Vehicle Instance Segmentation From Aerial Image and Video Using a Multitask Learning Residual Fully Convolutional Network

Lichao Mou, *Student Member, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*— Object detection and semantic segmentation are two main themes in object retrieval from high-resolution remote sensing images, which have recently achieved remarkable performance by surfing the wave of deep learning and, more notably, convolutional neural networks. In this paper, we are interested in a novel, more challenging problem of vehicle instance segmentation, which entails identifying, at a pixel level, where the vehicles appear as well as associating each pixel with a physical instance of a vehicle. In contrast, vehicle detection and semantic segmentation each only concern one of the two. We propose to tackle this problem with a semantic boundary-aware multitask learning network. More specifically, we utilize the philosophy of residual learning to construct a fully convolutional network that is capable of harnessing multilevel contextual feature representations learned from different residual blocks. We theoretically analyze and discuss why residual networks can produce better probability maps for pixelwise segmentation tasks. Then, based on this network architecture, we propose a unified multitask learning network that can simultaneously learn two complementary tasks, namely, segmenting vehicle regions and detecting semantic boundaries. The latter subproblem is helpful for differentiating "touching" vehicles that are usually not correctly separated into instances. Currently, data sets with a pixelwise annotation for vehicle extraction are the ISPRS data set and the IEEE GRSS DFC2015 data set over Zeebrugge, which specializes in a semantic segmentation. Therefore, we built a new, more challenging data set for vehicle instance segmentation, called the *Busy Parking Lot Unmanned Aerial Vehicle Video data set*, and we make our data set available at http://www.sipeo.bgu.tum.de/downloads so that it can be used to benchmark future vehicle instance segmentation algorithms.**

***Index Terms*— Boundary-aware multitask learning network, fully convolutional network (FCN), high-resolution remote sensing image/video, instance semantic segmentation, residual neural network (ResNet), vehicle detection.**

Manuscript received November 18, 2017; revised April 2, 2018 and May 22, 2018; accepted May 23, 2018. Date of publication July 9, 2018; date of current version October 25, 2018. This work was supported in part by the China Scholarship Council, in part by the European Research Council through the European Union's Horizon 2020 Research and Innovation Programme under Grant ERC-2016-StG-714087 (So2Sat), in part by the Helmholtz Association through the Framework of the Young Investigators Group (SiPEO) under Grant VH-NG-1018, and in part by the Bavarian Academy of Sciences and Humanities through the Framework of Junges Kolleg. *(Corresponding author: Xiao Xiang Zhu.)*

The authors are with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany, and also with Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: lichao.mou@dlr.de; xiao.zhu@dlr.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2018.2841808

I. Introduction

THE last decade has witnessed dramatic progress in modern remote sensing technologies—along with the launch of small and cheap commercial high-resolution satellites and the now widespread availability of unmanned aerial vehicles (UAVs)—which facilitates a diversity of applications, such as urban management [1]–[4], monitoring of land changes [5]–[8], and traffic monitoring [9], [10]. Among these applications, object extraction from very high-resolution remote sensing images/videos has gained increasing attention in the remote sensing community in recent years, particularly vehicle extraction, due to successful civil applications. Vehicle extraction, however, is still a challenging task, mainly because it is easily affected by several factors, e.g., vehicle appearance variation, the effects of shadow, illumination, and a complicated and cluttered background. Existing vehicle extractionapproaches can be roughly divided into two categories: vehicle detection and vehicle semantic segmentation.

A. Vehicle Detection

The goal of vehicle detection is to detect all instances of vehicles and localize them in the image, typically in the form of bounding boxes with confidence scores. Traditionally, this topic was addressed by works that use low-level, hand-crafted visual features [e.g., color histogram, texture feature, scale-invariant feature transform (SIFT), and histogram of oriented gradients (HOG)] and classifiers. For example, Shao *et al.* [11] incorporate multiple visual features, local binary patterns, HOG, and opponent histogram for vehicle detection from high-resolution aerial images. Moranduzzo and Melgani [12] first use SIFT to detect the interest points of vehicles and then train a support vector machine (SVM) to classify these interest points into vehicle and nonvehicle categories based on the SIFT descriptors. They later present an approach [13] that performs filtering operations in the horizontal and vertical directions to extract HOG features and yield vehicle detection after the computation of a similarity measure, using a catalog of vehicles as a reference. Liu and Mattyus [14] make use of an integral channel concept with Haar-like features and an AdaBoost classifier in a soft-cascade structure to achieve fast and robust vehicle detection.

