the three detectors.

We trained SSD and YOLOv3 on the same training set for 100 epochs. For SSD and YOLOv3, we monitored the performance of its network on the same validation set. For SSD and YOLOv3, the model which had the best performance was saved and then used to predict the test data. We used the performance of RetinaNet in Table 1 and we get Table 2.

The UA-DETRAC dataset mainly consists of medium and small objects. SSD[14] performs poor on the small objects. This is mainly because the small objects may not have even information at the very top layers. YOLOv3[13] performs much better than SSD, especially on the medium and small objects. As shown in Table 2, RetinaNet trained by minimizing the focal loss achieves a 1.5 point AP gap (47.1 vs. 45.6) with YOLOv3 and achieves a 9.8 point AP gap (47.1 vs. 37.3) with SSD. RetinaNet trained by minimizing our dual cross-entropy focal loss achieves a 1.6 point AP gap (48.7 vs. 47.1) further.

Table 2. Comparison of traditional object detection methods on our dataset

methods	backbone	mAP50
SSD[14]	ResNet101-SSD	0.373
YOLOv3[13]	Darknet-53	0.456
RetinaNet[16] + Focal Loss	ResNet101-FPN	0.471
RetinaNet[16] + Dual Cross-Entropy Focal Loss	ResNet101-FPN	0.487

4 Conclusions

In this paper, we integrated the dual cross-entropy loss and focal loss to create a new loss function named Dual Cross-Entropy Focal Loss. As compared to the focal loss, our proposed loss considers loss on both both the ground-truth classes and the false classes by adding a regularization term which places a constraint on the probability that the example belongs to a false class. Fig. 2 shows that our new loss can accelerate the convergence of the network and improve the detection accuracy. Table 2 shows that RetinaNet[16] trained by minimizing our new loss achieves best accuracy on our dataset compared to the baselines.

References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton.: ImageNet classification with deep con-volutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097-1105 (2012)
2. K. Simonyan and A. Zisserman.: Very deep convolutional networks for large-scale im-age recognition. arXiv preprint arXiv:1409.1556 (2014)
3. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems, pp. 487-495 (2014)

4. K. He, X. Zhang, S. Ren, and J. Sun.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778 (2016)
5. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel.: Backpropagation applied to handwritten zip code recognition. Neural Computation(4), 541-551 (1989)
6. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9 (2015)
7. R. B. Girshick, J. Donahue, T. Darrell, and J. Malik.: Rich feature hierarchies for accu-rate object detection and semantic segmentation. In: Proceedings of the IEEE Confer-ence on Computer Vision and Pattern Recognition, pp. 580-587 (2014)
8. R. B. Girshick.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440-1448 (2015)
9. S. Ren, K. He, R. Girshick, and J. Sun.: Faster R-CNN: Towards real-time object detec-tion with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91-99 (2015)
10. J. Dai, Y. Li, K. He, and J. Sun.: R-FCN: Object detection via region-based fully convo-lutional networks. In: Advances in Neural Information Processing Systems, pp. 379-387 (2016)
11. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788 (2016)
12. J. Redmon and A. Farhadi.: YOLO9000: Better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263-7271 (2017)
13. J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement. arXiv preprint arXiv: 1804.02767 (2018)
14. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg.: SSD: Single shot multibox detector. In: Proceedings of the European Conference on Com-puter Vision 2016, LNCS, vol. 9905, pp. 21-37. Springer, Heidelberg (2016)
15. C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg.: DSSD: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2016)
16. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar.: Focal Loss for Dense Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980-2988 (2017)
17. T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie.: Feature pyra-mid networks for object detection. In: Proceedings of the IEEE Conference on Comput-er Vision and Pattern Recognition, pp.2117-2125 (2017)
18. J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al.: Speed/accuracy trade-offs for modern convolutional ob-ject detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7310-7321 (2017)
19. A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta.: Beyond skip connections: Top-down modulation for object detection. arXiv preprint arXiv:1612.06851 (2016)
20. X. Li, L. Yu, D. Chang, Z. Ma, and J. Cao.: Dual Cross-Entropy Loss for Small-Sample Fine-Grained Vehicle Classification. IEEE Transactions on Vehicular Technology 68(5), 4204-4212 (2019)