Triply Supervised Decoder Networks for Joint Detection and Segmentation

Jiale Cao[1], Yanwei Pang[1*], Xuelong Li[2]
[1]School of Electrical and Information Engineering, Tianjin University
[2]Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University
connor@tju.edu.cn, pyw@tju.edu.cn, li@nwpu.edu.cn

Abstract

Joint object detection and semantic segmentation is essential in many fields such as self-driving cars. An initial attempt towards this goal is to simply share a single network for multi-task learning. We argue that it does not make full use of the fact that detection and segmentation are mutually beneficial. In this paper, we propose a framework called TripleNet to deeply boost these two tasks. On the one hand, to deeply join the two tasks at different scales, triple supervisions including detection-oriented supervision and class-aware/agnostic segmentation supervisions are imposed on each layer of the decoder. Class-agnostic segmentation provides an objectness prior to detection and segmentation. On the other hand, to further intercross the two tasks and refine the features in each scale, two light-weight modules (i.e., the inner-connected module and the attention skip-layer fusion) are incorporated. Because segmentation supervision on each decoder layer are not performed at the test stage and two added modules are light-weight, the proposed TripleNet can run at a real-time speed (16 fps). Experiments on the VOC 2007/2012 and COCO datasets show that TripleNet outperforms all the other one-stage methods on both two tasks (e.g., 81.9% mAP and 83.3% mIoU on VOC 2012, and 37.1% mAP and 59.6% mIoU on COCO) by a single network.

1. Introduction

Object detection and semantic segmentation are two fundamental tasks in the field of computer vision. Most state-of-the-art methods merely focus on one single task (i.e., object detection [42, 31, 28, 4] or semantic segmentation [24, 45, 14, 18]). However, simultaneous object detection and semantic segmentation is very important in many applications, such as self-driving cars and unmanned surface vessels. Thus, joint detection and segmentation is necessary because simply using two state-of-the-art networks for detection and segmentation is inefficient and time-consuming.

Moreover, object detection and semantic segmentation are highly related. On the one hand, semantic segmentation used as a multi-task supervision can provide context information and semantic features for object detection [35, 25]. On the other hand, object detection can be used as an object prior knowledge to help improve the performance of semantic segmentation [14, 40, 39].

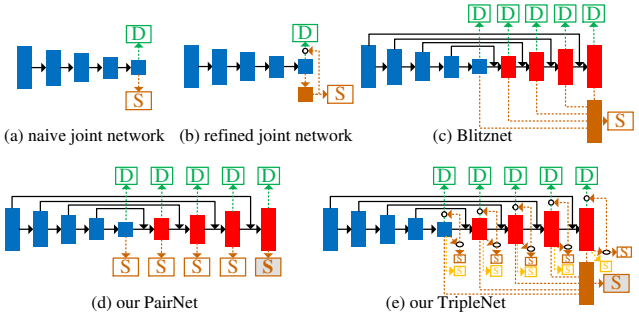


Figure 1. The architectures of joint object detection and semantic segmentation. (a) Two branches for detection and segmentation are added to the last encoder layer [2]. (b) The branch for detection is refined by the branch for segmentation [35, 53]. (c) Each layer of the decoder detects the objects of different scales, and the layer of fusing multi-scale features is for segmentation [8]. (d) Our proposed PairNet. Each layer of the decoder is simultaneously used for detection and segmentation. (e) Our proposed TripleNet, which has three types of supervisions and two light-weight modules. During inference, only segmentation in the gray rectangle is used.

Due to the application requirement and task relevance, joint object detection and semantic segmentation is very necessary, which has attracted the attentions of researchers. Fig. 1 summarizes three typical architectures of joint detection and segmentation. Fig. 1(a) shows the most naive way where two branches of object detection and semantic segmentation are in parallel attached to the last layer of the encoder [2]. In Fig. 1(b), the branch for object detection is further refined by the features from the branch for semantic segmentation [35, 53]. Recently, the encoder-decoder network is further proposed for joint detection and segmentation. In Fig. 1(c), each layer of the decoder is used for multi-scale object detection, and the concatenated

feature maps from the different layers of the decoder are used for semantic segmentation [8]. The above methods have achieved great success on detection and segmentation. However, the performance is still far from the strict demand of real applications. We argue that these methods simply sharing a single network for multi-task learning do not fully exploit the mutual benefit between the two tasks.

To deeply exploit the mutual benefits for joint object detection and semantic segmentation, we propose a new framework called TripleNet in this paper (see Fig. 1(e)). On the one hand, to deeply join the two tasks at different scales, three types of supervisions (i.e., detection-oriented supervision, class-aware/agnostic segmentation supervisions) are imposed on each layer of the decoder network. On the other hand, to further intercross the two related tasks and refine the decoder features, two light-weight modules (i.e., the inner-connected module and attention skip-layer fusion) are incorporated. As a simplified version of TripleNet, PairNet in Fig. 1(d) is also proposed, which only imposes the detection supervision and class-aware segmentation supervision on each layer of the decoder. Finally, the contributions of this paper can be summarized as follows:

(1) Two novel frameworks (i.e., PairNet and TripleNet) are proposed to deeply join detection and segmentation. In TripleNet, the detection-oriented supervision and class-aware/agnostic segmentation supervisions are imposed on each layer of the decoder. Meanwhile, two light-weight modules (i.e., inner-connected module and attention skip-layer fusion) are also incorporated in each decoder layer.

(2) The synergies in both accuracy and speed are gained. On the one hand, both the detection and segmentation accuracies are significantly improved. On the other hand, TripleNet can run at a real-time speed because the improvement is not at expense of the extra computational costs.

