



MDCN: Multi-scale Dilated Convolutional Enhanced Residual Network for Traffic Sign Detection

Yan Ke¹, Wanghao Mo¹, Zhe Li², Ruyi Cao¹, and Wendong Zhang¹(✉)

¹ XinJiang University, Urumqi, China

{keyan, whm734097706, 107552101634}@stu.xju.edu.cn, zwd@xju.edu.cn

² The Hong Kong Polytechnic University, Hong Kong SAR, China
lizhe.li@connect.polyu.hk

Abstract. Detecting small, multi-scale, and easily obscured traffic signs in real-world scenarios presents a persistent challenge. This paper proposes an approach that utilizes a multi-scale feature pyramid module to capture hierarchical features, facilitating robust detection of traffic signs across varying viewing angles and scales. To aggregate features at different scales and eliminate background interference, we employ a superposition of null convolution kernels with varying dilation rates, expanding the perceptual field from small to large. This effectively covers the object distribution across multiple scales while enhancing the resolution of the final output feature map for improved small target localization. Our method has demonstrated its effectiveness and superiority over several state-of-the-art approaches through extensive experiments conducted on two public traffic sign detection datasets.

Keywords: Traffic sign detection · Multi-scale feature pyramid · Hierarchical features · Perceptual field expansion · Robust detection

1 Introduction

The rapid advancement of autonomous driving and intelligent driver assistance systems has spurred extensive research on traffic sign detection [4]. However, this task is beset by various challenges, including interference from external factors such as illumination, occlusion, weather conditions, and shooting angles. Furthermore, traffic sign targets are typically small and exhibit variations in scale, exacerbating the difficulty of detection.

To overcome these challenges, researchers have proposed innovative approaches. Wang *et al.* [17] have replaced the original feature pyramid network in YOLOv5, resulting in improved real-time detection performance. Similarly, Yao *et al.* [19] have enhanced the feature fusion method of YOLOv4-Tiny through the introduction of an AFPN (Adaptive Feature Pyramid Network).

Y. Ke and W. Mo—Contribute equally to this work.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

X. Yang et al. (Eds.): ADMA 2023, LNAI 14176, pp. 584–597, 2023.

https://doi.org/10.1007/978-3-031-46661-8_39

While single-stage networks are commonly employed in traffic sign detection studies due to computational limitations in real-world applications, this approach is not without shortcomings. Environmental changes and occlusions often impact the visibility of traffic signs, leading to a decline in detection performance.

In recent times, researchers have made significant advancements in improving the performance of ATDR (Automatic Traffic Sign Detection and Recognition) in real-world scenarios. One such approach involves the utilization of multiscale pre-trained networks, which have shown promising results. In this context, a novel traffic sign detection network called TSingNet has been introduced. TSingNet leverages scale-aware and context-rich features to effectively detect and identify small or obscured traffic signs [11]. Furthermore, Shen *et al.* [16] propose a population multiscale attention pyramid network that facilitates optimal feature fusion patterns and the construction of information-rich feature pyramids to detect traffic signs of various sizes. Although these approaches have demonstrated strong performance, they primarily incorporate high-level semantic information in the earlier layers. Consequently, they face the challenge of foreground semantics, particularly regarding small traffic signs, which are prone to vanishing at higher levels of the FPN (Feature Pyramid Network).

We introduce MDCN (Multi-Scale Dilated Convolutional Enhanced Residual Network), a novel traffic sign detection network that leverages scale awareness and context-rich feature representation to detect multi-scale and small-object traffic signs efficiently. Our paper makes the following key contributions:

- We employ several novel data augmentation methods to increase the diversity and difficulty of the data, thereby enhancing the generalization ability of the model and effectively addressing the detection of various scales.
- We propose MDRNet (Multi-scale Deep Residual network), a new backbone architecture designed to learn scale-aware and context-rich features for traffic sign detection in outdoor environments. MDRNet aims to narrow the semantic gap between multiple scales, leading to improved detection performance.
- To overcome hardware limitations and accommodate high-resolution images, we introduce GN to remove the batch size limitation. Additionally, we utilize WS to further normalize the data from a weight perspective, accelerating model convergence and improving accuracy.

2 Methodology

We developed MDCN based on the Faster R-CNN framework [15]. The architecture of our model is illustrated in Fig. 1, which comprises three primary components: feature extraction, feature fusion, and detection.

