CONTEXT-AWARE CASCADE NETWORK FOR SEMANTIC LABELING IN VHR IMAGE

Yongcheng Liu[1,2], *Bin Fan*[1], *Lingfeng Wang*[1], *Jun Bai*[3], *Shiming Xiang*[1], *Chunhong Pan*[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2] School of Computer and Control Engineering, University of Chinese Academy of Sciences
[3] Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences

ABSTRACT

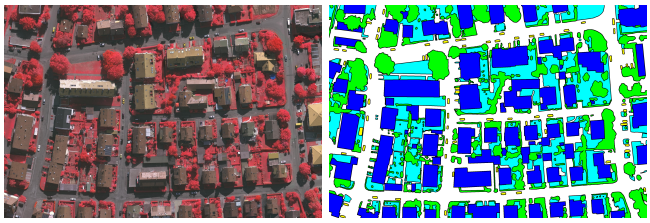
Semantic labeling for the very high resolution (VHR) image of urban areas is challenging, because of many complex manmade objects with different materials and fine-structured objects located together. Under the framework of convolutional neural networks (CNNs), this paper proposes a novel end-to-end network for semantic labeling. Specifically, our network not only improves the labeling accuracy of complex manmade objects by aggregating multiple context semantics with a cascaded architecture, but also refines fine-structured objects by utilizing the low-level detail in shallow layers of CNNs with a hierarchical pyramid structure. Throughout the network, a dedicated residual correction scheme is employed to amend the latent fitting residual. As a result of these specific components, the whole model works in a global-to-local and coarse-to-fine manner. Experimental results show that our network outperforms the state-of-the-art methods on the large-scale *ISPRS Vaihingen 2D Semantic Labeling Challenge* dataset.

Index Terms— Semantic Labeling, Convolutional Neural Networks, Context, Residual Correction, VHR Image

1. INTRODUCTION

Semantic labeling for the VHR image, which is to assign each pixel to a given object class, is a long-standing research problem in image processing, as it plays a vital role in infrastructure planning, territorial planning, urban change detection, and so on. However, as Fig. 1 shows, in urban areas, many manmade object categories are composed of a large number of different materials with similar color and texture. Meanwhile, fine-structured objects in cities (such as cars, trees) are small or threadlike and interact with each other through occlusions and cast shadows. Both result in that semantic labeling for this kind of image poses additional challenge.

Recently, most of the state-of-the-art methods for semantic labeling are developed on deep convolutional neural networks (CNNs). The fully convolutional networks (FCNs), which outputs the class likelihoods for each pixel in an image [1, 2], has boosted the accuracy of semantic labeling a lot than the patch-based approaches [3, 4]. Nevertheless, the feature map output by FCNs is coarse due to sub-sampling, resulting in inaccurate pixelwise labeling results, especially for complex urban images. Some researches try to mitigate the problem of coarse labeling by FCNs, either by utilizing local detail in CNNs' shallow layers [5–8], or by introducing boundary detection to improve localization [9, 10]. However, these methods usually directly perform less effective stack of local detail and require extra boundary supervision.



(a) Image (b) Ground truth

Fig. 1: Illustration of complex surroundings on *ISPRS* dataset. The ground truth contains five classes (white: impervious surface, blue: building roof, cyan: low vegetation, green: tree, yellow: car).

On the other hand, to improve recognition accuracy of various objects, some recent works concentrate on utilizing the context of the image. To exploit multi-view context, [11–15] directly input multi-view images around the objects, but this operation is usually less efficient. Another way is to acquire context by CNNs, such as dilated convolution [16, 17], spp-net [18], multi-kernel convolution [19] and multi-stage features fusion [20,21]. However, their stack-fusion strategies of multi-context ignore the hierarchical dependencies among the objects and scenes in contexts of different scales.

In this paper, we propose a novel context-aware cascade network as shown in Fig. 2. The aim of this work is to further advance semantic labeling in VHR image by focusing on three key aspects: context aggregating, fine-structured objects refinement and multi-feature fusion. Our contributions can be highlighted as follows: (1) A multi-context aggregating cascade model is proposed to capture global and local context with hierarchical dependencies. (2) An effective refinement model is proposed to refine the labeling map progressively, especially for fine-structured objects. (3) A residual correction scheme is proposed to amend the latent fitting residual in multi-feature fusion. (4) All the proposed models are embed-

This work was supported by the National Natural Science Foundation of China under Grants 91646207, 61403375, 61573352, 61403376, 91338202 and 91438105.