The aforementioned approaches mainly rely on the hand-crafted features for constructing a classification system.

0196-2892 © 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Fig. 1. Illustration of different vehicle extraction methods. (From left to right and top to bottom) Input image, vehicle detection, semantic segmentation, and vehicle instance segmentation. The challenge of vehicle instance segmentation is that some vehicles are segmented incorrectly. While most pixels belonging to the category are identified correctly, they are not correctly separated into instances (see arrows in the bottom-left image).

Recently, as an important branch of the deep learning family, the convolutional neural network (CNN) has become the method of choice in many computer vision and remote sensing problems [15]–[19] (e.g., object detection) due to its ability to automatically extract midlevel and high-level abstract features from raw images for pattern recognition purposes. Chen *et al.* [20] propose a vehicle detection model, called the hybrid deep neural network, which consists of a sliding window technique and CNN. The main insight behind their model is to divide the feature maps of the last convolutional layer into different scales, allowing for the extraction of multiscale features for vehicle detection. Ammour *et al.* [21] segment an input image into homogeneous superpixels that can be considered as vehicle candidate regions, making use of a pretrained deep CNN to extract features, and train a linear SVM to classify these candidate regions into vehicle and nonvehicle classes.

B. Vehicle Semantic Segmentation

Vehicle semantic segmentation aims to label each pixel in an image as belonging to the vehicle class or other categories (e.g., building, tree, and low vegetation). In comparison with vehicle detection, it can give more accurate pixelwise extraction results. More recently, progress in deep CNNs, particularly fully convolutional networks (FCNs), makes it possible to achieve end-to-end vehicle semantic segmentation. For instance, Audebert *et al.* [22] propose a deep-learning-based "segment-before-detect" method for semantic segmentation and subsequent classification of several types of vehicles in high-resolution remote sensing images. The use of SegNet [23] in this method is capable of producing pixelwise annotations for vehicle semantic mapping. In addition, several recent works in the semantic segmentation of high-resolution aerial imaging also involve vehicle segmentation. Kampffmeyer *et al.* [24] focus on the class imbalance which often represents a problem for semantic segmentation in remote sensing images, since small objects (e.g., vehicles) are less prioritized in an effort to achieve a good overall accuracy (OA). To address this problem, they train FCNs using the cross-entropy loss function weighted with median frequency balancing, which is proposed by Eigen and Fergus [25].

C. Is Semantic Segmentation Good Enough for Vehicle Extraction?

The existence of "touching" vehicles in a remote sensing image makes it quite hard for most vehicle semantic segmentation methods to separate objects individually, while in most cases, we need to know not only which pixels belong to vehicles (vehicle semantic segmentation problem) but also the exact number of vehicles (vehicle detection task). This drives us to examine an instance-oriented vehicle segmentation.

The vehicle instance segmentation seeks to identify the semantic class of each pixel (i.e., vehicle or nonvehicle) as well as associate each pixel with a physical instance of a vehicle. This is contrasted with the vehicle semantic segmentation which is only concerned with the above-mentioned first task. Fig. 1 shows differences among vehicle detection, semantic segmentation, and instance segmentation. In this paper, we are interested in the vehicle instance segmentation in a complex, cluttered, and challenging background from aerial images and videos. Moreover, since deep networks have recently been very successful in a variety of remote sensing applications, from hyperspectral/multispectral image analysis to interpretation of high-resolution aerial images to multimodal data fusion [15], in this paper, we would like to use an end-to-end network to achieve the vehicle instance segmentation. This paper contributes to the literature in three major respects.

1) So far, most studies in the remote sensing community have focused on the object detection and semantic segmentation in high-resolution remote sensing imagery. The instance segmentation has rarely been addressed. In a pioneer work moving from semantic segmentation to instance segmentation, Audebert *et al.* [22] developed a three-stage segment-before-detect framework. In this paper, we try to address the vehicle instance segmentation problem by an end-to-end learning framework.
2) In order to facilitate progress in the field of vehicle instance segmentation in high-resolution aerial images/videos, we provide a new, challenging data set that presents a high range of variation—with a diversity of vehicle appearances, the effects of shadow, a cluttered background, and extremely close vehicle distances—for producing quantitative measurements and comparing among approaches.
3) We present a semantic boundary-aware unified multitask learning FCN, which is end-to-end trainable, for vehicle instance segmentation. Inspired by several recent works [26]–[28], we exploit residual neural network (ResNet) [29] to construct the feature extractor

of the whole network. In this paper, we theoretically analyze and discuss why residual networks can produce better probability maps for pixelwise prediction tasks. The proposed multitask learning network creates two separate, yet identical branches to jointly optimize two complementary tasks—namely, vehicle semantic segmentation and semantic boundary detection. The latter subproblem is beneficial for differentiating vehicles with an extremely close distance and further improving the instance segmentation performance.