2.1 Multiple Data Augmentation Fusion

In our experiments, we utilized the CTSD and GTSDDB datasets, which exhibit an imbalanced distribution of target scales and contain small target samples.

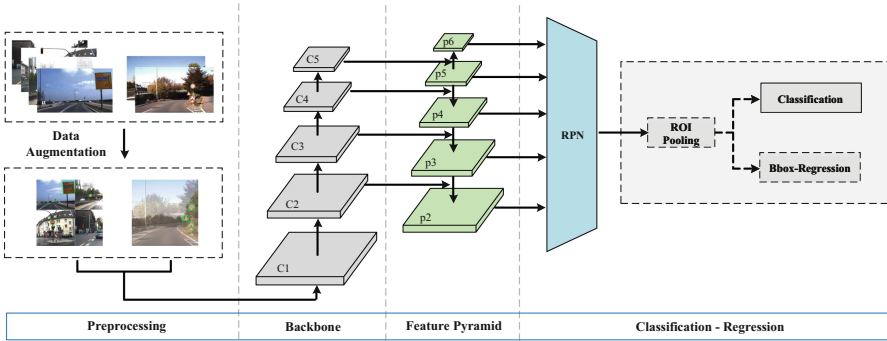


Fig. 1. The Overall architecture of MDCN. The feature maps C1 to C5, obtained from the backbone network, serve as the input for feature extraction. The feature maps P2 to P6 are then generated to extract features at different scales. The RPN (Region Proposal Network) is responsible for extracting regions of interest.

To address this issue and enhance the model’s attention towards small targets during training, we expanded and augmented the datasets with additional data specifically focused on small targets.

To achieve this, we employed various data augmentation techniques, including Mixup [22], Mosaic [1], and Random affine [14]. These methods allowed us to synthesize samples and expand the dataset effectively. Fig. 2 provides a visualization of the data augmentation techniques employed.

Mixup involves overlaying two images, resulting in improved image detection accuracy without incurring significant computational overhead. Mosaic, on the other hand, stitches together four images by randomly cropping, scaling, and aligning them. This approach not only enhances the richness of the image background but also increases the diversity of target scales within a single image. These augmentation methods significantly enrich the detection dataset and contribute to the network’s robustness.

2.2 MDRNet

To address the challenges posed by small-scale traffic sign image targets, multiple scale levels, and potential occlusion, we propose MDRNet. Our approach involves replacing all 3×3 regular convolutions in the conv4-conv5 layers of ResNet50 with dilated convolutions [20]. This modification expands the model’s perceptual field without increasing computational complexity or compromising resolution. Additionally, the different perceptual fields obtained from various convolutional layers facilitate the extraction of multi-scale contextual information, thereby enabling the effective detection of traffic sign targets at different scales. Please refer to Fig. 3 for a visual representation of the details.

The regular convolution operation can be represented by Eq. 1, while a dilated convolution is defined as Eq. 2.



Fig. 2. Visualization of multiple data augmentation fusion. The ground truth is represented by the green box. (a) demonstrates the application of the Mosaic method, where four images are randomly stitched together. This technique enhances the richness of the image background and increases the diversity of target scales within a single image. (b) depicts the use of Mixup, which involves scaling two sample-label data pairs together to generate a new sample with an adjusted label count. This method effectively improves the accuracy of image detection. (Color figure online)

Let $\mathbf{y}(m, n)$ denote the result of dilated convolution between an input signal $\mathbf{I}(m, n)$ and a filter $\mathbf{F}(i, j)$, where \mathbf{I} has a length and width of M and N , respectively. The formulation of the regular convolution operation is given by Eq. 1:

$$\mathbf{y}(m, n) = \sum_{i=1}^M \sum_{j=1}^N \mathbf{I}(m+i, n+j) * \mathbf{F}(i, j) \quad (1)$$

In contrast, the dilated convolution introduces a hyperparameter known as the dilation rate, denoted by r . This parameter defines the spacing between values as the convolution kernel processes the data. The formulation of the dilated convolution can be expressed as Eq. 2:

$$\mathbf{y}(m, n) = \sum_{i=1}^M \sum_{j=1}^N \mathbf{I}(m+r \times i, n+r \times j) * \mathbf{F}(i, j) \quad (2)$$

When the dilation rate r is set to 1, the dilated convolution reduces to a regular convolution.

