MSPNet: Multi-level Semantic Pyramid Network for Real-Time Object Detection

Ji Li and Yingdong Ma(✉)

Inner Mongolia University, The Inner Mongolia Autonomous Region, College Road no. 235, Hohhot, China
csmyd@imu.edu.cn

Abstract. With increasing demand of running Convolutional Neural Networks (CNNs) on mobile devices, real-time object detection has made great progress in recent years. However, modern approaches usually compromise detection accuracy to achieve real-time inference speed. Some light weight top-down CNN detectors suffer from problems of spatial information loss and lack of multi-level semantic information. In this paper, we introduce an efficient CNN architecture, the Multi-level Semantic Pyramid Network (MSPNet), for real-time object detection on devices with limited resource and computational power. The proposed MSPNet consists of two main modules to enhance spatial details and multi-level semantic information. The multi-scale feature fusion module integrates different level features to tackle the problem of spatial information loss. Meanwhile, a light weight multi-level semantic enhancement module is developed which transforms multiple layer features to strengthen semantic information. The proposed light weight object detection framework has been evaluated on CIFAR-100, PASCAL VOC and MS COCO datasets. Experimental results demonstrate that our method achieves state-of-the-art results while maintains a compact structure for real-time object detection.

Keywords: Real-time object detection · Multi-scale feature fusion · Multi-level semantic information

1 Introduction

Real-time object detection is a fundamental computer vision task. With rapid development of mobile devices, there are increasing interests in designing Convolutional Neural Network models (CNNs) for speed sensitive applications, such as robotics, video surveillance, autonomous driving and augmented reality. Real-time object detection on mobile devices is a challenging task due to state-of-the-art CNNs require high computational resources beyond the capabilities of many mobile and embedded devices.

To tackle the problem, some light weight networks adopt small backbone and simple structure that compromise detection accuracy to inference speed. For example, the Light-head R-CNN [1] implemented real-time detection by using a small backbone. However, small backbone makes the network prone to overfitting. Iandola et al. proposed the

A. Campilho et al. (Eds.): ICIAR 2020, LNCS 12132, pp. 76–88, 2020.
https://doi.org/10.1007/978-3-030-50516-5_7

SqueezeNet [2] which uses a fire module to reduce parameters and computational cost. Though the method is simple and effective, the lack of effective spatial details leads to accuracy degradation. Meanwhile, it is difficult for light weight networks with small backbone to provide feature maps with large receptive field. Global Convolution Network [3] utilizes "large kernel" to enlarge the receptive field, while it leads to a sharp increase in computational cost. ICNet [4] adopts a multi-branch framework, in which coarse prediction map obtained from deep feature maps are refined by medium- and high-resolution features. This method enhances spatial details but the semantic information is computed mainly from deep feature maps. In fact, feature maps in previous layers not only contain spatial detail but have different level semantic information. With these observations, we aim to implement a CNN to achieve accurate object detection while maintain compact architecture. The proposed multi-level semantic pyramid network consists of two modules to integrate multiple layer features. The overall framework is shown in Fig. 1. The main contributions of this paper are summarized as follows:

1. We propose a light weight network architecture that consists of a multi-scale feature fusion (MFF) module to preserve spatial information and a multi-level semantic enhancement (MSE) module to extract different level semantic features. The new model enhances network representation ability for both fine-level spatial details and high-level semantic information. Meanwhile, the proposed model maintains a compact structure for real-time object detection.
2. The multi-scale feature fusion module integrates different scale features to enrich spatial information. In the light weight multi-level semantic enhancement module, features of various layers are transformed to shallow, medium and deep features. Shallow features and deep features are further combined to compute global semantic clues. These features are aggregated to generate semantic segmentation maps which are used as semantic guidance to improve detection performance.
3. With 304×304 input images, MSPNet achieves 78.2% mAP on the PASCAL VOC 07/12 and 30.1% AP on the MS COCO datasets, outperforming state-of-the-art light weight object detectors. Furthermore, experiments demonstrate the efficiency of MSPNet, e.g. our module operates at 126 frames per second (FPS) on PASCAL VOC 2007 with a single GTX1080Ti GPU.

