ISPRS Journal of Photogrammetry and Remote Sensing xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Semantic labeling in very high resolution images via a self-cascaded convolutional neural network

Yongcheng Liu [a,b], Bin Fan [a,*], Lingfeng Wang [a], Jun Bai [c], Shiming Xiang [a], Chunhong Pan [a]

[a] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, PR China
[b] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, PR China
[c] Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, PR China

ARTICLE INFO

Article history:
Received 30 May 2017
Received in revised form 8 November 2017
Accepted 17 December 2017
Available online xxxx

Keywords:
Semantic labeling
Convolutional neural networks (CNNs)
Multi-scale contexts
End-to-end

ABSTRACT

Semantic labeling for very high resolution (VHR) images in urban areas, is of significant importance in a wide range of remote sensing applications. However, many confusing manmade objects and intricate fine-structured objects make it very difficult to obtain both coherent and accurate labeling results. For this challenging task, we propose a novel deep model with convolutional neural networks (CNNs), i.e., an end-to-end self-cascaded network (ScasNet). Specifically, for confusing manmade objects, ScasNet improves the labeling coherence with sequential global-to-local contexts aggregation. Technically, multi-scale contexts are captured on the output of a CNN encoder, and then they are successively aggregated in a self-cascaded manner. Meanwhile, for fine-structured objects, ScasNet boosts the labeling accuracy with a coarse-to-fine refinement strategy. It progressively refines the target objects using the low-level features learned by CNN's shallow layers. In addition, to correct the latent fitting residual caused by multi-feature fusion inside ScasNet, a dedicated residual correction scheme is proposed. It greatly improves the effectiveness of ScasNet. Extensive experimental results on three public datasets, including two challenging benchmarks, show that ScasNet achieves the state-of-the-art performance.

1. Introduction

Semantic labeling in very high resolution (VHR) images is a long-standing research problem in remote sensing field. It plays a vital role in many important applications, such as infrastructure planning, territorial planning and urban change detection (Lu et al., 2017a; Matikainen and Karila, 2011; Zhang and Seto, 2011). The target of this problem is to assign each pixel to a given object category. Note that it is not just limited to building extraction (Li et al., 2015a), road extraction (Cheng et al., 2017b) and vegetation extraction (Wen et al., 2017) which only consider labeling one single category, semantic labeling usually considers several categories simultaneously (Li et al., 2015b; Xu et al., 2016; Xue et al., 2015). As a result, this task is very challenging, especially for the urban areas, which exhibit high diversity of manmade objects. Specifically, on one hand, many manmade objects (e.g., buildings) show various structures, and they are composed of a large number of different materials. Meanwhile, plenty of different manmade objects (e.g., buildings and roads) present much similar visual characteristics. These confusing manmade objects with high intra-class variance and low inter-class variance bring much difficulty for coherent labeling. On the other hand, fine-structured objects in cities (e.g., cars, trees and low vegetations) are quite small or threadlike, and they also interact with each other through occlusions and cast shadows. These factors always lead to inaccurate labeling results. Furthermore, it poses additional challenge to simultaneously label all these size-varied objects well.

To accomplish such a challenging task, features at different levels are required. Specifically, abstract high-level features are more suitable for the recognition of confusing manmade objects, while labeling of fine-structured objects could benefit from detailed low-level features. Convolutional neural networks (CNNs) (Lecun et al., 1990) in *deep learning* field are well-known for feature learning (Mas and Flores, 2008). CNNs consist of multiple trainable layers which can extract expressive features of different levels (Lecun et al., 1998). Moreover, recently, CNNs with *deep learning* have demonstrated remarkable learning ability in computer vision field, such as scene recognition (Yuan et al., 2015) and image segmentation (Long et al., 2015). Meanwhile, the development of remote sensing has also been greatly promoted by numerous

* Corresponding author at: National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, PR China.
E-mail address: bfan@nlpr.ia.ac.cn (B. Fan).

https://doi.org/10.1016/j.isprsjprs.2017.12.007
0924-2716/

CNNs-based methods (Cheng et al., 2017a). For example, deconvolution networks (Zeiler et al., 2010) are investigated by Lu et al. (2017b) for remote sensing scene classification, and Chen et al. (2016b) perform target classification using CNNs for SAR Images.

Based on CNNs, many patch-classification methods are proposed to perform semantic labeling (Mnih, 2013; Mostajabi et al., 2015; Paisitkriangkrai et al., 2016; Nogueira et al., 2016; Alshehhi et al., 2017; Zhang et al., 2017). These methods determine a pixel's label by using CNNs to classify a small patch around the target pixel. However, they are far from optimal, because they ignore the inherent relationship between patches and their time consumption is huge (Maggiori et al., 2017). Typically, fully convolutional networks (FCNs) have boosted the accuracy of semantic labeling a lot (Long et al., 2015; Sherrah, 2016). FCNs perform pixel-level classification directly and now become the normal framework for semantic labeling. Nevertheless, due to multiple *sub-samplings* in FCNs, the final *feature maps* are much coarser than the input image, resulting in less accurate labeling results.

Accordingly, a tough problem locates on how to perform accurate labeling with the coarse output of FCNs-based methods, especially for fine-structured objects in VHR images. To solve this problem, some researches try to reuse the low-level features learned by CNNs' shallow layers (Zeiler and Fergus, 2014). The aim is to utilize the local details (e.g., corners and edges) captured by the *feature maps* in fine resolution. Technically, they perform operations of multi-level feature fusion (Ronneberger et al., 2015; Long et al., 2015; Hariharan et al., 2015; Pinheiro et al., 2016), *deconvolution* (Noh et al., 2015) or *up-pooling* with recorded *pooling* indices (Badrinarayanan et al., 2015). Most of these methods use the strategy of direct stack-fusion. However, this strategy ignores the inherent semantic gaps in features of different levels. An alternative way is to impose boundary detection (Bertasius et al., 2016; Marmanis et al., 2016). It usually requires extra boundary supervision and leads to extra model complexity despite boosting the accuracy of object localization.