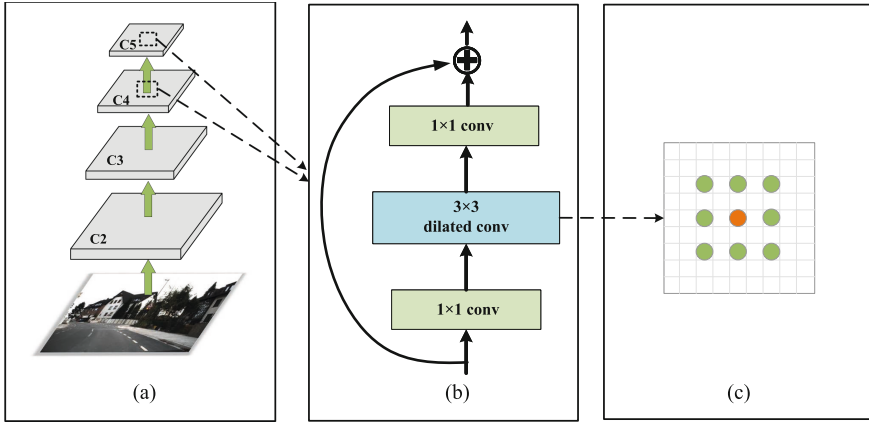


Fig. 3. A detailed design of MDRNet. (a)The network architecture of MDRNet. (b) A residual module is employed in MDRNet to enhance feature extraction and localization. This module consists of two 1×1 regular convolutional modules and a 3×3 dilated convolutional module. (c) The dilated convolutional module showcases the spacing between dots.

2.3 Normalization Methods

Smaller batch sizes have been shown to be more robust to variations between training and test sets [8]. For pixel-level image generation tasks like object detection and image segmentation, memory consumption limits the batch size to be small [6]. Moreover, the large image size of traffic signs and the constraints of general hardware resources further support the use of smaller batch sizes.

However, the effectiveness of BN (Batch Normalization) decreases significantly with small batches, limiting its applicability in micro-batch training. To address this limitation, we employ GN (Group Normalization), which divides channels into groups and calculates the mean and variance within each group for normalization. GN demonstrates stable accuracy across a wide range of batch sizes compared to BN. GN differs from BN in terms of the statistical range over which the mean and variance are calculated. We also introduce WS (Weight Standardization) [13] to further enhance the model’s generalization ability and network performance. WS provides regularization without compromising information exchange, thereby improving model generalization. The joint application of GN and WS is illustrated in Fig. 4. The fusion of GN and WS is expressed as follows:

$$\hat{\mathbf{x}} = \left[\hat{\mathbf{x}}_{i,j} \mid \hat{\mathbf{x}}_{i,j} = \frac{1}{\sigma_{i,\cdot}} \left[\sum_{k,l} \mathbf{x}_{i-k,j-l} \cdot \left[\frac{1}{\sigma_w} (\mathbf{\Gamma}_{k,l} - \boldsymbol{\mu}_w) \right] - \boldsymbol{\mu}_i \right] \right] \quad (3)$$

where \mathbf{x} is the input tensor, $\mathbf{\Gamma}$ is the convolution kernel, and \mathbf{y} is the output tensor. i and j denote the indexes of the output tensor, and k and l denote the

indexes of the convolution kernel, respectively. In WS, the statistical domain of the mean and standard deviation of the weight parameters is each channel. The $\mu_{i,\cdot}$ and $\sigma_{i,\cdot}$ in Eq. 4 are the mean and standard deviation, calculated as σ_i and μ_i .

$$\mu_{i,\cdot} = \frac{1}{m} \sum_{k \in \mathcal{S}_i} \mathbf{x}_k, \quad \sigma_i = \sqrt{\frac{1}{m} \sum_{k \in \mathcal{S}_i} (\mathbf{x}_k - \mu_i)^2 + \epsilon} \quad (4)$$

\mathcal{S}_i is the set of pixels for which the mean and variance are calculated, and \mathcal{S}_i of GN is defined as

$$\mathcal{S}_i = \left\{ k \mid k_N = i_N, \left\lfloor \frac{k_c}{C/G} \right\rfloor = \left\lfloor \frac{i_c}{C/G} \right\rfloor \right\} \quad (5)$$

where G is the number of groups (default value is 32) and C/G is the number of channels per group. $\lfloor \cdot \rfloor$ represents the floor operation. GN computes μ and σ along the (H, W) axes and along a group of C/G channels.