2 Related Work

2.1 Light Weight Deep Neural Network

To construct a compact detector, some models either compress and prune typical CNNs or adopt a light weight structure. As an example, the MobileNet [5, 6] utilizes depthwise separable convolution to build a light weight deep neural network. Other light weight networks, such as ShuffleNet [7, 8] reduces computational cost by using the pointwise group convolution and adopt the channel shuffle operation to obtain feature maps from different groups. The ThunderNet adopts a light weight backbone and a compressed RPN network with discriminative feature representation to realize effective two-stage detector [9]. In [10], Wang et al. proposed the PeleeNet which consists of a stem block and a set of dense blocks. Different from other light weight networks, the PeleeNet adopts conventional convolutions to achieve efficient architecture.

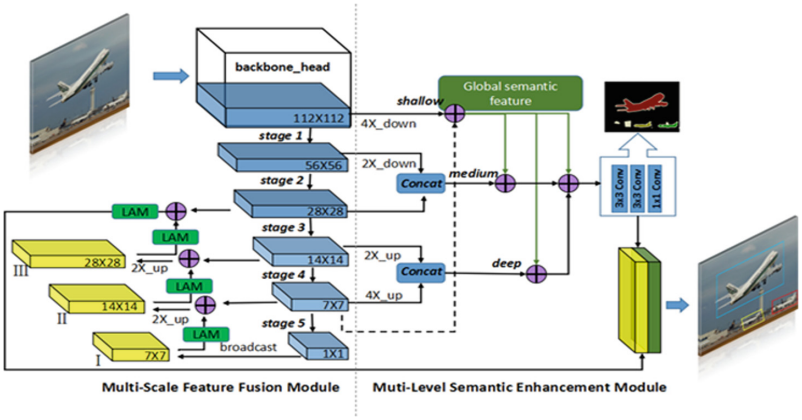

Fig. 1. The multi-level semantic pyramid network. The main structure is composed of the feature fusion module (left half part) and the semantic enhancement module (right half part). LAM: light weight attention module.

2.2 Multi-layer Feature Fusion

Combination of multi-layer feature maps is a common method to make better use of different level features. Lin et al. developed a top-down feature pyramid architecture with lateral connections to construct the FPN model [11]. The feature pyramid enhances semantic information by combing semantically strong deep feature maps with high-resolution shallow feature maps. In the SSD network [12], additional layers are added to the baseline model to enrich semantic clues. The network makes prediction from multiple layer feature maps with different resolutions. DSSD adds deconvolutional layers to the top of SSD network to build a U-shape structure [13]. These feature maps are then combined with different scale feature maps to enhance different level features. The STDN [14] adopts multiple layer features with different resolutions and the scale-transfer layers to generate a large size feature map. In [15], Kong et al. proposed a light weight reconfiguration architecture to combine multiple level features. The architecture employs global attention to extract global semantic features which followed by a local reconfiguration to model low-level features.

2.3 Attention Mechanism

The attention mechanism has been successfully applied in many computer vision tasks to boost the representational power of CNNs. The residual attention network [16] is built by stacking multiple attention modules to generate attention-aware features. The trunk branch performs feature processing and the soft mask branch learns weight for output features. In [17], Hu et al. proposed a compact squeeze and excitation structure to adjust output response by modeling the relationship between channel features. The channel-wise attention learned by SENet is used to select important feature channels. In CBAM [18] and BAM [19], attention modules are integrated with CNNs to compute attention maps in both channel and spatial dimensions. These attention modules sequentially

transmit input feature map to channel attention module and spatial attention module for feature refinement. Li et al. propose the pyramid attention network which combines attention mechanism with spatial pyramid to provide pixel-level attention for high-level features extraction [20].