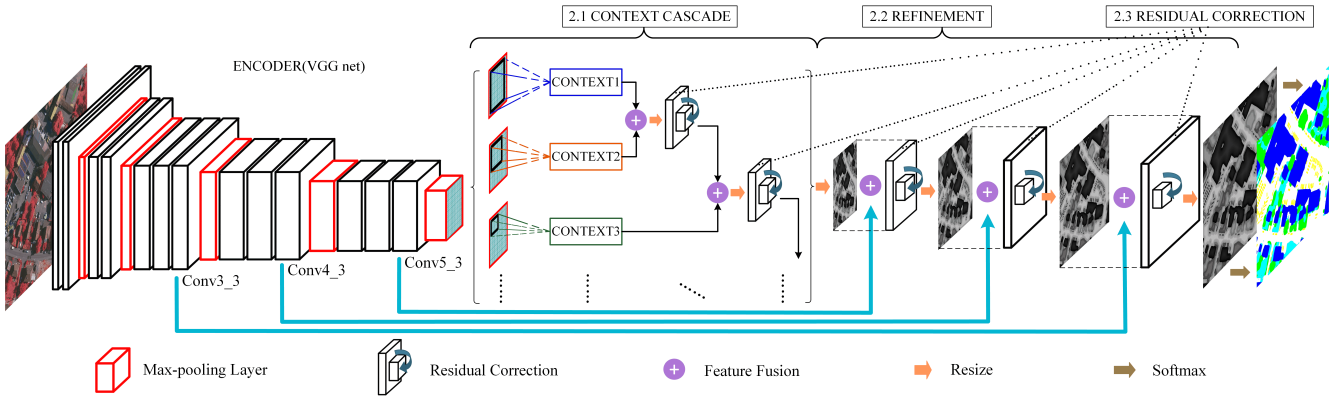


Fig. 2: The architecture of the proposed semantic labeling network for VHR image.

ded collaboratively in an end-to-end network, which achieves the state of the art on the large-scale *ISPRS Vaihingen 2D Semantic Labeling Challenge* dataset.

2. CONTEXT-AWARE CASCADE NETWORK

Since there are many complex scenes, semantic labeling for the image of urban areas is particularly difficult. For such a challenging task, we propose a context-aware cascade network (CAC-NET) as shown in Fig. 2, whose four aspects are described in the following: *1) Multi-context aggregating cascade*, *2) Fine-structured objects refinement*, *3) Residual correction* and *4) Network training and labeling*.

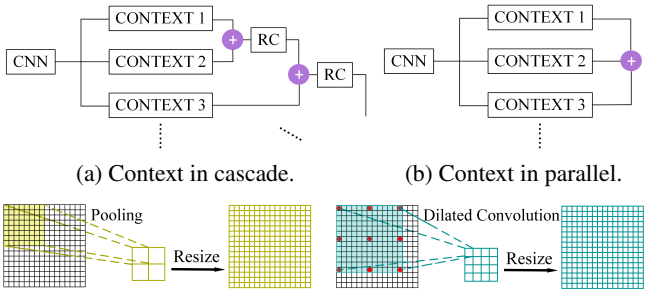
2.1. Multi-context aggregating cascade

Context is a critical factor to recognize complex manmade objects. Since deeper layers in CNNs contain wider (larger *receptive field* on the input image) and stronger (higher nonlinearity) semantics, the context acquired from deep layers can capture strong semantic information. Meanwhile, multi-context can capture hierarchical dependencies. To aggregate multi-context with well-kept hierarchical dependencies, a novel cascade model is proposed, as shown in Fig. 3(a).