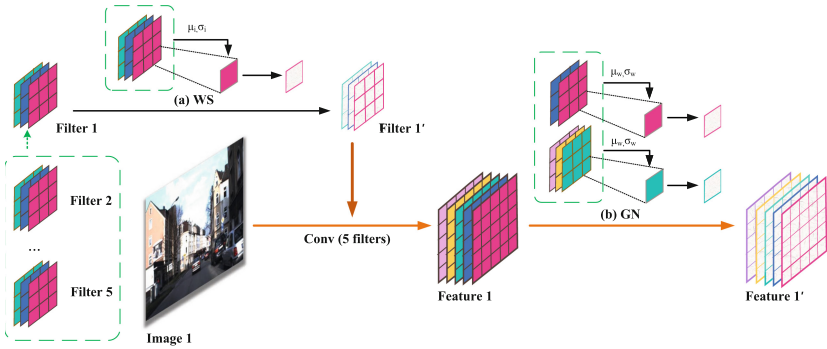


Fig. 4. The joint application of GN and WS involves utilizing μ_i , σ_i , μ_w , and σ_w , which represent the mean and variance of the respective statistical intervals. The normalized convolution kernel of Filter 1 is denoted as Filter 1', while the normalized feature layer of Feature 1 is denoted as Feature 1'. It is important to note that the same operation performed on Filter 1 is also repeated for Filter 2 to 5, ensuring consistency across all filters.

3 Experiment

3.1 Implementation Details

Datasets GTSDb [7]: The traffic signs are classified into three categories: triangular warning signs (yellow or red), circular prohibitory signs (red or black), and mandatory signs (blue circles or squares). The GTSDb dataset consists of 900 images with a size of 1360×800 , including 43 types of traffic signs. Similarly,

the CTSD dataset contains 1100 images captured on highways, urban, and rural roads, with sizes of 1024×768 and 1280×720 . It also classifies traffic signs in China into the same three categories as the GTSDB dataset.

Both the CTSD and GTSDB datasets were selected for their inclusion of small targets and multiple scales in complex environments. Figure 5 shows some typical examples from these datasets, where small targets and multiple scales of traffic signs are often present in challenging conditions such as blurred, poorly lit, damaged, and obscured signs. The datasets were divided into training and test sets in a 7:1 ratio.

Evaluation Metrics. In this paper, the evaluation metrics used are AP_{50} , AP_{75} , AP_S , AP_M , AP_L , and AR, which are calculated following the methods described in COCO [10].

Setting. All experiments were conducted on Ubuntu 20.04 using PyTorch 1.9.1. The training process utilized an NVIDIA GeForce GTX 3090 GPU with 24 GB of memory. The SGD optimizer was employed with an initial learning rate of 0.00125, momentum of 0.9, and weight decay of 0.0001.



Fig. 5. Some difficult examples for traffic sign detection.

3.2 Results and Analysis

Comparison with State-of-the-Art Methods on CTSD. We compared our MDCN method with several state-of-the-art approaches on the CTSD dataset. The comparison results are presented in Table 1 and Table 2.

Our proposed MDCN method achieves an impressive 83.0 AP and 87.6 AR, surpassing all other methods in the table. It demonstrates a strong balance between minimizing false detections and reducing missed detections. In terms of detection metrics, MDCN not only exhibits significant improvements over the

original Faster R-CNN, but also outperforms both two-stage and single-stage target detection algorithms. It outperforms the second-best DH Faster R-CNN by 1.7 AP and surpasses the lowest AP YOLOF by a remarkable 14.9 AP. MDCN outperforms other methods in all metrics, except for the APL metric. This demonstrates the superiority of our method in multi-scale detection.

Furthermore, our proposed method achieves outstanding results for all three major categories of traffic signs, particularly for the mandatory category, where it outperforms all other detection methods by at least 3 points. This is due to the variable shapes and scales of directional traffic signs, with their rectangular aspect ratios often being more extreme compared to other traffic signs. Consequently, the detection of multi-scale signs in this category poses greater challenges.

Overall, the results validate the effectiveness of our MDCN method in achieving superior performance in multi-scale traffic sign detection.

Table 1. Comparison with the state-of-the-art methods on the CTSD dataset. Set Wa is the abbreviation of Warning, Pr is the abbreviation of Prohibitory, and Ma is the abbreviation of Mandatory.