Table 1. Architecture of the multi-level semantic pyramid network

Stage	Output size	Operation	
Input	224 × 224 × 3		
PeleeNet_head			
Multi-scale feature fusion module			
MFF feature I	7 × 7 × 704	Broadcasting	
		Element-wise sum	
MFF feature II	14 × 14 × 704	2× up sampling	
		Element-wise sum	
MFF feature III	28 × 28 × 704	2× up sampling	
		Element-wise sum	
Global semantic feature			
Multi-level semantic enhancement module			
MSE deep feature	28 × 28 × 704	2/4× up sampling	Sum
		Concatenate	
MSE medium feature	28 × 28 × 704	2x down sampling	
		Concatenate	
MSE shallow feature	28 × 28 × 704	4× down sampling	
Classification			

2.4 Light Weight Semantic Segmentation

Wu et al. introduced a light weight context guided network for semantic segmentation [21]. The CGNet contains multiple context guided blocks which learns joint features by using a local feature extractor, a surrounding context extractor, and a global context extractor. The Light weight RefineNet implements real-time segmentation by using light weight residual convolutional units and light weight chained residual pooling [22]. By replacing 3 × 3 convolutions with 1 × 1 convolutions, the method reduces model parameters while achieving similar performance to the original RefineNet [23]. Zhang et al. proposed the detector with enriched semantics network which consists of a detection branch and a segmentation branch [24].

Different to these works, we aim to improve the discriminative power of light weight network by enriching spatial details and high-level semantic information. The proposed network extracts multi-scale spatial details and different level semantic information in two modules simultaneously. These enriched features are further integrated to generate semantic segmentation map which is used to guide object detection.

3 Multi-level Semantic Pyramid Network

In the task of object detection, both spatial details and object-level semantic information are crucial to achieve high accuracy. However, it is difficult to meet these demands simultaneously in a light weight top-down CNN structure. In this work, we introduce the multi-level semantic pyramid network to solve the problem. The proposed MSPNet applies a feature fusion module to strengthen multi-scale spatial features. A light weight multi-level semantic enhancement module is developed, which aggregates global semantic features with different features to enhance multiple level semantic information. The overall structure of the MSPNet is shown in Table 1.

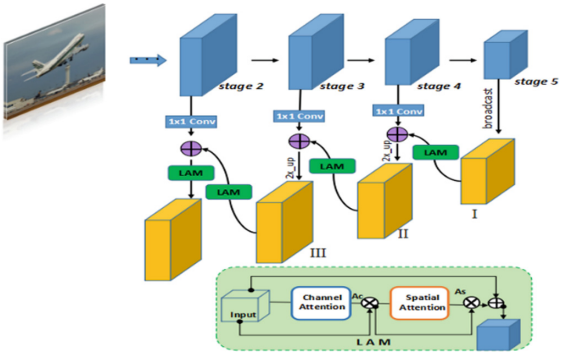

Fig. 2. The multi-scale feature fusion module.

3.1 Multi-scale Feature Fusion Module

Most light weight networks make prediction mainly on deep feature maps. While deep features provide rich semantic information, the top-down CNN structure suffers from fine-level spatial information loss. Thus, lack of suitable strategy to preserve multi-scale spatial details is one of the main issues of light weight CNNs.

We present a feature fusion module to tackle this problem. Specifically, a multi-level feature pyramid is built in the feature fusion module. The proposed MFF module refines features by aggregating multiple scale spatial features from different pyramid layers. The structure of the MFF module is shown in Fig. 2. In the module, we build a four-stage feature pyramid to fuse multiple level features. Firstly, the broadcasting is applied to the global average pooling layer. Secondly, feature maps with smaller sizes are $2\times$ upsampled by the bilinear interpolation to match spatial dimensions of previous layer feature maps. Then feature maps of two levels are merged by using element-wise add.