(3) Experiments on the VOC 2007, VOC 2012, and COCO datasets [9, 29] demonstrate the effectiveness and efficiency of the proposed TripleNet. For example, using a single network, it achieves 81.9% mAP and 83.3% mIoU on VOC 2012 without COCO pretraining at 16 fps.

2. Related works

Object detection It aims to classify and locate objects in an image. Generally, the methods can be divided into two main classes: two-stage methods and one-stage methods.

Two-stage methods firstly extract some candidate object proposals from an image and then classify these proposals into the specific object categories. R-CNN [12] and its variants (e.g., Fast RCNN [11] and Faster RCNN [42]) are the most representative frameworks among the two-stage methods. Based on R-CNN series, researchers have done many improvements [7, 27, 3]. To accelerate detection speed, Dai *et al.* [7] proposed R-FCN which uses position-sensitive feature maps for proposal classification and bounding box regression. To output multi-scale feature maps with strong semantics, Lin *et al.* [27] proposed feature pyramid network (FPN) based on skip-layer connection and top-down pathway. Recently, Cai *et al.* [3] trained a sequence of detectors with increasing IoU thresholds to improve detection quality.

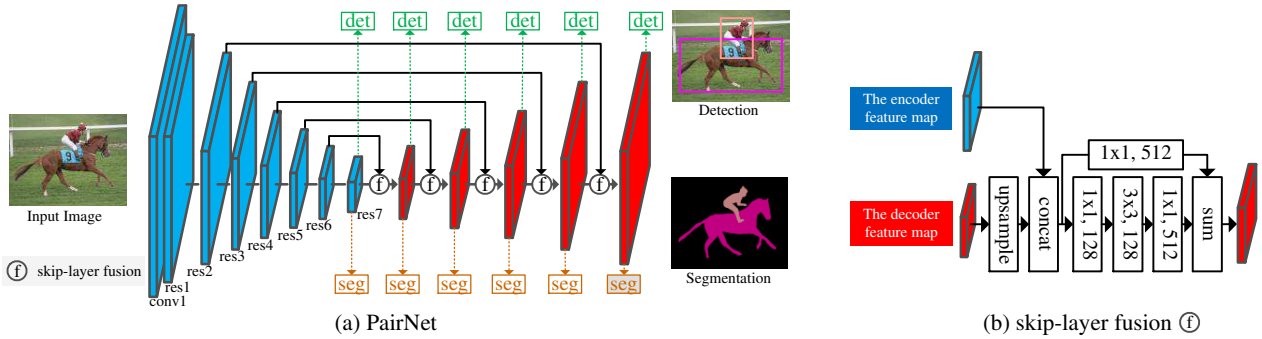
One-stage methods directly predict object class and bounding box in a single network. YOLO [41] and SSD [31] are two earliest one-stage methods. After that, many variants were proposed [10, 23, 44, 55, 37]. DSSD [10] and RON [23] use the encoder-decoder network to add context information for multi-scale object detection. To train the object detector from scratch, DSOD [44] uses the dense layer-wise connections on SSD for deep supervision. Instead of using the in-network feature maps of different resolutions for multi-scale object detection, STDN [55] uses scale-transferable module to generate the different high-resolution feature maps based on the last layer. To solve class imbalance, RetinaNet [28] introduces focal loss to down-weight the contribution of easy samples.

Semantic segmentation It aims to predict the semantic label of each pixel in an image, which has achieved the significant progress with Fully Convolutional Networks (i.e., FCN [34]). Generally, the methods of semantic segmentation can also be divided into two main classes: encoder-decoder methods and spatial pyramid methods.

The encoder-decoder methods contain two subnetworks: an encoder subnetwork and a decoder subnetwork. The encoder subnetwork extracts strong semantic features and reduces spatial resolution of feature maps, which is usually based on the deep CNN models (e.g., VGG [45], ResNet [16], DenseNet [18]) pre-trained on ImageNet [43]. The decoder subnetwork gradually upsamples the feature maps of encoder subnetwork. For example, DeconvNet [36] and SegNet [1] use max-pooling indexes to upsample the feature maps. To extract context information, some methods [38, 26, 50] adopt the skip-layer connection to combine the feature maps of the encoder and decoder subnetworks.

Spatial pyramid methods adopt the idea of spatial pyramid pooling [15] to extract multi-scale information from the feature maps of last layer. Chen *et al.* [5, 6, 47, 48] proposed to use multiple convolutional layers of different atrous rates in parallel (called ASPP) to extract multi-scale features. Instead of using the convolutional layers of different atrous rates, Zhao *et al.* [54] proposed pyramid pooling module (called PSPnet), which downsamples and upsamples the feature maps in parallel. Yang *et al.* [48] proposed to use dense connection to densely cover object scale range.

Joint object detection and semantic segmentation It aims to simultaneously detect objects and predict pixel semantic labels by a single network. Recently, researchers have made some attempts. Yao *et al.* [49] proposed to use the graphical model to understand holistic scene. Teichmann *et al.* [46] proposed to join object detection and



(a) PairNet (b) skip-layer fusion ⓕ

Figure 2. The proposed PairNet for joint object detection and semantic segmentation. (a) The detailed architecture of PairNet. Each layer of the decoder is simultaneously used for detection and segmentation. (b) The skip-layer fusion used in PairNet. Please note only the segmentation in gray rectangle is used for final segmentation during inference.

semantic segmentation by sharing the encoder subnetwork. Kokkinos [21] proposed to integrate multiple vision tasks together. In [35, 2], it was found that joint semantic segmentation and pedestrian detection can help detection. Meanwhile, joint instance semantic segmentation and object detection is proposed [40, 39]. Dvornik *et al.* [8] proposed a real-time framework (BlitzNet) for joint object detection and semantic segmentation. Compared with BlitzNet, the proposed method deeply joins the two tasks by triple supervisions (detection supervision and class-aware/agnostic segmentation supervisions) at each decoder layer.