Method	Backbone	Epo	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	Wa	Pr	Ma
Dynamic rcnn [15]	ResNet-50	1x	81.2	97.9	95.0	60.9	82.8	88.9	85.3	79.9	83.1	80.4
Retinanet [9]		2x	76.5	92.4	90.1	49.6	79.4	82.2	84.3	73.0	78.3	78.1
Cascade rcnn [2]		2x	81.2	96.4	96.0	61.9	83.4	91.4	85.4	80.4	83.6	79.6
Baseline [15]		1x	79.0	96.2	94.0	61.0	81.7	85.3	84.1	77.4	81.7	78.1
DH Faster rcnn [18]		1x	81.3	98.3	95.3	64.3	83.1	86.3	84.5	80.8	82.0	81.1
Libra Faster rcnn [12]		1x	80.0	97.3	94.9	63.7	82.4	87.2	85.6	78.7	82.5	78.8
YOLOF [3]		1x	68.1	94.5	81.6	39.4	71.4	77.4	74.7	69.0	64.9	70.2
VfNet [21]		1x	78.4	97.0	93.7	62.5	80.5	88.7	84.7	78.7	80.0	76.4
ATSS [23]		1x	81.1	97.6	96.7	65.5	82.4	87.0	85.6	80.3	81.1	81.9
TOOD [5]		1x	79.8	98.4	95.5	63.4	81.5	88.7	83.8	80.5	79.9	79.1
MDCN		MDRNet	1x	83.0	98.3	98.1	70.5	83.9	90.8	87.6	81.7	84.6

Comparison with State-of-the-Art Methods on GTSDB. We conducted a comparison between MDCN and several popular detection algorithms on the GTSDB dataset, which is widely recognized as a representative dataset in the field of traffic sign object detection. Table 1 presents the results of this comparison.

Among the two-stage algorithms, DH Faster R-CNN and Cascade R-CNN achieved commendable detection performance with 77.9 AP and 76.6 AP, respectively. However, MDCN surpassed them by achieving an impressive 78.7 AP. Furthermore, MDCN outperformed the baseline by 5.5 AP and surpassed YOLOF, which had the lowest AP, by a significant factor of 1.48. Overall, MDCN exhibited superior detection accuracy compared to other methods, particularly for the

warning class of traffic signs that typically occupy a smaller proportion of the image.

Additionally, MDCN demonstrated higher AP scores in the AP_S and AP_L metrics compared to all other methods. Although its AP_M was slightly lower than Dynamic R-CNN, Cascade R-CNN, and DH Faster R-CNN by 0.6 AP_M , 1.2 AP_M , and 2.0 AP_M , respectively, MDCN still outperformed them by significant margins in the AP_S metric (10.1 AP_S , 13.1 AP_S , and 6.8 AP_S , respectively). Moreover, MDCN’s AP_L was higher than that of these methods. These results indicate that our method exhibits superior and more stable performance in multi-scale detection.

In conclusion, MDCN demonstrated excellent detection performance on the GTSDB dataset, affirming the effectiveness and generalization capabilities of our proposed model.

Table 2. Comparison with the state-of-the-art methods on the GTSDB dataset. Set Wa is the abbreviation of Warning, Pr is the abbreviation of Prohibitory, and Ma is the abbreviation of Mandatory.

Method	Backbone	Epo	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	Wa	Pr	Ma
Dynamic rcnn [15]	ResNet-50	1x	76.2	96.1	92.5	64.3	82.7	85.6	80.2	76.6	76.3	75.7
Retinanet [9]		2x	67.3	85.5	79.8	44.4	79.5	78.8	76.3	72.3	75.7	53.8
Cascade rcnn [2]		2x	76.6	95.9	91.2	61.3	83.3	82.5	80.6	75.9	78.5	75.4
Baseline [15]		1x	73.2	96.5	92.1	56.9	81.6	85.4	77.8	74.8	77.5	67.3
DH Faster rcnn [18]		1x	77.9	97.8	93.5	67.6	84.1	76.3	82.1	73.9	79.7	80.3
Libra Faster rcnn [12]		1x	75.8	98.9	92.6	63.8	80.9	82.9	79.9	75.3	75.5	76.7
YOLOF [3]		1x	53.3	93.3	62.2	39.7	61.7	59.6	64.0	47.4	58.6	53.9
VfNet [21]		1x	72.2	95.6	85.6	56.6	80.6	78.4	78.0	67.1	76.7	73.0
ATSS [23]		1x	75.4	97.1	92.3	57.7	82.0	81.7	79.8	73.2	77.1	75.8
TOOD [5]		1x	76.1	98.0	92.7	67.4	82.0	83.4	82.1	76.5	76.6	75.3
MDCN	MDRNet	1x	78.7	98.0	95.2	74.4	82.1	85.9	83.2	81.2	77.0	78.1

3.3 Ablation Studys

We performed a series of ablation experiments to demonstrate the effectiveness of MDCN for the detection of small traffic sign targets at multiple scales in real-world complex environments.