The Light weight Attention Module. In the proposed multi-scale feature fusion module, a light weight attention module (LAM) is applied to learn weight for multiple level features. The LAM consists of a channel attention module (CAM) and a spatial attention module (SAM) as shown in Fig. 2. The LAM sequentially transforms input features to the channel attention module and the spatial attention module. The channel and spatial attention are computed as:

$$Ac = Sigmoid(MaxPool(F)) \tag{1}$$

$$Ac' = F \otimes Ac \tag{2}$$

$$\text{As} = \text{Sigmoid}\left\{Conv^1\left(Concat\left(M^3, M^6\right)\right)\right\} \tag{3}$$

$$M^3 = MaxPool\left\{Conv^{d3}\left(Conv^1\left(Ac'\right)\right)\right\} \tag{4}$$

$$M^6 = MaxPool\left\{Conv^{d6}\left(Conv^1\left(Ac'\right)\right)\right\} \tag{5}$$

$$As' = Ac' \otimes As \tag{6}$$

Where, Conv^1 is 1×1 convolution, Conv^{d3} and Conv^{d6} are dilated convolution with dilation rate of 3 and 6 respectively. For the given input feature F, the refined feature map F' is computed as:

$$F' = F \otimes As' \tag{7}$$

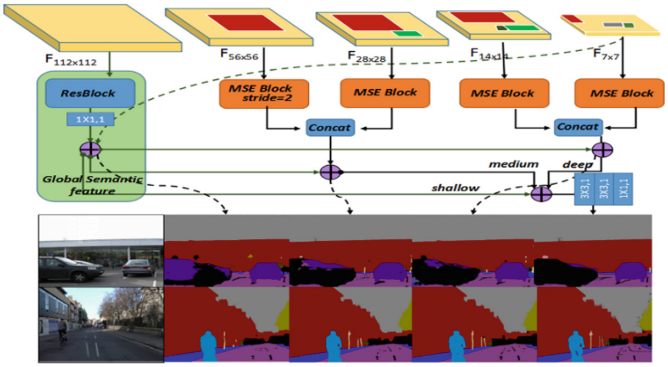

Fig. 3. The multi-level semantic enhancement module.

3.2 Multi-level Semantic Enhancement Module

We develop a light weight multi-level semantic enhancement module to further improve detection accuracy. In the semantic module, multi-level features are combined to form

shallow, medium and deep features. Figure 3 shows the architecture. Specifically, features in different levels are resized to 28 × 28 pixels by downsampling and upsampling. The upsampled $F_{7\times7}$ and $F_{14\times14}$ are concatenated to get the deep features. Likewise, the medium features are obtained by integrating $F_{28\times28}$ features and $F_{56\times56}$ features. $F_{112\times112}$ is used as the shallow features. However, only deep features are not enough as segmentation also require object boundaries information to facilitate different scale objects localization. In addition, when feature maps propagate from top level to low level, progressively upsampling might cause information dilution. we combine the $F_{7\times7}$ features and the $F_{112\times112}$ features to yield global semantic features. These global semantic features not only contain high-level semantic cues, but also provide fine-level object location information. The deep features and medium features are then combined with global semantic features to form output features.

In the MSE module, we apply a two-path MSE block to reduce training time. For each level feature maps, the input is transformed by a stack of three convolutional layers. The two 1 × 1 convolutions reduces and then restores dimensions that makes the 3 × 3 DW_Conv [5] running on features with small dimensions. MSE Block has a shortcut path which downsamples input features by using the max pooling. The shortcut path is designed to enhance semantic information of different level features.

4 Experiments

In this section, we evaluate the effectiveness of MSPNet on CIFAR-100 [25], PASCAL VOC [26] and MS COCO [27] benchmarks. The CIFAR-100 consists of 60,000 32 × 32 color images in 100 classes where the training and testing sets contain 50,000 and 10,000 images respectively. We use VOC 2007 trainval and VOC 2012 trainval as the training data, and use VOC 2007 testval as the test data. For MS COCO, we use a popular split which takes trainval35k for training, minival for validation, and we report results on test-dev 2017. Ablation experiments are also conducted to verify the effectiveness of different components.

Table 2. Experimental results on CIFAR-100 dataset.