Recently, panoptic segmentation [19] is further proposed, which is indeed complete. But the methods (e.g., Mask-RCNN [14] and Panoptic FPN [20]) usually have a heavy ROI head and tedious post-processing, which are relatively time-consuming. Joint detection and segmentation is relatively simple, which is enough and significant for some real-time applications (e.g., car detection and road segmentation in autonomous driving). Thus, in this paper more explorations on joint object detection and semantic segmentation are given.

3. The proposed methods

In recent years, Fully Convolutional Networks (FCN [34]) with encoder-decoder structure have achieved the immense success on object detection [28, 10] and semantic segmentation [1]. For example, DSSD [10, 40] and RetinaNet [28] use the different layers of the decoder to detect the objects of different scales. By using the encoder-decoder structure, SegNet [1] and LargeKernel [38] generate the high-resolution logits for semantic segmentation. Based on the above observations, a very natural and simple idea is that FCN with the encoder-decoder is suitable for joint object detection and semantic segmentation.

Based on FCN with the encoder-decoder, the paired supervision decoder network (i.e., PairNet) and triply supervised decoder network (i.e., TripleNet) for joint object detection and semantic segmentation are proposed.

3.1. Paired supervision decoder network (PairNet)

Based on the encoder-decoder structure, a very simple PairNet is firstly proposed. The supervisions of detection and segmentation are added at each decoder layer to guide feature learning at different scales. On the one hand, PairNet uses different layers of the decoder to detect objects of different scales. On the other hand, instead of using the last high-resolution layer for semantic segmentation which is adopted by most state-of-the-art methods [1, 38], PairNet uses each layer of the decoder to respectively parse pixel semantic labels. Though the proposed PairNet is very simple and naive, it has not been explored for joint object detection and semantic segmentation to the best of our knowledge.

Fig. 2(a) gives the detailed architecture of PairNet. The input image firstly goes through a fully convolutional network with encoder-decoder structure. The encoder gradually down-samples the feature maps. In this paper, ResNet [16] (i.e., res1-res4) and some new added residual blocks (i.e., res5-res7) construct the encoder. The decoder gradually maps the low-resolution feature maps to the high-resolution feature maps. To enhance context information, skip-layer fusion is used to combine the feature maps from the decoder and the corresponding feature maps from the encoder. Fig. 2(b) gives the illustration of skip-layer fusion. The feature maps in the decoder is firstly upsampled by bilinear interpolation and then concatenated with the corresponding feature maps of the same resolution in the encoder. After that, the concatenated feature maps go through a residual unit to generate the output feature maps.

To join object detection and semantic segmentation, each layer of the decoder is split into two different branches. The branch of object detection consists of a 3×3 convolutional layer and two sibling 1×1 convolutional layers for object classification and bounding box regression. The branches of object detection at different layers are used to detect objects of different scales. Specifically, the branch at front

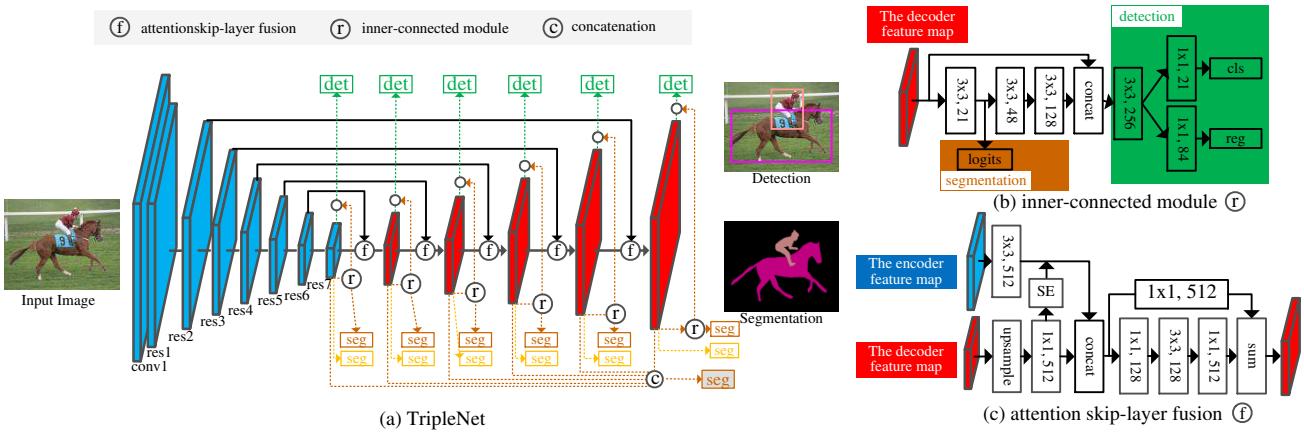


Figure 3. The proposed TripleNet for joint object detection and semantic segmentation. (a) The detailed architecture of TripleNet. (b) inner-connected module. (c) attention skip-layer fusion. The yellow rectangle means the class-agnostic segmentation supervision, and the segmentation in gray rectangle is for final segmentation during inference.

layer of the decoder with low-resolution maps is used to detect large-scale objects, while the branch at latter layer with high-resolution maps is to detect small-scale objects.

The branch of semantic segmentation consists of a 3×3 convolutional layer to generate the logits. There are two different ways to generate the logits. The first is that the segmentation logits are upsampled to the same resolutions of the ground-truths, and the second is that the ground-truths are downsampled to the same resolutions of the logits. We find that the first strategy has a slightly better performance, which is adopted as follows.