The remainder of this paper is organized as follows. After Section I, detailing vehicle extraction from high-resolution remote sensing imagery, we enter Section II, dedicated to the details of the proposed semantic boundary-aware multitask learning network for vehicle instance segmentation. Section III then provides the data set information, the network setup, and the experimental results and discussion. Finally, Section IV concludes this paper.

II. Methodology

We formulate the vehicle instance segmentation task by two subproblems, namely, vehicle detection and semantic segmentation. The training set is denoted by $\{(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)\}$, where $i = 1, 2, \ldots, N$ and N is the number of training samples. Since we consider each image independently, the subscript i is dropped hereafter for notational simplicity. $\boldsymbol{x} = \{x_j, j = 1, 2, \ldots, |\boldsymbol{x}|\}$ represents a raw input image, $\boldsymbol{y} = \{y_j, j = 1, 2, \ldots, |\boldsymbol{x}|, y_j \in \{0, 1\}\}$ denotes its corresponding manually annotated pixelwise segmentation mask, and $\boldsymbol{z} = \{\boldsymbol{r}_k, k = 0, 1, \ldots, K\}$ is the instance label, where $\boldsymbol{r}_k$ indicates a set of pixels inside the kth region.[1] K is the total number of vehicle instances in the image, and $\boldsymbol{r}_0$ is the background area. When k takes other values, it denotes the corresponding vehicle instance. Note that the instance labels only count vehicle instances, and thus, they are commutative. Our aim is to segment vehicles while ensuring that all instances are differentiated. In this paper, we approximate the vehicle detection by the semantic boundary detection.[2] We generate the semantic boundary labels $\boldsymbol{b}$ through $\boldsymbol{z}$ to train a boundary detector, in which $\boldsymbol{b} = \{b_j, j = 1, 2, \ldots, |\boldsymbol{x}|, b_j \in \{0, 1\}\}$ and b_j equals 1 when it belongs to boundaries.

In this section, we describe our proposed semantic boundary-aware multitask learning network for accurate vehicle instance segmentation in detail. We start by introducing the FCN architecture for end-to-end semantic segmentation in Section II-A. Furthermore, we propose to exploit multilevel contextual feature representations, generated by different stages of a residual network, to construct a residual FCN (ResFCN) for producing better likelihood maps of vehicle regions or semantic boundaries (see Section II-B). Then, in Section II-C, we elaborate the semantic boundary-aware unified multitask learning network drawn from the ResFCN for effective instance segmentation by jointly optimizing the complementary tasks.

[1]Regions in the image satisfy $\boldsymbol{r}_k \cap \boldsymbol{r}_t = \varnothing, \forall k \neq t$ and $\cup \boldsymbol{r}_k = \Omega$, where Ω is the whole image region.

[2]The semantic boundary detection is to detect the boundaries of each object instance in the images. Compared with edge detection, it focuses more on the association of boundaries and their object instances.

A. Fully Convolutional Network for Semantic Segmentation

Long *et al.* [30] first proposed the FCN architecture for semantic segmentation tasks which is both efficient and effective. Later, some extensions of the FCN model have been proposed to improve a semantic segmentation performance. To name a few, Chen *et al.* [31] removed some of the max-pooling operations and, accordingly, introduced atrous/dilated convolutions in their network, which can expand the field of view without increasing the number of parameters. As postprocessing, a dense conditional random field (CRF) was trained separately to refine the estimated category score maps for further improvement. Zhang *et al.* [32] introduced a new form of network that combines FCN- and CRF-based probabilistic graphical modeling to simulate a mean-field approximate inference for the CRF with Gaussian pairwise potentials as the recurrent neural network.