Module	Params	Error (%)
ResNet 50 [28]	23.71M	21.49
ResNet 101 [28]	42.70M	20.00
WideResNet 28 [29]	23.40M	20.40
ResNeXt 29 [30]	34.52M	18.18
DenseNet-BC-250 (k = 24) [31]	15.30M	19.64
MobileNet [5]	3.30M	18.30
ShuffleNet [7]	2.50M	16.60
PreResNet 110 [32]	1.73M	22.2
Pelee [10]	1.60M	15.90
MSPNet	**2.20**M	**10.3**

4.1 Experiments on CIFAR-100

The MSPNet is trained on the public platform TensorFlow with batch size of 128. We set the initial learning rate to 10^{-3}. The learning rate changes to 10^{-4} at 60k iterations and 10^{-5} at 100k iterations. Experimental results on CIFAR-100 dataset are shown in Table 2. The MSPNet has 10.3% error rate, about 5.6% lower than the baseline model with slight network parameter increase of 0.6M. Our model outperforms other light weight networks, e.g. compared to MobileNet and ShuffleNet, the MSPNet achieves 8.0% and 6.3% performance improvement with fewer model parameters.

Table 3. Experimental results on PASCAL VOC.

Module	Backbone	Input dimension	MFLOPs	Data	mAP (%)
R-FCN [35]	ResNet-101	600 × 1000	58900	12	77.4
HyperNet [34]	VGG-16	600 × 1000	–	07 + 12	76.3
RON384 [36]	VGG-16	384 × 384	–	12	73.0
SSD300 [12]	VGG-16	300 × 300	31750	07 + 12	77.5
SSD321 [12]	ResNet-101	321 × 321	15400	12	77.1
DSSD321 [13]	ResNet-101	321 × 321	21200	07 + 12	78.6
DES300 [24]	VGG16	300 × 300	–	12	77.1
RefineDet320 [37]	VGG-16	320 × 320	–	12	78.1
DSOD300 [38]	DenseNet	300 × 300	–	07 + 12	77.7
YOLOv2 [33]	Darknet-19	416 × 416	17400	07 + 12	76.8
YOLOv2 [33]	Darknet-19	288 × 288	8360	07 + 12	69.0
PFPNet-R320 [39]	VGG-16	320 × 320	–	12	77.7
Tiny-YOLOv2 [33]	DarkNet-19	416 × 416	3490	07 + 12	57.1
MobileNet-SSD [5]	MobileNet	300 × 300	1150	07 + 12	68.0
MobileNet-SSD [5]	MobileNet	300 × 300	1150	07 + 12 + coco	72.7
Tiny-DSOD [40]	–	300 × 300	1060	07	72.1
Pelee [10]	DenseNet-41	304 × 304	1210	07 + 12 + coco	76.4
Pelee [10]	DenseNet-41	304 × 304	1210	07 + 12	70.9
ThunderNet [9]	SNet146	320 × 320	–	07 + 12	75.1
MSPNet	**Pelee**	**304 × 304**	**1370**	**07 + 12**	**78.2**
MSPNet	**Pelee**	**512 × 512**	**1370**	**07 + 12**	**79.4**

4.2 Experiments on PASCAL VOC

We train the model with an initial learning rate of 0.05 and batch size is 32. The learning rate reduces to 0.005 at 80k iterations. Table 3 lists the experimental results on Pascal VOC dataset. We use the standard mean average precision (mAP) scores with IoU thresholds of 0.5 as evaluation metric. Our model achieves 78.2% mAP, higher than the baseline model by 7.3%. Compared to other light weight models, the proposed MSPNet also has competitive results. For example, we observe 3.1% performance improvement than the ThunderNet [9] with SNet146. Furthermore, our model has better performance than some state-of-the-art CNNs, such as SSD300 [12], YOLOv2 [33], HyperNet [34] and R-FCN [35] with significant computational cost reduction.