3.2. Triply supervised decoder network (TripleNet)

To deeply exploit the mutual benefits between the two tasks, triply supervision decoder network (called TripleNet) is further proposed. On the one hand, to deeply join two tasks at different scales, detection-oriented supervision, class-aware/agnostic segmentation supervisions are added on each layer of the decoder. On the other hand, to further intercross the two tasks and refine the features in each decoder layer, the inner-connected module and attention skip-layer fusion are proposed. Fig. 3(a) gives the detailed architecture of TripleNet. Compared to PairNet, TripleNet adds some new modules as follows.

Multi-scale fused segmentation In [8, 54, 5], it has been demonstrated that multi-scale features can extract context information for semantic segmentation. To use multi-scale features for better semantic segmentation, the feature maps of different layers in the decoder go through a 3×3 convolutional layer for channel reduction. The output feature maps are upsampled to the same spatial resolution and concatenated together. After that, a 3×3 convolutional layer is used to generate the segmentation logits. Compared to the features from one layer of the decoder, multi-scale fused features make better use of context information. Thus, multi-scale fused segmentation is used for final prediction during inference. Meanwhile, the semantic segmentation on each layer of the decoder can be seen as a deep supervision for feature learning.

The inner-connected module PairNet only shares each layer of the decoder for detection and segmentation, while the branches of detection and segmentation in each layer of PairNet have no cross. To further intercross the two tasks, an inner-connected module is proposed. Fig. 3(b) shows the inner-connected module in layer i. The feature map in layer i first goes through a 3×3 convolutional layer to produce the segmentation logits for the branch of semantic segmentation. Meanwhile, the segmentation logits further go through two 3×3 convolutional layers to generate the new feature maps which are concatenated with the feature maps in layer i. Based on the concatenated feature maps, a 3×3 convolutional layer is used to generate the feature maps for the branch of object detection.

Class-agnostic segmentation supervision Semantic segmentation mentioned above is class-aware, which aims to simultaneously identify specific object categories and the background. We argue that class-aware semantic segmentation may ignore the discrimination between objects and the background. Therefore, class-agnostic segmentation supervision is further added to each layer of the decoder. Specifically, a 3×3 convolutional layer is used to generate the logits of class-agnostic semantic segmentation. To generate the ground-truths, all the objects are set as one category, and the background is set as another category.

Attention skip-layer fusion In Section 3.1, PairNet simply fuses the feature maps of the decoder and the encoder. Generally, the features from the layer of the encoder have relatively low-level semantic, and that from the layer of the decoder have relatively high-level semantic. To enhance informative features and suppress less useful features from the

Method	backbone	input size	det	seg_fine	seg_all	MFS	IC	CAS	ASF	mAP	mIoU
(a) only detection	ResNet50	300×300	✓							78.0	N/A
(b) only segmentation with fine layer	ResNet50	300×300		✓						N/A	72.5
(c) only segmentation with all layers	ResNet50	300×300			✓					N/A	72.9
(d) PairNet	ResNet50	300×300	✓		✓					78.9	73.1
(e) add MFS	ResNet50	300×300	✓		✓	✓				79.0	73.5
(f) add MFS and IC	ResNet50	300×300	✓		✓	✓	✓			79.5	73.6
(g) add MFS, IC, and CAS	ResNet50	300×300	✓		✓	✓	✓	✓		79.7	74.4
(h) TripleNet	ResNet50	300×300	✓		✓	✓	✓	✓	✓	80.0	74.8
(i) TripleNet	ResNet50	512×512	✓		✓	✓	✓	✓	✓	83.2	77.3
(j) TripleNet†	ResNet50	512×512	✓		✓	✓	✓	✓	✓	87.3	79.6
(k) TripleNet†	ResNet101	512×512	✓		✓	✓	✓	✓	✓	**88.1**	**80.4**

Table 1. Ablation experiments of PairNet and TripleNet on the `VOC2012-val-seg` set. The backbone is ResNet50 or ResNet 101 [16], and the input image is of 300×300 or 512×512. † means that VOC 2007 is use for training. "MFS" means multi-scale fused segmentation, "IC" means inner-connected module, "CAS" means class-agnostic segmentation, and "ASF" means attention skip-layer fusion.

encoder by the decoder, Squeeze-and-Excitation (SE) [17] block in Fig. 3(c) is incorporated. The input of a SE block is the feature maps from the decoder, and the output of SE block is used to scale the feature maps from the encoder. After that, the feature maps from the encoder and decoder layers are concatenated for skip-layer fusion as PairNet.

4. Experiments

4.1. Datasets and evaluation

To demonstrate the effectiveness and efficiency of proposed methods and compare with some state-of-the-art methods, the experiments on the famous VOC 2007, VOC 2012 [9], and COCO [29] datasets are conducted

The PASCAL VOC challenge [9] has been held annually since 2006 and consists of three principal challenges (i.e., image classification, object detection, and semantic segmentation). Among these annual challenges, the VOC 2007 and VOC 2012 datasets are usually used to evaluate the performance of object detection and semantic segmentation, which have 20 object categories. The VOC 2007 dataset contains 5011 `trainval` images and 4952 `test` images. The VOC 2012 dataset is split into three subsets (i.e., `train`, `val`, and `test`). The `train` set contains 5717 images for detection and 1464 images for segmentation (called `VOC12-train-seg`). The `val` set contains 5823 images for detection and 1449 images for segmentation (called `VOC12-val-seg`). The test set contains 10991 images for detection and 1456 for segmentation. To enlarge the training data for semantic segmentation, the augmented segmentation data provided by [13] is used, which contains 10582 training images (called `VOC12-trainaug-seg`).