Component Ablation Studies of MDCN. The effectiveness of various optimization components in improving the performance of the baseline model was evaluated through experiments on the CTSD and GTSDB datasets, and the results are summarized in Table 3 and Table 4.

The addition of GN led to improvements of 1.4 and 3.4 in AP on the CTSD and GTSDB datasets, respectively, highlighting its positive impact on detecting

small targets. Combining GN with WS further enhanced the detection performance, resulting in improvements of 1.5 AP and 1.2 AP, as well as 8.9 and 15.4 in APS, respectively, compared to GN alone. This demonstrates that the combination of GN and WS contributes to improved detection of small targets.

The integration of MDRNet significantly improved the AP values by 3.7 AP and 5.3 AP on both datasets compared to the baseline. Moreover, it led to notable improvements in APS, APM, and APL, with gains of 8.9, 2.6, and 4.1, respectively, on both datasets. The largest improvement was observed in APS, indicating the effectiveness of MDRNet in detecting small targets. This improvement can be attributed to the ability of MDRNet to capture more contextual information, which helps reduce the rates of false detections and missed detections for small targets.

Furthermore, the inclusion of the MDEF method resulted in additional improvements of 4.0 AP and 5.5 AP on the two datasets compared to the baseline. MDEF effectively expanded the number of small targets in the dataset and disrupted the regular positioning of traffic signs on the road, thereby preventing overfitting to specific road environments.

Visualizations in Fig. 6 further demonstrate the superior performance of MDCN compared to the baseline. MDCN exhibits greater sensitivity to traffic sign objects with uneven positional distribution and a wide range of scales. Additionally, it demonstrates better detection performance for small objects, as evident from the second row of the visualization.

Table 3. Ablation study on the effectiveness of the various MDCN component modules on CTSD dataset. MDEF is short for Multiple Data augmentation Fusion.

Method	GN	WS	MDRNet	MDEF	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MDCN	-	-	-	-	79.0	96.2	94.0	61.0	81.7	85.3
	✓	-	-	-	80.4	97.3	96.2	70.2	82.6	84.7
	✓	✓	-	-	81.9	97.1	97.0	68.9	83.6	90.3
	✓	✓	✓	-	82.7	97.6	97.2	69.9	84.3	89.4
	✓	✓	✓	✓	83.0	98.3	98.1	70.5	83.9	90.8

Ablation Experiments of Dilated Convolution Embedding Positions.

In our ablation experiments, we investigated the impact of dilated convolution on the task of traffic sign detection. Dilated convolution is known for its ability to expand the receptive field while preserving resolution. However, it is crucial to carefully select the positions where regular convolutions are replaced with dilated convolutions to achieve optimal results.

We examined the effect of incorporating dilated convolutions from the bottom-up, starting from the conv2 to conv5 layers in the backbone network. The results, as shown in Table 5 and Table 6, revealed that the addition of dilated

Table 4. Ablation study on the effectiveness of the various MDCN component modules on GTSDB dataset. MDEF is short for Multiple Data augmentation Fusion.

Method	GN	WS	MDRNet	MDEF	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MDCN	-	-	-	-	73.2	96.5	92.1	56.9	81.6	85.4
	✓	-	-	-	76.6	95.9	95.1	69.9	82.0	84.6
	✓	✓	-	-	77.8	97.7	94.3	72.3	81.2	82.5
	✓	✓	✓	-	78.5	97.1	95.6	73.1	82.2	89.6
	✓	✓	✓	✓	78.7	98.0	95.2	74.4	82.1	85.9