Table 4. Experimental results on MS COCO

Module	Backbone	Input dimension	AP	AP_{50}	AP_{75}
ResNet-50 [28]	–	320 × 320	26.5	47.6	–
YOLOv2 [33]	DarkNet-19	416 × 416	21.6	44.0	19.2
SSD300 [12]	VGG-16	300 × 300	25.1	43.1	25.8
SSD512 [12]	VGG-16	512 × 512	28.8	48.5	30.3
DSSD321 [13]	ResNet-101	321 × 321	28.0	46.1	29.2
MDSSD300 [41]	VGG-16	300 × 300	26.8	46.0	27.7
Light-head r-cnn [1]	ShuffleNetv2	800 × 1200	23.7	–	–
RefineDet320 [37]	VGG-16	320 × 320	29.4	49.2	31.3
EFIPNet [42]	VGG-16	300 × 300	30.0	48.8	31.7
DES300 [24]	VGG-16	300 × 300	28.3	47.3	29.4
DSOD300 [38]	DenseNet	300 × 300	29.3	47.3	30.6
RON384++ [36]	VGG-16	384 × 384	27.4	40.5	27.1
PFPNet-S300 [39]	VGG-16	300 × 300	29.6	49.6	31.1
MobileNet-SSD [5]	MobileNet	300 × 300	19.3	–	–
MobileNetv2-SSDLite [6]	MobileNet	320 × 320	22.1	–	–
MobileNetv2 [6]	–	320 × 320	22.7	–	–
Tiny-DSOD [40]	–	300 × 300	23.2	40.4	22.8
Pelee [10]	DenseNet-41	304 × 304	22.4	38.3	22.9
ThunderNet [9]	SNet146	320 × 320	23.6	40.2	24.5
ShuffleNetv1 [7]	–	320 × 320	20.8	–	–
ShuffleNetv2 [8]	–	320 × 320	22.7	–	–
MSPNet	**Pelee**	**304 × 304**	**30.1**	**48.9**	**31.5**
MSPNet	**Pelee**	**512 × 512**	**35.2**	**52.1**	**36.7**

4.3 Experiments on MS COCO

We train the MSPNet with an initial learning rate of 0.05 and the batch size is 32. The learning rate changes to 0.01 at 40k iterations and 0.001 at 45k iterations. The evaluation metric of MS COCO is the average precision (AP) scores, which includes AP_{50} and AP_{75}, with IoU thresholds of 0.5 and 0.75 respectively. As shown in Table 4, MSPNet achieves 30.1% AP, surpasses most light weight networks, such as MobileNet-SSD [5] and PeleeNet [10] with the mostly same computational cost.

4.4 Ablation Experiments

Table 5 shows the results of ablation experiments. Compared to Pelee (70.9% mAP), utilizing feature fusion module alone obtains 3.4% accuracy improvement. Similarly, the third and the fourth row show that using attention module and semantic enhancement module separately increases performance by 1.2% and 3.7%, respectively. In the attention experiment, the feature fusion module and semantic enhancement module are removed and the attention module is applied to refine each stage backbone feature maps. It can be seen from these experiments that multi-scale feature fusion module is necessary to object detection as it enhances both spatial information and high-level semantic features. Utilization of the multi-scale feature fusion module and the multi-level semantic enhancement module separately yields 75.6% mAP (the sixth row) and 75.1% mAP (the seventh row), respectively. The last two experiments show that our model achieves 77.7% mAP on the PASCAL VOC dataset by utilizing the feature fusion module, the attention module, and the semantic enhancement module (without global semantic features). Integrating the global semantic features further improve detection performance to 78.2%.

Table 5. Ablation study of different modules on PASCAL VOC.

	Multi-level semantic pyramid network				
	MFF		MSE		
	Feature fusion module	Attention module	Semantic enhancement module	Global semantic feature	*mAP (%)*
1	Pelee [10]				70.9
2	√				74.3
3		√			72.1
4			√		74.6
5				√	71.8
6	√	√			75.6
7			√	√	75.1
8	√	√	√		77.7
9	√	√	√	√	**78.2**

5 Conclusion

In this paper, we propose a light weight network, the multi-level semantic pyramid network, to implement real-time object detection. The MSPNet consists of a multi-scale feature fusion module and a multi-level semantic enhancement module to improve network representation ability for both spatial details and multi-level semantic information. Specifically, the MFF integrates different level features to collect multi-scale spatial information. The light weight MSE transforms different level features to shallow, medium and deep features. These features are combined with global semantic features to generate semantic segmentation maps which are used as semantic guidance to improve detection performance. Experiments on different datasets demonstrate superior object detection performance of the proposed method as compared with state-of-the-art works.