The COCO benchmark [29] is a large-scale dataset for object detection, instance segmentation, and image caption, which has 80 object categories. It is usually split into three subsets (i.e., `trainval35k`, `minival`, and `test`). The `trainval35k` set contains about $115k$ images, the `minival` set contains $5k$ images, and the `test` set contains about $20k$ images. Usually, the `trainval35k` set is used for training, the `minival` set is for validation experiments, and the `test` set is for comparison with state-of-the-art methods. For semantic segmentation, the ground-truths are generated by assigning the objects which belong to the same category with the same semantic label.

Evaluation metric For object detection, mean average precision (i.e., mAP) is used for performance evaluation. On the VOC datasets, mAP is calculated under the IoU threshold of 0.5. On the COCO benchmark, mAP is calculated by averaged over the IoU thresholds of 05:0.95. For semantic segmentation, mean intersection over union (i.e., mIoU) is used for performance evaluation.

4.2. Ablation experiments on the VOC 2012 dataset

In this subsection, experiments are conducted on the PASCAL VOC 2012 to validate the effectiveness of proposed method. The set of `VOC12-trainaug-seg` is used for training and the set of `VOC12-val-seg` is for performance evaluation. The input images are rescaled to the size of 300×300, and the size of a mini-batch is 32. The total number of iteration in the training stage is $40k$, and the initial learning rate is 0.0001. The learning rate decreases by a factor of 10 at the $25k$ and $35k$ iterations.

PairNet The top part of Table 1 shows the ablation experiments of PairNet. When the different layers of the decoder are only used for multi-scale object detection (i.e., Table 1(a)), mAP of object detection is 78.0%. When the fine layer of the decoder is used for semantic segmentation (i.e., Table 1(b)), mIoU of semantic segmentation is 72.5%. When all the different layers of the decoder are used for joint object detection and semantic segmentation (i.e., Table 1(d)), mAP and mIoU of PairNet are 78.9% and 73.1%. Namely, PairNet can improve the performance of both detection and segmentation, which indicates that joint detection and segmentation on each layer of the decoder is useful.

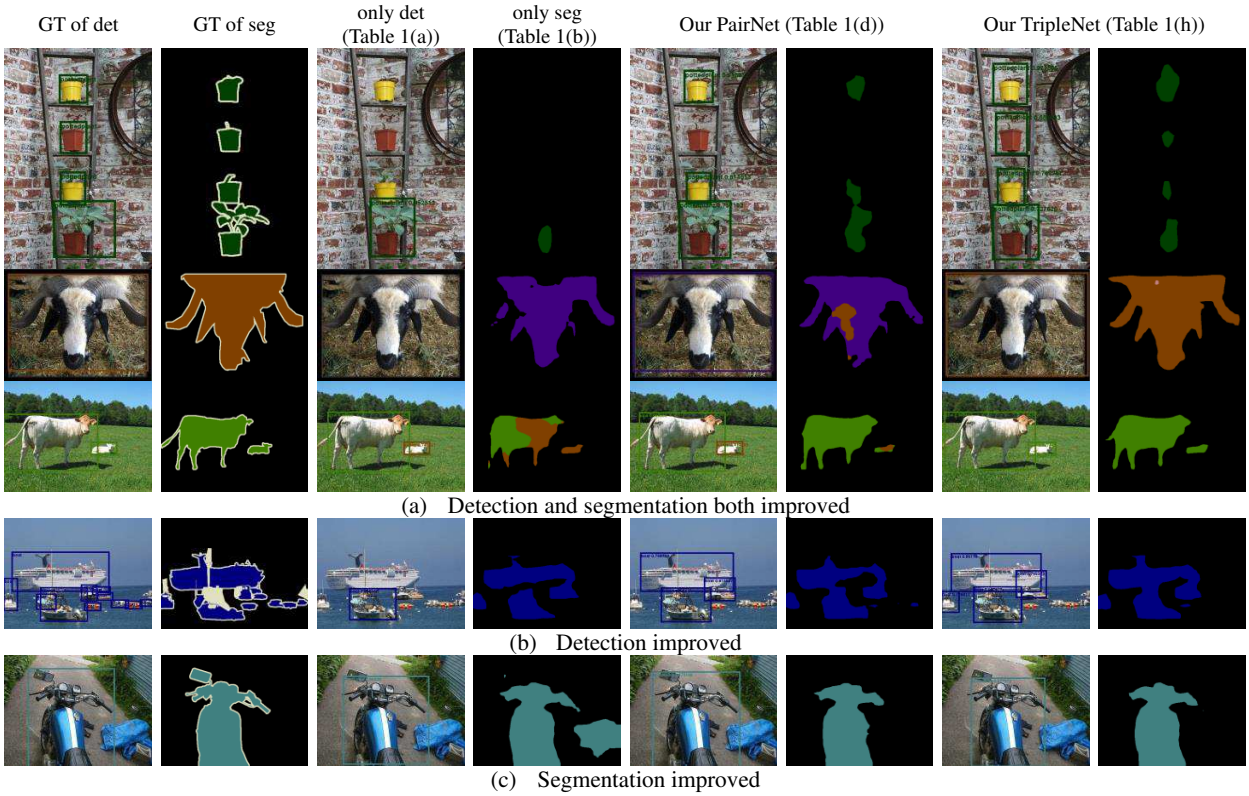


Figure 4. Detection or segmentation results of the methods in Table 1 (i.e., "only det", "only seg", PairNet, and TripleNet). (a) demonstrates that detection and segmentation can be both improved by PairNet and TripleNet. (b) demonstrates that detection is mainly improved by PairNet or TripleNet. (c) demonstrates that segmentation is mainly improved by PairNet or TripleNet.