Another tricky problem is the labeling incoherence of confusing objects, especially of the various manmade objects in VHR images. To tackle this problem, some researches concentrate on leveraging the multi-context to improve the recognition ability of those objects. They use multi-scale images (Farabet et al., 2013; Mostajabi et al., 2015; Cheng et al., 2016; Liu et al., 2016b; Chen et al., 2016a; Zhao and Du, 2016) or multi-region images (Gidaris and Komodakis, 2015; Luus et al., 2015) as input to CNNs. However, these methods are usually less efficient due to a lot of repetitive computation. Differently, some other researches are devoted to acquire multi-context from the inside of CNNs. They usually perform operations of multi-scale *dilated convolution* (Chen et al., 2015), multi-scale pooling (He et al., 2015b; Liu et al., 2016a; Bell et al., 2016) or multi-kernel convolution (Audebert et al., 2016), and then fuse the acquired multi-scale contexts in a direct stack manner. Nevertheless, this manner not only ignores the hierarchical dependencies among the objects and scenes in different scales, but also neglects the inherent semantic gaps in contexts of different-level information.

In summary, although current CNN-based methods have achieved significant breakthroughs in semantic labeling, it is still difficult to label the VHR images in urban areas. The reasons are as follows: (1) Most existing approaches are less efficient to acquire multi-scale contexts for confusing manmade objects recognition; (2) Most existing strategies are less effective to utilize low-level features for accurate labeling, especially for fine-structured objects; (3) Simultaneously fixing the above two issues with a single network is particularly difficult due to a lot of fitting residual in the network, which is caused by semantic gaps in different-level contexts and features.

In this paper, we propose a novel self-cascaded convolutional neural network (ScasNet), as illustrated in Fig. 1. The aim of this work is to further advance the state of the art on semantic labeling in VHR images. To this end, it is focused on three aspects: (1) multi-scale contexts aggregation for distinguishing confusing manmade objects; (2) utilization of low-level features for fine-structured objects refinement; (3) residual correction for more effective multi-feature fusion. Specifically, a conventional CNN is adopted as an encoder to extract features of different levels. On the *feature maps* outputted by the encoder, global-to-local contexts are sequentially aggregated for confusing manmade objects recognition. Technically, multi-scale contexts are first captured by different convolutional operations, and then they are successively aggregated in a self-cascaded manner. With the acquired contextual information, a coarse-to-fine refinement strategy is performed to refine the fine-structured objects. It progressively reutilizes the low-level features learned by CNN's shallow layers with long-span connections. In addition, to correct the latent fitting residual caused by semantic gaps in multi-feature fusion, several residual correction schemes are employed throughout the network. As a result of residual correction, the above two different solutions could work collaboratively and effectively when they are integrated into a single network. Extensive experiments demonstrate the effectiveness of ScasNet. Moreover, the three submodules in ScasNet could not only provide good solutions for semantic labeling, but are also suitable for other tasks such as object detection

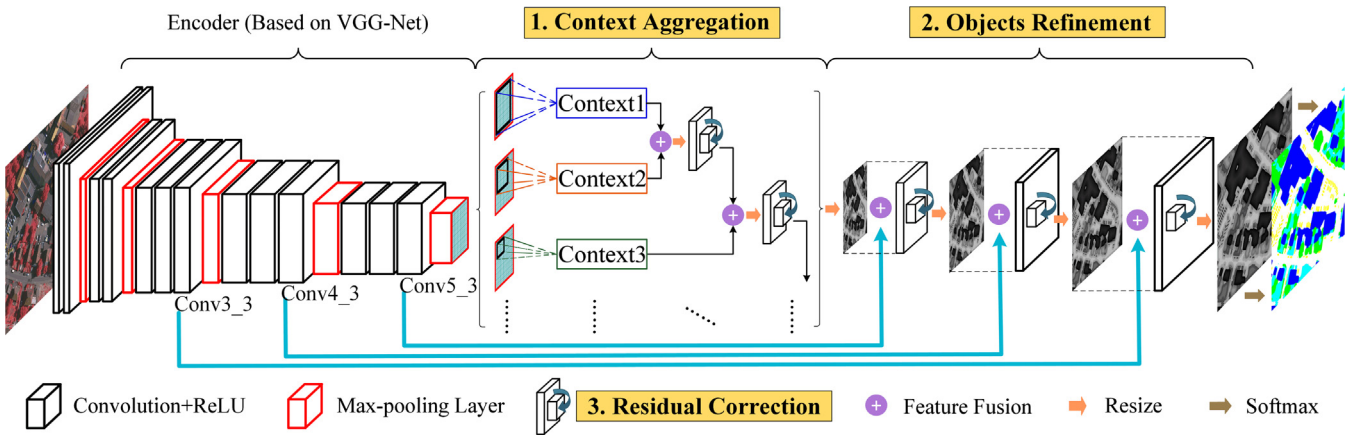

Fig. 1. Overview of the proposed ScasNet. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Liu, Y., et al. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. ISPRS J. Photogram. Remote Sensing (2017), https://doi.org/10.1016/j.isprsjprs.2017.12.007