In our cascade model, we take multi-context by performing multi-scale unit operation on the last layer of CNNs, as shown in Fig. 2. Fig. 3(c) illustrates two different ways to perform unit operation, pooling and dilated convolution [17], where *pooling_size* equals 8 and *dilation_rate* equals 6. Multi-scale unit operation corresponds to multi-size regions on the last layer of CNNs. large region (high-level context) contains wide information and small region (low-level context) otherwise. To hold the hierarchical dependencies in multi-level context, we aggregate them in a cascade manner, i.e., high-level context is aggregated first and low-level context next. Formally, it is described as:

$$\begin{cases} T = \mathcal{R}\Big(\cdots\mathcal{R}\Big(\mathcal{R}(T_1 \oplus T_2) \oplus T_3\Big) \oplus \cdots \oplus T_n\Big), \\ s_{T_1} > s_{T_2} > s_{T_3} > \cdots > s_{T_n}. \end{cases} \quad (1)$$

Where $T_1, T_2, \cdots, T_n$ denote n-level context and T is the final aggregated context. s_{T_n} is the scale size of unit operation for context T_n. $\oplus$ denotes fusion. $\mathcal{R}$ denotes the residual correction, which will be described in Subsection 2.3.



(a) Context in cascade. (b) Context in parallel.

(c) Pooling(left) and dilated convolution (right).

Fig. 3: Multi-context aggregating cascade. RC: residual correction.

Our cascade model has two advantages. First, the context is acquired from deep layers in CNNs, which is more efficient than directly using multiple images as input (e.g. ten regions around each object are used in [12]). Second, multi-context is aggregated in a cascade manner. This is more effective than simply operating in a parallel stack [16, 21] as shown in Fig. 3(b), which loses the hierarchical dependencies.

2.2. Fine-structured objects refinement

The coarse output by FCNs-based methods results in much difficulty for precise semantic labeling, especially for fine-structured objects. Shallow layers in CNNs represent much local detail (such as edge and texture) of the image, which could be utilized for refinement of labeling map. Thus, an effective refinement model is proposed, as shown in Fig. 2. Specifically, the coarse output from the encoder is refined progressively with a hierarchical pyramid structure. Fig. 4 illustrates single process of refinement. Formally, it can be described as:

$$M^{i+1} = \mathcal{R}e\Big(\mathcal{R}\Big(\mathbf{L}(M^i \otimes W_{M^i}) \oplus \mathbf{L}(F^i \otimes W_{F^i})\Big)\Big), \quad (2)$$

where M^i is refined map of last process and F^i is the map of this process in shallow layers. W_{M^i} and W_{F^i} are the convolutional weights of M^i and F^i respectively. $\otimes$ denotes convolution and $\oplus$ is fusion. $\mathbf{L}$ is the ReLU activation function. $\mathcal{R}e$ denotes the resize process and $\mathcal{R}$ is the residual correction.

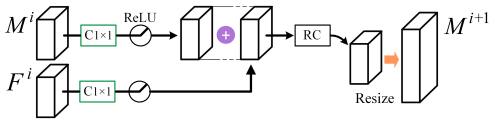


Fig. 4: Single process of refinement. C1×1: convolution with *kernel_size* 1×1, RC: residual correction.

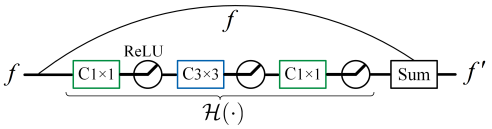


Fig. 5: Residual correction scheme. C1×1: convolution with *kernel_size* 1×1, C3×3: convolution with *kernel_size* 3×3

To fuse finer detail in the next shallow layer, we resize current map to the corresponding higher resolution with bilinear interpolation to generate M^{i+1}.

The hierarchical pyramid structure is beneficial to progressively reintroduce the local detail of multi-stage features in shallow layers. This is more effective than directly stacking these features [6,14], since the latter neglects the semantic gaps in multi-stage features.

The most relevant work with our refinement model is proposed in [8], however, they are quite different. Our model is focused on refinement with the specific properties (e.g., small dataset and complex scenes) of VHR image. Specifically, as shown in Fig. 2, only a few specific shallow layers are chosen considering complexity. Moreover, structurally, they are combined with several dedicatedly designed residual correction schemes, which are described in the following.