Meanwhile, the method using all the different layers of the decoder for segmentation (i.e., Table 1(c)) has better performance than the method using only the last layer of the decoder for segmentation (i.e., Table 1(b)). The reason is that using all the layers of the decoder for semantic segmentation provides a deeper supervision for feature learning.

TripleNet The medium part of Table 1 shows the ablation experiments of TripleNet. Based on PairNet, TripleNet adds four modules (i.e., MFS, IC, CAS, and ASF). When adding the MFS module, TripleNet outperforms PairNet by 0.1% on detection and 0.4% on segmentation. When adding the MFS and IC modules, TripleNet outperforms PairNet by 0.6% on detection and 0.5% on segmentation. When adding all four modules, TripleNet has the best performance on both detection and performance, which outperforms the baselines (i.e., Table 1(a) and (b)) by 2.0% on object detection and 2.3% on semantic segmentation.

Larger input and more data The bottom part of Table 1 shows the affects of larger input size and more training data. It can be seen that with the larger input size (i.e., 512×512) and more training data (i.e., adding the VOC 2007 dataset), TripleNet can further improve performance of object detection and semantic segmenation. For example, TripleNet512 with more extra training data achieves 88.1% mAP on object detection and 80.4% mIoU on semantic segmentation.

Qualitative results Fig. 4 shows the detection and segmentation results of some methods in Table 1. The first two columns are the ground-truths of object detection and semantic segmentation. The results of only detection (i.e., "only det" in Table 1(a)) and only segmentation (i.e., "only seg" in Table 1(b)) are shown in the third and fourth columns. The results of PairNet in Table 1(d) and TripleNet in Table 1(h) are shown in fifth to eighth columns. In Fig. 4(a), the examples of detection and segmentation both improved by PairNet or TripleNet are given. For example, in the first row, "only det" and "only seg" both miss three potted plant, while PairNet only misses one potted plant and TripleNet does not miss any potted plant. In Fig. 4(b), the example of object detection improved is shown. "only detect" can only detect one ship, PairNet can detect three ships, and TripleNet can detect four ships. In Fig. 4(c), the example of semantic segmentation improved is shown. "only seg" recognizes the blue bag as a motorbike, while PairNet and TripleNet correctly classify the blue bag into the background.

Feature visualization Fig. 5 further visualizes the feature maps of the methods in Table 1 (i.e., "only det", "only seg", and TripleNet). On object detection and semantic seg-

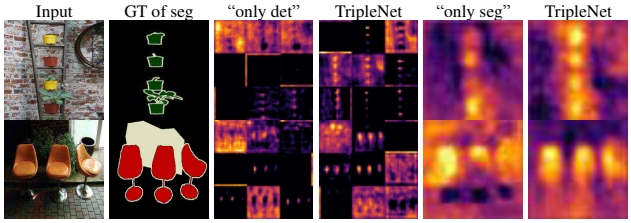


Figure 5. Feature visualization of the methods in Table 1 (i.e., "only det", "only seg", and TripleNet). For detection, the first 9 feature maps of maximum response values where objects are detected are selected. For segmentation, the feature map of corresponding object category is shown.

method	backbone	input size	mAP	mIoU
baseline	ResNet50	300 × 300	78.0	72.5
TripleNet	ResNet50	300 × 300	80.0	**74.8**
TripleNet-SS	ResNet50	300 × 300	**80.2**	73.9

Table 2. TripleNet *vs* TripleNet-SS. TripleNet-ss means that scale-specific semantic segmentation is used in each layer, which is similar to detection.

method	backbone	input size	mAP
SSD512 [31]	VGG16	512 × 512	79.5
DES512 [53]	VGG16	512 × 512	81.7
DSSD513 [10]	ResNet101	512 × 512	81.5
STDN513 [55]	DenseNet169	512 × 512	80.9
BlitzNet512 [8]	ResNet50	512 × 512	81.5
RefineDet512 [52]	VGG16	512 × 512	81.8
RFBNet512 [30]	VGG16	512 × 512	82.2
DFPR512 [22]	ResNet101	512 × 512	82.4
TripleNet512	ResNet50	512 × 512	82.4
TripleNet512	ResNet101	512 × 512	**83.0**

Table 3. Detection results of some single-stage and state-of-the-art methods on the VOC 2007 `test` set without using COCO pre-training. All the methods are based on single scale test.

mentation, the feature maps generated by TripleNet are less noisy. We argue that the deep supervisions make the learned features more robust.

Scale-specific semantic supervision Another possible strategy of semantic segmentation supervision in each layer is that each layer only focuses on the scale-specific objects which is similar to object detection (called TripleNet-SS). Table 2 compares TripleNet-SS and TripleNet. Compared with TripleNet, TripleNet-SS achieves the better detection performance but worse segmentation performance. It means that scale-specific semantic segmentation in each layer is much more useful on object detection, while scale-unspecific semantic segmentation in each layer is much more useful on semantic segmentation. Thus, how to combine the advantages of them to further improve joint detection and segmentation is an open problem and a future work. In the following experiments, TripleNet with scale-unspecific semantic segmentation is adapted.

method	backbone	mAP	mIoU	speed (fps)
SSD512 [31]	VGG	79.4	N/A	19
RON384 [23]	VGG16	75.4	N/A	15
DSSD513 [10]	ResNet101	80.0	N/A	5
DES512 [53]	VGG16	80.3	N/A	31
RefineDet512 [52]	ResNet101	80.0	N/A	24
DFPR512 [22]	ResNet101	81.1	N/A	16
FCN [34]	VGG16	N/A	62.2	16
ParseNet [32]	VGG16	N/A	69.8	-
Deeplab [5]	VGG16	N/A	71.6	8
DPN [33]	VGG-like	N/A	74.1	-
RefineNet [26]	ResNet101	N/A	82.4	-
PSPNet [54]	ResNet101	N/A	82.6	6
DFN [50]	ResNet101	N/A	82.7	-
EncNet [51]	ResNet101	N/A	82.9	-
BlitzNet512 [8]	ResNet50	79.0	75.6	19
TripleNet512	ResNet50	80.3	82.4	16
TripleNet512	ResNet101	**81.9**	**83.3**	14

Table 4. Results of object detection (mAP) and semantic segmentation (mIoU) on the VOC 2012 `test` set without using COCO pre-training. Inference time is tested on NVIDIA TitanX.