B. Residual Fully Convolutional Network

Here, we first explain how to construct a ResFCN according to the existing works in the literature, mainly, the ResNet [29] and FCN [30]. Then, we theoretically analyze why ResFCN is able to offer better performance than other FCNs based on the traditional feedforward network architectures (e.g., VGG Nets [33]).

Network Design: Several recent studies in computer vision [26]–[28] have shown that ResNet [29] is capable of offering better features for pixelwise prediction tasks, such as semantic segmentation [26], [27] and depth estimation [28]. We, therefore, make use of ResNet to construct the segmentation network in this paper. We initialize a ResFCN from the original version of ResNet [29], instead of the newly presented preactivation version [34]. Unlike [30], we directly remove the fully connected layers from the original ResNet but do not convolutionalize these layers so as to make one prediction per spatial location. Moreover, we keep the 7×7 convolutional layer and 3×3 max-pooling layer, which can enlarge the field of view for feature representations. One of the recent trends in a network architecture design is stacking convolutional layers with small convolution kernels (e.g., 3×3 and 1×1) in the entire network, because the stacked small kernels are more efficient than a large filter, given the same computational complexity. However, a recent study [35] found that the large filter also plays an important role when classification and localization tasks are performed simultaneously. This can be easily understood through the analogy of individuals commonly confirming the category of a pixel by referring to its surrounding context region.

By now, the output feature maps are only $1/32$ the resolution of their original input image, which is apparently too low to precisely differentiate individual pixels. To deal with this problem, Long *et al.* [30] made use of backward-strided convolutions that upsample the feature maps and output score masks. The motivation behind this is that the convolutional layers and max-pooling layers focus on extracting high-level abstract features, whereas the backward-strided convolutions estimate the score masks in a pixelwise way. Ghiasi and Fowlkes [36] proposed a multiresolution recon-

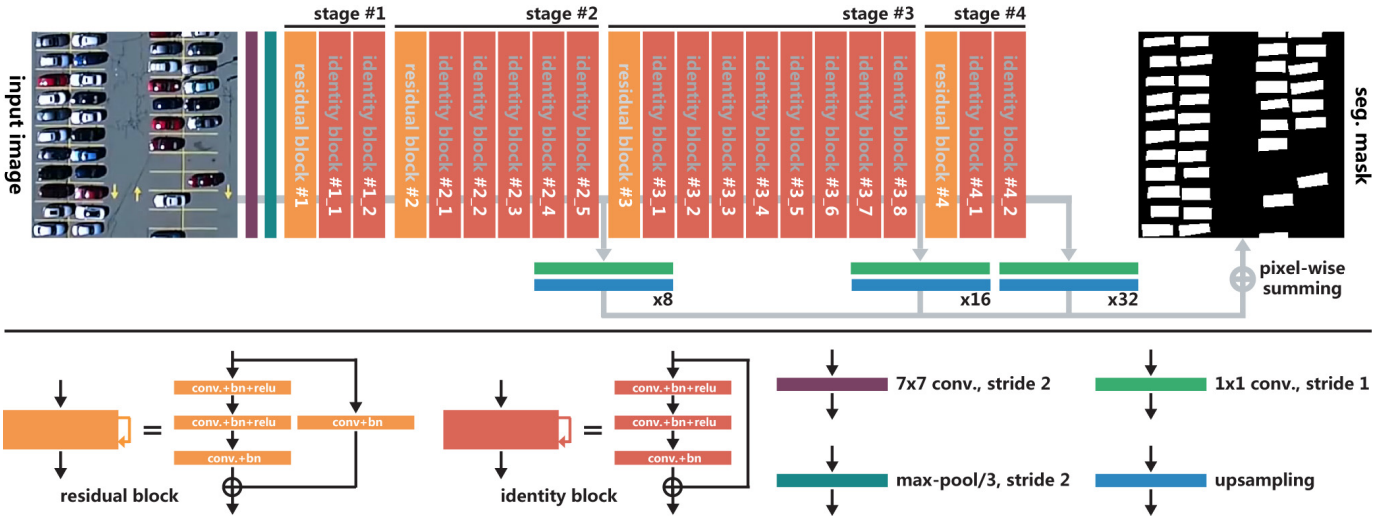


Fig. 2. Network architecture of the ResFCN we use, as illustrated in Section II-B. We incorporate the multilevel contextual features from the last 32×32, 16×16, and 8×8 layers of a classification ResNet since making use of information from fairly early fine-grained layers is beneficial to segmenting small objects such as vehicles. To get the desired full resolution output, we use 1×1 convolutional layers followed by upsampling operations to upsample back to the spatial resolution of the input image. Then, predictions from different residual blocks are fused together with a summing operation.