References

1. Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., Sun, J.: Light-head r-cnn: In defense of two-stage object detector. arXiv preprint arXiv:1711.07264 (2017)
2. Iandola, F.N., Han, S., Moskewicz, M.W., et al.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv preprint arXiv:1602.07360 (2016)
3. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4353–4361 (2017)
4. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: ICNet for real-time semantic segmentation on high-resolution images. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 418–434. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_25
5. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
6. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
7. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
8. Ma, N., Zhang, X., Zheng, H.-T., Sun, J.: ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 122–138. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_8
9. Qin, Z., et al.: ThunderNet: towards real-time generic object detection on mobile devices. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6718–6727 (2019)
10. Wang, R.J., Li, X., Ling, C.X.: Pelee: a real-time object detection system on mobile devices. In: Advances in Neural Information Processing Systems, pp. 1963–1972 (2018)
11. Lin, T.Y., Dollár, P., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
12. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2

13. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017)
14. Tang, Y., et al.: Visual and semantic knowledge transfer for large scale semi-supervised object detection. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 3045–3058 (2017)
15. Kong, T., Sun, F., Huang, W., Liu, H.: Deep feature pyramid reconfiguration for object detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11209, pp. 172–188. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01228-1_11
16. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
18. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
19. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: Bam: Bottleneck attention module. arXiv preprint arXiv:1807.06514. (2018)
20. Li, H., Xiong, P., An, J., Wang, L.: Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180 (2018)
21. Wu, T., Tang, S., Zhang, R., Zhang, Y.: CGNET: a light-weight context guided network for semantic segmentation. arXiv preprint arXiv:1811.08201 (2018)
22. Nekrasov, V., Shen, C., Reid, I.: Light-weight refinenet for real-time semantic segmentation. arXiv preprint arXiv:1810.03272 (2018)
23. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925–1934 (2017)
24. Zhang, Z., Qiao, S., Xie, C., Shen, W., Wang, B., Yuille, A.L.: Single-shot object detection with enriched semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5813–5821 (2018)
25. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Computer Science Department, University of Toronto, vol. 1, no. 4, pp. 7 (2009)
26. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., et al.: The pascal visual object classes (Voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010). https://doi.org/10.1007/s11263-009-0275-4
27. 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
28. He, K., Zhang, X., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
29. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
30. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
31. Huang, G., Liu, Z., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
32. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
33. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)

34. Kong, T., Yao, A., Chen, Y., Sun, F.: HyperNet: towards accurate region proposal generation and joint object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 845–853 (2016)
35. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, pp. 379–387 (2016)
36. Kong, T., Sun, F., Yao, A., Liu, H., Lu, M., Chen, Y.: RON: reverse connection with objectness prior networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5936–5944 (2017)
37. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4203–4212 (2018)
38. Shen, Z., Liu, Z., Li, J., Jiang, Y.G., Chen, Y., Xue, X.: DSOD: learning deeply supervised object detectors from scratch. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1919–1927 (2017)
39. Kim, S.-W., Kook, H.-K., Sun, J.-Y., Kang, M.-C., Ko, S.-J.: Parallel feature pyramid network for object detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11209, pp. 239–256. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01228-1_15
40. Li, Y., Li, J., Lin, W., Li, J.: Tiny-DSOD: lightweight object detection for resource-restricted usages. arXiv preprint arXiv:1807.11013 (2018)
41. Xu, M., et al.: MDSSD: multi-scale deconvolutional single shot detector for small objects. arXiv preprint arXiv:1805.07009 (2018)
42. Pang, Y., Wang, T., Anwer, R.M., Khan, F.S., Shao, L.: Efficient featurized image pyramid network for single shot detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7336–7344 (2019)