2.3. Residual correction

Usually, it is difficult to well fuse multi-level features in CNNs. The reasons are two-fold. First, CNNs have difficulty to directly fit a desired underlying fusion mapping when network deepens. Second, there exists latent fitting residual when directly fusing multi-level features. Thus, a residual correction scheme is proposed, as shown in Fig. 5, to amend fitting residual of multi-level feature fusion inside CAC-NET.

Building on the idea of deep residual learning [22], we explicitly let the stacked layers fit an inverse residual mapping instead of letting them directly fit a desired underlying fusion. Formally, let f denotes fused feature and $f^{'}$ denotes the desired underlying fusion. The stacked layers is expected to fit another mapping of $\mathcal{H}(\cdot) = f^{'} - f$, thus to achieve $f^{'} = f + \mathcal{H}(\cdot)$. Then, the impact of fitting residual can be offset to some extent. As demonstrated in [22], it is easier to fit the inverse residual mapping when network deepens, i.e., the residual can be well amended.

Specifically, the residual correction schemes are dedicatedly positioned in CAC-NET to mitigate the cumulate impact of fitting residual. Thus, CAC-NET can work effectively and efficiently in an end-to-end manner as shown in Fig. 2. Moreover, the skip connection is beneficial to the training of the proposed deep network.

2.4. Network training and labeling

The whole model is trained in end-to-end manner. Considering the GPU's memory, we crop raw images to small patches. Then, we minimize the normalized logistic loss of the softmax outputs over a given patch (*batchsize* equals 1) as:

$$Loss = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} -1(y^i = j)\log\left(\frac{\exp(z_j^i)}{\sum_{l=1}^{k}\exp(z_l^i)}\right), \quad (3)$$

where N is the number of pixels in a patch and k the number of classes. For the i-th pixel, y^i denotes its label and $(z_1^i, \cdots, z_i^k)$ is the output prediction vector. $1(y = j)$ is an indicator function, it takes 1 when $y = j$, and 0 otherwise.

In labeling stage, to mitigate the discontinuity caused by cropping, we take multi-scale labeling of 0.5, 1 and 1.5 of raw image, and average the final scores at multiple scales. The detailed labeling procedure for each patch is shown in Algorithm 1. Finally, the semantic labeling results of the whole image is constituted by the predicted class labels of all patches.

Algorithm 1 Semantic Labeling with CAC-NET

Input: The image patch $\mathbf{X}$
Output: The predicted class label $\mathbf{P}$

```
Calculate feature maps by inputting X into the encoder
Procedure multi-context aggregating
    Acquire multi-context T_1 ··· T_n at the last layer of encoder by the multi-scale unit operation
    Aggregate T_1 ··· T_n to obtain T by Eq. (1)
Procedure fine-structured objects refinement
    Identify feature maps F^1 ··· F^i in shallow layers
    Calculate M^{i+1} by fusing F^1 ··· F^i initialized with M^1 = Re(T) using Eq. (2)
Calculate P = argmax_{k∈{1,···,k}} (softmax(M^{i+1}))
return P
```

3. EXPERIMENT

We conduct experiments on the large-scale *ISPRS Vaihingen 2D semantic labeling challenge* [23] dataset. This is an open benchmark. Overall, there are 33 tiles of $\approx 2500\times2000$ pixels at a GSD of $\approx$ 9cm in image data. Among them, 16 tiles contain ground truth, and the remaining 17 tiles are withheld by the challenge organizers for online test. Throughout our experiments, only raw image data is used.

We first conduct offline experiment using the supplied 16 tiles with ground truth, where 8 tiles is randomly chosen as training set, and the rest as testing set. Raw images are cropped to a number of 320×320 patches with overlap 160 (no overlap in testing set). The training set is augmented by adding random noise and applying rotate and mirror operations to these patches. Multi-context is aggregated by multi-

Table 1: Comparison with the state-of-the-art models*(%)*. surf: impervious surface (roads), veg: low vegtation.