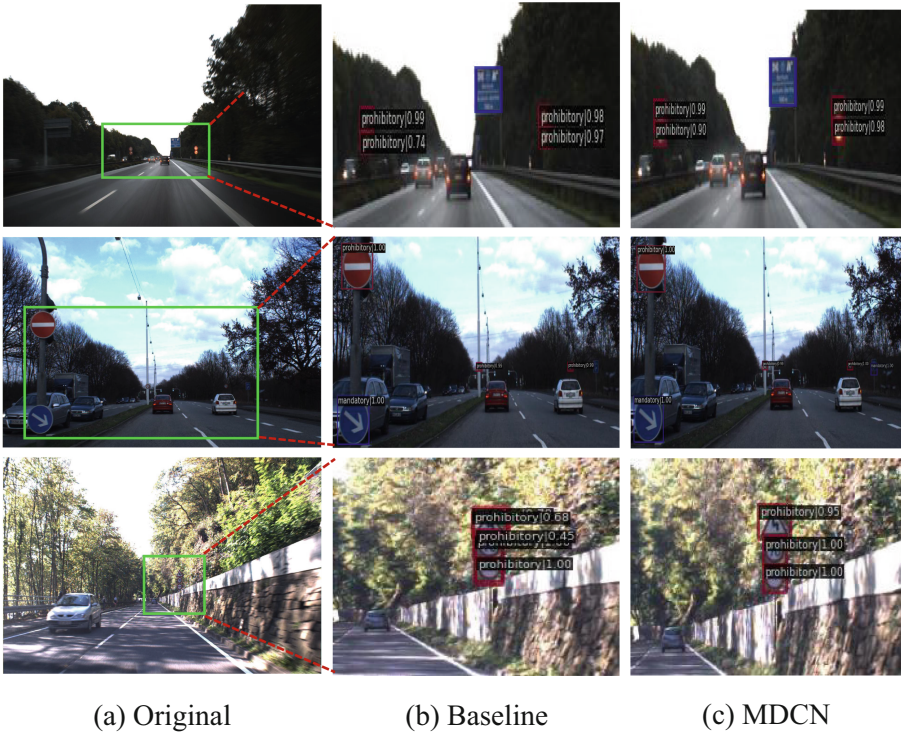


Fig. 6. Comparison of detection results between baseline and MDCN. (a) is the original image. (b) is the detection result using baseline. (c) is the detection result using MDCN. The images in the last two columns are taken from the green box in the first column. (Color figure online)

convolutions in the last two layers of the convolutional hierarchy yielded the most significant improvements. Specifically, on the CTSD dataset, the APs increased to 81.7 and 81.6, while on the other dataset, they improved to 76.4 and 77.6, respectively.

This observation can be attributed to the fact that deeper convolution layers tend to lose internal data structures, resulting in the loss of some crucial information related to small objects. By replacing the deeper convolutions with dilated convolutions, we were able to improve the localization accuracy of small targets.

Building upon these findings, we gradually added dilated convolution modules in pairs until all 3×3 convolutions between c4 and c5 were replaced. This configuration resulted in the optimal embedding position for MDRNet, achieving the best performance of 83.0 AP and 78.7 AP on both datasets.

These ablation experiments highlight the importance of selecting appropriate positions for dilated convolutions, and they provide valuable insights into the optimal design choices for our MDRNet architecture.

Table 5. Ablation study with gradually dilated modules on CTSD dataset. In the setting column, c2, c3, c4 and c5 stand for applying dilated convolution at c2, c3, c4 and c5 layers.

Method	c2	c3	c4	c5	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR
MDRNet	✓	-	-	-	81.3	96.9	95.9	66.0	83.2	91.1	86.7
	-	✓	-	-	81.4	96.8	96.4	66.6	83.5	88.7	86.1
	-	-	✓	-	81.7	97.9	97.8	68.0	82.8	89.4	86.4
	-	-	-	✓	81.6	97.0	96.0	67.1	83.6	90.2	87.4
	✓	✓	-	-	80.5	97.1	96.3	66.6	81.2	89.9	85.0
	-	✓	✓	-	80.4	97.6	97.1	66.5	82.3	87.6	85.9
	-	-	✓	✓	83.0	98.3	98.1	70.5	83.9	90.8	87.6

Table 6. Ablation study with gradually dilated modules on GTSDDB dataset. In the setting column, c2, c3, c4 and c5 stand for applying dilated convolution at c2, c3, c4 and c5 layers.