struction architecture based on a Laplacian pyramid that uses skip connections from higher resolution feature maps and multiplicative gating to successively refine segment boundaries reconstructed from lower resolution maps. Inspired by the existing works, in this paper, we exploit multilevel contextual feature representations that include information from different residual blocks (i.e., different levels of contextual information). Fig. 2 shows the illustration of the ResFCN architecture we use with multilevel contextual features. More specifically, we incorporate feature representations from the last 32×32, 16×16, and 8×8 layers of the original ResNet, since making use of information from fairly early fine-grained layers is beneficial to segmenting small objects such as vehicles. To get the desired full resolution output, we used a 1×1 convolutional layer, which adaptively squashes the number of channels down to the number of labels (1 for binary classification), takes advantage of the upsampling operation to upsample back to the spatial resolution of the input image, and makes predictions based on the contextual cues from the given fields of view. Then, these predictions are fused together with a summing operation, and the final segmentation results are generated after sigmoid classification.

Why Residual Learning? Until recently, the majority of feedforward networks, such as AlexNet [37] and VGG Nets [33], were made up of a linear sequence of layers. $\boldsymbol{x}_{n-1}$ and $\boldsymbol{x}_n$ are denoted as the input and output of the nth layer/block, respectively, and each layer in such a network learns the mapping function $\mathcal{F}$

$$\boldsymbol{x}_n = \mathcal{F}(\boldsymbol{x}_{n-1}; \boldsymbol{\Theta}_n) \tag{1}$$

where $\boldsymbol{\Theta}_n$ is the parameters of the nth layer. This kind of network is also often referred to as a traditional feedforward network.

According to a study by He *et al.* [29], simply deepening traditional feedforward networks usually leads to an increase in training and test errors (i.e., so-called degradation problem). A residual learning-based network is composed of a sequence of residual blocks and exhibits significantly improved training characteristics, providing the opportunity to make network depths that were previously unattainable. The output $\boldsymbol{x}_n$ of the nth residual block in a ResNet can be computed as

$$\boldsymbol{x}_n = \mathcal{H}(\boldsymbol{x}_{n-1}; \boldsymbol{\Theta}_n) + \boldsymbol{x}_{n-1} \tag{2}$$

where $\mathcal{H}(\boldsymbol{x}_{n-1}; \boldsymbol{\Theta}_n)$ is the residual, which is parametrized by $\boldsymbol{\Theta}_n$. The core insight of ResNet is that the addition of a shortcut connection from the input $\boldsymbol{x}_{n-1}$ to the output $\boldsymbol{x}_n$ bypasses two or more convolutional layers by performing identity mapping and is then added together with the output of stacked convolutions. By doing so, $\mathcal{H}$ only computes a residual instead of computing the output $\boldsymbol{x}_n$ directly.

In the experiments, we found that the ResFCN can offer a better performance than the other FCNs based on the traditional feedforward network architecture, such as VGG-FCN. What is the reason behind this? To answer this question, we need to go deeper. According to the characteristics of the ResFCN, we can easily get the following recurrence formula:

$$\boldsymbol{x}_m = \sum_{i=n-1}^{m-1} \mathcal{H}(\boldsymbol{x}_i; \boldsymbol{\Theta}_{i+1}) + \boldsymbol{x}_{n-1} \tag{3}$$

for any deeper residual block m and any shallower residual block n. Equation (3) shows that the ResFCN creates a direct path for propagating information of shallow layers (i.e., $\boldsymbol{x}_{n-1}$) through the entire network. Several recent studies [38], [39] that attempt to reveal what were learned by CNNs show that the deeper layers exploit filters to grasp global high-level information, while the shallower layers capture low-level details, such as object boundaries and edges, which are of great importance in small object detection/segmentation. In addition, when we dive into the backward propagation process,