Method	surf	roof	veg	tree	car	Mean
Segnet [5]	66.9	76.1	44.6	69.7	62.4	63.9
FCN-8s [1]	75.2	80.4	65.6	70.5	45.8	67.5
Deeplab-vgg [16]	80.0	87.9	70.0	75.4	36.1	69.9
Ours(vgg)	**81.3**	**89.3**	**70.3**	**75.5**	**66.4**	**76.6**
Deeplab-res101	81.6	90.7	71.4	**76.7**	58.9	75.9
Ours(res101)	**84.0**	**90.9**	**72.1**	76.6	**75.3**	**79.8**

Table 2: Ablation Experiment*(%)*. MPD: multiple average pooling and dilation, MCC: multi-context cascade, RC: residual correction.

Method	surf	roof	veg	tree	car	Mean
Ours(Deeplab_13)	76.7	82.3	67.8	72.6	40.7	68.0
+ MPD	79.7	86.5	68.3	74.6	47.2	71.3
+ Refinement	80.1	87.1	68.0	74.6	55.5	73.1
+ MCC	80.3	88.1	69.5	76.5	60.0	74.9
+ RC	81.3	89.3	70.3	75.5	66.4	76.6

pooling maps of 1×1, 2×2, 3×3, 6×6 and multi-dilation rates of 6, 12, 18, 24. We take sum fusion throughout the network. In all offline tests, we adopt the Intersection over Union (IoU) as evaluation criterion: $IoU(P_m, P_{gt}) = \frac{|P_m \cap P_{gt}|}{|P_m \cup P_{gt}|}$, where P_{gt} is ground truth and P_m the prediction.

The comparative results are shown in Table 1. Our vgg16-encoder network outperforms all the advanced models based on vgg16. More importantly, our model surpasses the deeplab which is the best method in the literature by a large margin for the fine-structured objects, e.g., cars. This demonstrates the effectiveness of our refinement strategy.

As deeper semantic context plays a more significant guiding role to capture global and local cues, the performance of our network is further improved by replacing vgg16 with the resnet101 [22]. Although deeplab could also benefit from deeper network, its performance is still worse than ours. To evaluate the effectiveness of different parts of our model, we make an ablation experiment. Here, we list the results of adding different parts progressively in Table 2. The deeplab with only the first 13 convolutional layers is taken as baseline. The results show that the proposed components significantly improve the performance, especially for cars.

To further demonstrate the effectiveness of our network, we make the online test which is evaluated by the challenge organizer. The experimental setup is the same as offline experiment, except that all 16 tiles are used for training. The benchmark metrics are the *Overall Accuracy* and the *F1 score* on each class given by:

$$F1_i = 2\frac{pre_i \times rec_i}{pre_i + rec_i} \text{ and } rec_i = \frac{tp_i}{C_i},\ pre_i = \frac{tp_i}{P_i}. \quad (4)$$

Where tp_i is the number of true positives for the i-th class, C_i is the number of pixels belonging to the i-th class, and P_i is the number of pixels attributed to the i-th class by the model.

The competing results obtained from the challenge website [24] (our method is denoted as 'CASIA') are shown in Table 3. Our model outperforms other sophisticated methods even though it uses a single network based on only raw image data. Other competitors either use extra data such as DSM or incorporate with structural models using CRF. Fig. 6 illustrates qualitative comparison on the challenge results.

Table 3: *ISPRS 2D Semantic Labeling Challenge* results*(%)*. OA: Overall Accuracy, DSM: Digital Surface Model

Method	surf	roof	veg	tree	car	OA
FCN+DSM('UZ_1')	89.2	92.5	81.6	86.9	57.3	87.3
CNN+RF+CRF+DSM [3]	89.5	93.2	82.3	88.2	63.3	88.0
FCN+RF+CRF [2]	90.5	93.7	83.4	89.2	72.6	89.1
FCN+Edge+DSM [10]	90.4	93.6	83.9	89.7	76.9	89.2
Segnet+DSM [19]	91.0	94.5	**84.4**	**89.9**	77.8	89.8
Ours(res101)	**92.7**	**95.3**	84.3	89.6	**80.8**	**90.6**