Method	c2	c3	c4	c5	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR
MDRNet	✓	-	-	-	75.7	95.6	90.3	70.2	79.6	84.2	80.8
	-	✓	-	-	75.9	96.1	93.4	65.0	80.1	84.6	81.5
	-	-	✓	-	76.4	95.9	92.7	66.3	81.8	86.7	80.7
	-	-	-	✓	77.6	97.4	93.1	72.1	81.3	85.4	82.6
	✓	✓	-	-	74.9	95.4	94.0	68.0	78.6	85.0	80.3
	-	✓	✓	-	77.2	96.7	93.7	65.7	81.6	84.2	82.7
	-	-	✓	✓	78.7	98.0	95.2	74.4	82.1	85.9	83.2

4 Conclusion

The primary objective of this paper is to enhance the detection performance of small multi-scale traffic signs in complex real-world environments. To achieve this goal, we propose an MDCN traffic sign detection method based on Faster R-CNN. Our approach incorporates MDRNet as the backbone network, which effectively increases the perceptual field and sampling rate. This improvement enhances the feature representation capability for targets of different scales and those that may be partially obscured. Considering the characteristics of the traffic sign dataset, hardware resource limitations, and the dataset's high resolution, we introduce a normalized combination method at the backbone and pooling layers. This technique accelerates the convergence speed and improves the generalization ability of the model. Furthermore, we employ a multi-data augmentation fusion strategy to expand small targets and enhance the network's robustness. The effectiveness and generalization of our proposed MDCN method are validated through comprehensive evaluations on the CTSD and GTSDB datasets. The results demonstrate its superiority in detecting small traffic signs and establish its competitiveness compared to other state-of-the-art methods.

For future research, we intend to explore traffic sign recognition techniques tailored specifically for mobile terminals. By adapting the methodology to mobile devices, we aim to further expand the applicability and practicality of traffic sign detection in real-world scenarios.

Acknowledgment. This work is supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (2020D01C33).

References

1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934) (2020)
2. Cai, Z., Vasconcelos, N.: Cascade r-CNN: high quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(5), 1483–1498 (2019)
3. Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., Sun, J.: You only look one-level feature. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13039–13048 (2021)
4. Elsaagheer Mohamed, S.A., AlShalfan, K.A.: Intelligent traffic management system based on the internet of vehicles (IoV). *J. Adv. Transp.* **2021**, 1–23 (2021)
5. Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: TOOD: task-aligned one-stage object detection. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3490–3499. IEEE Computer Society (2021)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969 (2017)
7. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: the German traffic sign detection benchmark. In: *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE (2013)

8. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: generalization gap and sharp minima. arXiv preprint [arXiv:1609.04836](https://arxiv.org/abs/1609.04836) (2016)
9. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of The IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
10. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
11. Liu, Y., Peng, J., Xue, J.H., Chen, Y., Fu, Z.H.: Tsingnet: scale-aware and context-rich feature learning for traffic sign detection and recognition in the wild. *Neurocomputing* **447**, 10–22 (2021)
12. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra R-CNN: towards balanced learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 821–830 (2019)
13. Qiao, S., Wang, H., Liu, C., Shen, W., Yuille, A.: Micro-batch training with batch-channel normalization and weight standardization. arXiv preprint [arXiv:1903.10520](https://arxiv.org/abs/1903.10520) (2019)
14. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint [arXiv:1804.02767](https://arxiv.org/abs/1804.02767) (2018)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems* 28 (2015)
16. Shen, L., You, L., Peng, B., Zhang, C.: Group multi-scale attention pyramid network for traffic sign detection. *Neurocomputing* **452**, 1–14 (2021)
17. Wang, J., Chen, Y., Dong, Z., Gao, M.: Improved yolov5 network for real-time multi-scale traffic sign detection. *Neural Comput. Appl.* **35**(10), 7853–7865 (2022)
18. Wu, Y., et al.: Rethinking classification and localization for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10186–10195 (2020)
19. Yao, Y., Han, L., Du, C., Xu, X., Jiang, X.: Traffic sign detection algorithm based on improved yolov4-tiny. *Signal Process.: Image Commun.* **107**, 116783 (2022)
20. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint [arXiv:1511.07122](https://arxiv.org/abs/1511.07122) (2015)
21. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: Varifocalnet: an iou-aware dense object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8514–8523 (2021)
22. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: beyond empirical risk minimization. arXiv preprint [arXiv:1710.09412](https://arxiv.org/abs/1710.09412) (2017)
23. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9759–9768 (2020)