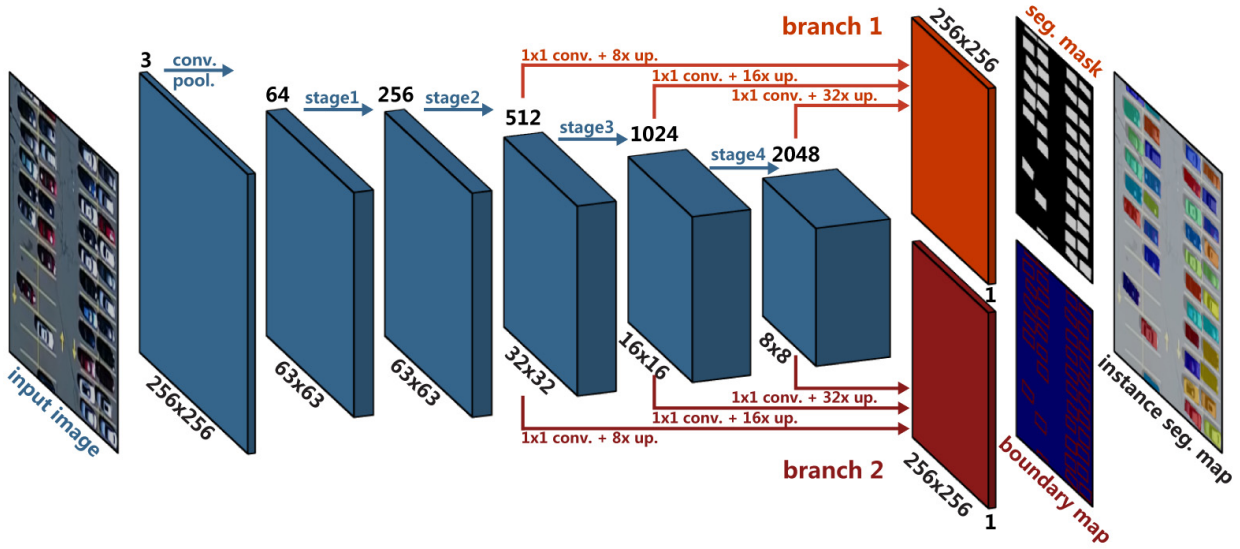


Fig. 3. Overall architecture of the proposed semantic B-ResFCN. We propose to use such a unified multitask learning network for vehicle instance segmentation which creates two separate, yet identical branches to jointly optimize two complementary tasks, namely, vehicle semantic segmentation and semantic boundary detection. The latter subproblem is beneficial for differentiating "touching" vehicles and further improving the instance segmentation performance.

according to the chain rule of backpropagation, we can obtain

$$\frac{\partial \mathcal{E}}{\partial \boldsymbol{x}_{n-1}} = \frac{\partial \mathcal{E}}{\partial \boldsymbol{x}_m}\frac{\partial \boldsymbol{x}_m}{\partial \boldsymbol{x}_{n-1}} = \frac{\partial \mathcal{E}}{\partial \boldsymbol{x}_m}\left(1 + \frac{\partial}{\partial \boldsymbol{x}_{n-1}}\sum_{i=n-1}^{m-1}\mathcal{H}(\boldsymbol{x}_i;\boldsymbol{\Theta}_{i+1})\right) \quad (4)$$

where $\mathcal{E}$ is the loss function of the network. As exhibited in (4), the gradient $(\partial\mathcal{E}/(\partial \boldsymbol{x}_{n-1}))$ can be decomposed into two additive terms: the term $(\partial\mathcal{E}/(\partial \boldsymbol{x}_m))((\partial/(\partial \boldsymbol{x}_{n-1}))\sum_{i=n-1}^{m-1}\mathcal{H})$ that passes information through the weight layers, and the term $(\partial\mathcal{E}/(\partial \boldsymbol{x}_m))$ that directly propagates without concerning any weight layers. The latter term ensures that the information can also be directly propagated back to any shallower residual block n.

In brief, the properties of the forward and backward propagation procedures of the ResFCN make it possible to shuttle the low-level visual information directly across the network, which is quite helpful for our vehicle (small object) instance segmentation tasks.

C. Semantic Boundary-Aware ResFCN

By exploiting the multilevel contextual features, the ResFCN is capable of producing good likelihood maps of vehicles. However, it is still difficult to differentiate vehicles with a very close distance by only leveraging the probability of vehicles due to the ambiguity in the "touching" regions. This is rooted in the loss of spatial details caused by max-pooling layers (downsampling) along with the feature abstraction. The semantic boundaries of vehicles provide good complementary cues that can be used for separating the instances.