4.3. Comparison with state-of-the-art methods on the VOC 2007/2012 test dataset

In this section, TripleNet is compared with some state-of-the-art methods on the VOC 2007/2012 datasets [9]. Among these methods, SSD [31], RON [23], DSSD [10], DES [53], STDN [55], RefineDet [52], and DFPR [22] are only for object detection. FCN [34], ParseNet [32], Deeplab [5], DPN [33], RefineNet [26], PSPNet [54], DFN [50], and EncNet [51] are only for semantic segmentation. BlitzNet [8] is for joint detection and segmentation.

VOC 2007 The proposed TripleNet is firstly compared with these methods on the VOC 2007 dataset. The training data is VOC 2007 trainval set and VOC 2012 trainval set. The total number of iterations is $75k$, and the initial learning rate is 0.0001. The learning rate decreases at the $45k$ and $60k$ iterations by a factor of 10. As only the ground-truth of object detection is provided, these methods are only evaluated on object detection. Table 3 shows mAP of these methods. TripleNet with the relatively small ResNet50 has already the same performance as DFPR512 [22] with the deep ResNet101. With the deep ResNet101, TripleNet has 83.0% mAP, which outperforms all the state-of-the-art methods.

VOC 2012 Then, the proposed method is compared with these methods on the VOC 2012 dataset. The training data is VOC 2007 dataset and VOC 2012 trainval set. The other parameter settings are the same as that used in the VOC 2007 dataset. Table 4 shows detection or segmentation results. It can be concluded as follows: (1) By using a single network, the proposed TripleNet achieves the state-of-the-art performance on both object detection and semantic seg-

Method	backbone	input size	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
SSD512 [31]	VGG16	512×512	28.8	48.5	30.3	10.9	31.8	43.5
DSSD512 [10]	ResNet101	512×512	33.2	53.3	35.2	13.0	35.4	51.1
STDN512 [55]	VGG16	512×512	31.8	51.0	33.6	14.4	36.1	43.4
DES512 [53]	VGG16	512×512	32.8	53.2	34.6	13.9	36.0	47.6
RetinaNet [28]	ResNet101	500×500	34.4	53.1	36.8	14.7	38.5	49.1
RefineDet512 [52]	ResNet101	512×512	36.4	57.5	39.5	16.6	**39.9**	51.4
RFBNet512 [30]	VGG16	512×512	34.4	55.7	36.4	17.6	37.0	47.6
DFPR512 [22]	ResNet101	512×512	34.6	54.3	37.3	14.7	38.1	51.9
TripleNet512	ResNet50	512×512	35.9	57.8	38.0	17.7	37.2	50.7
TripleNet512	ResNet101	512×512	**37.4**	**59.3**	**39.6**	**18.5**	39.0	**52.7**

Table 5. Detection results of single-stage methods on COCO `test-dev` set. All the methods are based on single-scale test.

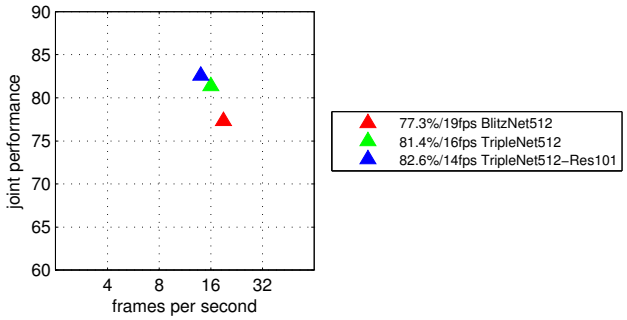


Figure 6. Joint performance & speed on the VOC 2012 `test` set.

Method	backbone	AP	AP_{50}	AP_{75}	mIoU
BlitzNet512 [8]	ResNet50	34.1	55.1	35.9	53.5
TripleNet512	ResNet50	36.0	57.7	37.8	58.3
TripleNet512	ResNet101	**37.1**	**58.7**	**39.4**	**59.6**

Table 6. Comparison of joint object detection and semantic segmentation on the COCO `minival` test.

mentation. (2) The proposed TripleNet and BlitzNet both output the results of object detection and semantic segmentation. TripleNet outperforms BlitzNet by 2.9% on object detection and 7.3% on semantic segmentation.

Speed Based on Table 4, it can also be seen that TripleNet can simultaneously output detection and segmentation results without much additional computation cost. Specifically, TripleNet runs at a real-time speed (14-16fps). Fig. 6 further plots joint performance and speed on the VOC 2012 `test` set. The joint performance is defined as the averaged performance on object detection and semantic segmentation. Compared with BlitzNet, TripleNet has higher joint performance with little extra computation cost.