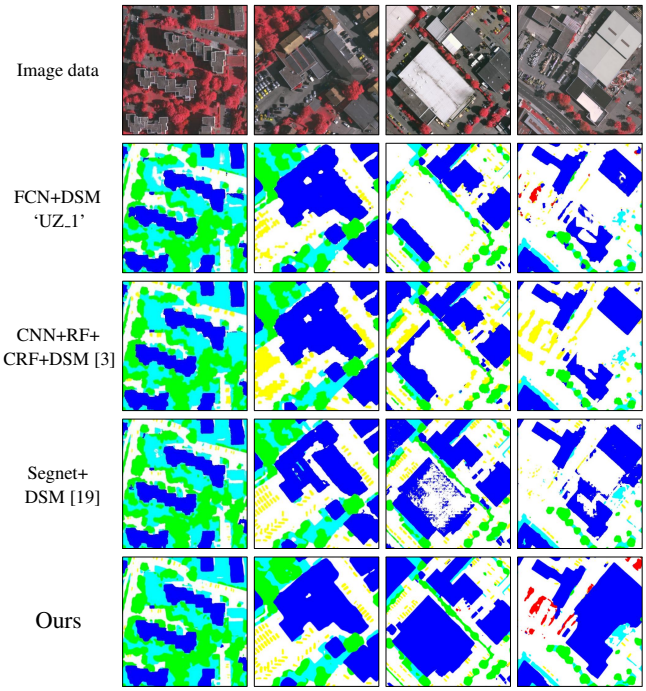


Fig. 6: Qualitative comparison on the challenge results (white: impervious surface, blue: building roof, cyan: low vegetation, green: tree, yellow: car, red: clutter/background).

4. CONCLUSION

In this work, a novel end-to-end network for semantic labeling in VHR image has been proposed. It consists of three key components in addition to a standard CNN. First, by aggregating multi-context with a cascaded architecture, the labeling accuracy of complex manmade objects is improved. Second, a hierarchical pyramid structure is proposed to refine the labeling results of fine-structured objects. Thirdly, a residual correction scheme is employed throughout the network to alleviate the impact of fitting residual. Thanks to these specific designs, the proposed network achieves the state-of-the-art performance on the large-scale *ISPRS Semantic Labeling Challenge* dataset, outperforming all the other participants.

5. REFERENCES

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[2] Jamie Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *arXiv preprint arXiv:1606.02585*, 2016.

[3] Sakrapee Paisitkriangkrai, Jamie Sherrah, Pranam Janney, and Anton van den Hengel, "Semantic labeling of aerial and satellite imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 7, pp. 2868–2881, 2016.

[4] K. Nogueira, M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Learning to semantically segment high-resolution remote sensing images," in *IEEE Conference on Pattern Recognition (ICPR)*, 2016, pp. 3566–3571.

[5] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv:1505.07293*, 2015.

[6] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Hypercolumns for object segmentation and fine-grained localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 447–456.

[7] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2874–2883.

[8] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár, "Learning to refine object segments," *arXiv preprint arXiv:1603.08695*, 2016.

[9] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani, "Semantic segmentation with boundary neural fields," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3602–3610.

[10] Dimitrios Marmanis, Konrad Schindler, Jan Dirk Wegner, Silvano Galliani, Mihai Datcu, and Uwe Stilla, "Classification with an edge: improving semantic image segmentation with boundary detection," *arXiv preprint arXiv:1612.01337*, 2016.

[11] Sergey Zagoruyko, Adam Lerer, Tsung-Yi Lin, Pedro O Pinheiro, Sam Gross, Soumith Chintala, and Piotr Dollár, "A multipath network for object detection," *arXiv preprint arXiv:1604.02135*, 2016.

[12] Spyros Gidaris and Nikos Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1134–1142.

[13] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3376–3385.

[14] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 8, pp. 1915–1929, 2013.

[15] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3640–3649.

[16] Chen Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *International Conference on Learning Representations (ICLR)*, 2015.

[17] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations (ICLR)*, 2016.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 37, no. 9, pp. 1904–1916, 2015.

[19] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," *arXiv preprint arXiv:1609.06846*, 2016.

[20] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3626–3633.

[21] Wei Liu, Andrew Rabinovich, and Alexander C Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[23] "ISPRS Vaihingen 2D Semantic Labeling Challenge," http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html.

[24] "ISPRS 2D Semantic Labeling Challenge Benchmark Test Results," http://www2.isprs.org/vaihingen-2d-semantic-labeling-contest.html.