Some approaches in computer vision and remote sensing have been explored for modeling segmentation and boundary prediction jointly in a combinatorial framework. For example, Kirillov *et al.* [40] propose InstanceCut which represents instance segmentation by two modalities, namely, a semantic segmentation and all instance boundaries. The former is computed from a CNN for semantic segmentation, and the latter is derived from an instance-aware edge detector. However, this approach does not address end-to-end learning. In the remote sensing community, Marmanis *et al.* [41] propose a two-step model that learns a CNN to separately output edge likelihoods at multiple scales from color-infrared and height data. Then, the boundaries detected with each source are added as an extra channel to each source, and a network is trained for semantic segmentation purposes. The intuition behind this paper is that using predicted boundaries helps to achieve sharper segmentation maps. In contrast, we train one end-to-end network that takes as input color images and predicts segmentation maps and object boundaries in order to augment the performance of segmentation at the instance level.

To this end, we train a deep semantic boundary-aware ResFCN (B-ResFCN) for effective vehicle instance segmentation (i.e., segmenting the vehicles and splitting clustered instances into individual ones). Fig. 3 shows an overview of the proposed network. Specifically, we formulate it as a unified multitask learning network architecture by exploring the complementary information (i.e., vehicle region and semantic boundaries), instead of treating the vehicle segmentation problem as an independent and single task, which can simultaneously learn the detections of vehicle regions and corresponding semantic boundaries. As shown in Fig. 3, the feature representations extracted from multiple residual blocks are upsampled with two separate, yet identical branches to predict the semantic segmentation masks of vehicles and semantic boundaries, respectively. In each branch, the mask is estimated by the ResFCN with multilevel contextual features, as illustrated in Section II-B. Since we have only two categories (foreground/vehicles versus background and semantic boundaries versus nonboundaries), sigmoid and binary cross-entropy loss

are used to train these two branches. Formally, the network training can be formulated as a pixel-level binary classification problem regarding ground-truth segmentation masks, including vehicle instances and semantic boundaries, as shown in the following:

$$\mathcal{L}(x; \boldsymbol{W}) = \mathcal{L}_s(x; \boldsymbol{W}_n, \boldsymbol{W}_s) + \lambda \mathcal{L}_b(x; \boldsymbol{W}_n, \boldsymbol{W}_b) \tag{5}$$

where

$$\begin{aligned} \mathcal{L}_s &= -\sum_{x \in \boldsymbol{x}} [y \log \sigma_s(x) + (1-y) \log(1-\sigma_s(x))] \\ \mathcal{L}_b &= -\sum_{x \in \boldsymbol{x}} [b \log \sigma_b(x) + (1-b) \log(1-\sigma_b(x))]. \end{aligned} \tag{6}$$

$\mathcal{L}_s(x; \boldsymbol{W}_n, \boldsymbol{W}_s)$ and $\mathcal{L}_b(x; \boldsymbol{W}_n, \boldsymbol{W}_s)$ denote the losses for estimating vehicle regions and semantic boundaries, respectively. We train the network using this joint loss, and the final instance segmentation map is produced by the first branch of the network in the test phase. Vehicle instances are obtained by computing the connected regions in the predicted segmentation map. Inside a region, pixels belong to the same vehicle, while different regions mean different instances. Our motivation is that jointly estimating segmentation and boundary map in a multitask network with such a joint loss can offer a better segmentation result at the instance level for aerial images. Note that we do not make use of any postprocessing operations, such as fusing the segmentation and boundary map, as we want to directly evaluate the performance of this network architecture.

Note that the multitask learning network is optimized in an end-to-end fashion. This joint multitask training procedure has several merits. First, in the application of vehicle instance segmentation, the multitask learning network architecture is able to provide the complementary semantic boundary information, which is helpful in differentiating the clustered vehicles, improving the instance-level segmentation performance. Second, the discriminative capability of the network's intermediate feature representations can be improved by this architecture because of multiple regularizations on correlated tasks. Therefore, it can increase the robustness of instance segmentation performance.