4.4. Experiments on the COCO benchmark

In this section, experiments on the challenging COCO benchmark [29] are further conducted. Specifically, the training data is the `trainval35k` set. The input images are of 512×512. The total number of iterations is $700k$, and the initial learning rate is 0.0001. The learning rate decreases at the $400k$ and $550k$ iterations by a factor of 10. Table 5 compares detection results of some single-stage and state-of-the-art methods on the COCO `test-dev` set. TripleNet512 with ResNet101 [16] achieves 37.4% mAP, which outperforms all the other state-of-the-art methods. For example, based on the deep ResNet101 [16], TripleNet512 outperforms DFPR512 [22] by 2.8% and RFBNet512 [30] by 3.0%.

Meanwhile, TripleNet is compared with BlitzNet [8] on the COCO `minival` set for joint object detection and semantic segmentation in Table 6. Based on the relatively small ResNet50 [16], TripleNet512 has 36.0% mAP on detection and 58.3% mIoU on segmentation, while BlitzNet [8] has 34.1% AP on etection and 53.5% mIoU on segmentation. Namely, the proposed TripleNet outperforms BlitzNet [8] by 1.9% on object detection and 4.8% on semantic segmentation. With the much deeper ResNet101, TripleNet can further improve both detection performance and segmentation performance.

5. Conclusion

In this paper, we proposed two fully convolutional networks (i.e., PairNet and TripleNet) for joint object detection and semantic segmentation. PairNet simultaneously predicts the objects of different scales by different layers and parses pixel semantic labels by all different layers. To further exploit the mutual benefits between detection and segmentation, TripleNet adds four modules (i.e, multi-scale fused segmentation, inner-connected module, class-agnostic segmentation supervision, and attention skip-layer fusion) to PairNet. Experiments demonstrate that TripleNet can achieve state-of-the-art performance on both object detection and semantic segmentation at a real-time speed.

6. Acknowledgments

It was supported by National Natural Science Foundation of China (No. 61632018), Postdoctoral Program for Innovative Talents (No. BX20180214), China Postdoctoral Science Foundation (No. 2018M641647), and the Nokia.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[2] G. Brazil, X. Yin, and X. Liu. Illuminating pedestrians via simultaneous detection & segmentation. *Proc. IEEE International Conference on Computer Vision*, 2017.

[3] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] J. Cao, Y. Pang, and X. Li. Learning multilayer channel features for pedestrian detection. *IEEE Transactions on Image Processing*, 26(7):3210–3220, 2017.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[6] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017.

[7] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. *Proc. Advances in Neural Information Processing Systems*, 2016.

[8] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid. Blitznet: A real-time deep network for scene understanding. *Proc. IEEE International Conference on Computer Vision*, 2017.

[9] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[10] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. *arXiv:1701.06659*, 2017.

[11] R. Girshick. Fast r-cnn. *Proc. IEEE International Conference on Computer Vision*, 2015.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[13] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. *Proc. IEEE International Conference on Computer Vision*, 2011.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *Proc. IEEE International Conference on Computer Vision*, 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *Proc. European Conference on Computer Vision*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proc. IEEE International Conference on Computer Vision*, 2016.

[17] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[19] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. *arXiv:1801.00868*, 2018.

[20] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic feature pyramid networks. *arXiv:1901.02446*, 2019.

[21] I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[22] T. Kong, F. Sun, W. Huang, and H. Liu. Deep feature pyramid reconguration for object detection. *Proc. European Conference on Computer Vision*, 2018.

[23] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen. Ron: Reverse connection with objectness prior networks for object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Proc. Advances in Neural Information Processing Systems*, 2012.

[25] C. Lin, J. Lu, G. Wang, and J. Zhou. Graininess-aware deep feature learning for pedestrian detection. *Proc. European Conference on Computer Vision*, 2018.

[26] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *Proc. IEEE International Conference on Computer Vision*, 2017.

[29] T.-Y. Lin, M. Maire, S. Belongie, P. P. J. Hays, P. D. D. Ramanan, and C. L. Zitnick. Microsoft coco: Common objects in context. *Proc. European Conference on Computer Vision*, 2014.

[30] S. Liu, D. Huang, and Y. Wang. Receptive field block net for accurate and fast object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. *Proc. European Conference on Computer Vision*, 2016.

[32] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *Proc. International Conference on Learning Representations*, 2016.

[33] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. *Proc. IEEE International Conference on Computer Vision*, 2015.

[34] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[35] J. Mao, T. Xiao, Y. Jiang, and Z. Cao. What can help pedestrian detection? *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[36] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. *Proc. IEEE International Conference on Computer Vision*, 2015.

[37] Y. Pang, T. Wang, R. M. Anwer, F. S. Khan, and L. Shao. Efficient featurized image pyramid network for single shot detector. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[38] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters–improve semantic segmentation by global convolutional network. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[39] V.-Q. Pham, S. Ito, and T. Kozakaya. Biseg: Simultaneous instance segmentation and semantic segmentation with fully convolutional networks. *Proc. British Machine Vision Conference*, 2017.

[40] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. *Proc. European Conference on Computer Vision*, 2016.

[41] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[42] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proc. Advances in Neural Information Processing Systems*, 2015.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

[44] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue. Dsod: Learning deeply supervised object detectors from scratch. *Proc. IEEE International Conference on Computer Vision*, 2017.

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[46] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. *arXiv:1612.07695*, 2016.

[47] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. *Proc. IEEE Winter Conference on Applications of Computer Vision*, 2017.

[48] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[49] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[50] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Learning a discriminative feature network for semantic segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[51] H. Zhang, K. Dana, J. Shi, and Z. Zhang. Context encoding for semantic segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[52] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Single-shot refinement neural network for object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[53] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang, and A. L. Yuille. Single-shot object detection with enriched semantics. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[54] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[55] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu. Scale-transferrable object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.