III. Experimental Results and Discussion

A. Data Sets

1) ISPRS Potsdam: The ISPRS Potsdam Semantic Labeling data set [42] is an open benchmark data set provided online.[3] The data set consists of 38 orthorectified aerial IRRGB images (6000×6000 pixels) with a 5-cm spatial resolution and the corresponding DSMs generated by dense image matching, taken over the city of Potsdam, Germany. A comprehensive manually annotated pixelwise segmentation mask is provided as the ground truth for 24 tiles that are available for training and validation. The other 14 remain unreleased and are kept with the challenge organizers for testing purposes. We randomly selected five tiles (image number: 2_12, 5_12, 7_7, 7_8, 7_9) from 24 training images and used them as the test set in our experiments (see Fig. 4). The resolution is downsampled to 15 cm/pixels to match the subsequent video data set. The input to the networks contains only red, green, and blue channels, and all the results reported on this data set refer to the aforementioned test set. Table I provides the details about this data set for our experiments.

Fig. 4. Image #5_12 from the ISPRS Potsdam data set for vehicle instance segmentation as well as three zoomed-in areas.

2) Busy Parking Lot: The task of vehicle instance segmentation currently lacks a compelling and challenging benchmark data set to produce quantitative measurements and to compare with other approaches. While the ISPRS Potsdam data set has clearly boosted research in semantic segmentation of high-resolution aerial imagery, it is not as challenging as certain practical scenes, such as a busy parking lot, where vehicles are often parked so close that it is quite hard to separate them, particularly from an aerial view. To this end, in this paper, we propose our new challenging Busy Parking Lot UAV Video data set that we built for the vehicle instance segmentation task. The UAV video was acquired by a camera onboard, a UAV covering the parking lot of Woburn Mall, Woburn, MA, USA.[4] The video comprises 1920×1080 pixels with a spatial resolution of about 15 cm per pixel at 24 frames/s and a length of 60 s. We have manually annotated pixelwise instance segmentation masks for 5 frames (at 1, 15, 30, 45, and 59 s), i.e., the annotation is dense in space and sparse in time to allow for the evaluation of methods with this long sequence (see Fig. 6). The Busy Parking Lot data set is challenging because it presents a high range of variations with a diversity of vehicle colors, the effects of shadow, several slightly blurred regions, and vehicles that are parked too close. We train the networks on the ISPRS Potsdam data set and then perform vehicle instance segmentation using the trained networks on this video data set. Details regarding this data set are shown in Table II.

B. Training Details

The network training is based on the TensorFlow framework. We choose Nesterov Adam [43], [44] as the optimizer to train the network, since for this task, it shows much faster

[3] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[4] https://www.youtube.com/watch?v=yojapmOkIfg

TABLE I

VEHICLE COUNTS AND NUMBER OF VEHICLE PIXELS IN THE ISPRS POTSDAM DATA SET

	Training Set	Test Set				
		2_12	5_12	7_7	7_8	7_9
Vehicle Count	4,433	123	427	301	309	305
Number of Pixels	1,184,789	36,236	122,332	76,892	77,669	74,404

TABLE II

VEHICLE COUNTS AND NUMBER OF VEHICLE PIXELS IN THE BUSY PARKING LOT UAV VIDEO DATA SET

	Frame@1s	Frame@15s	Frame@30s	Frame@45s	Frame@59s
Vehicle Count	511	492	502	484	479
Number of Pixels	257,462	235,560	240,607	235,448	226,697

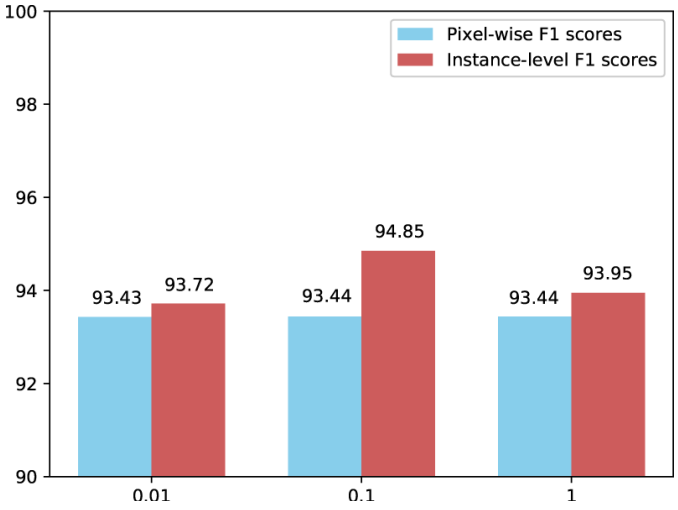


Fig. 5. Sensitivity analysis for the parameter λ on the ISPRS Potsdam data set.

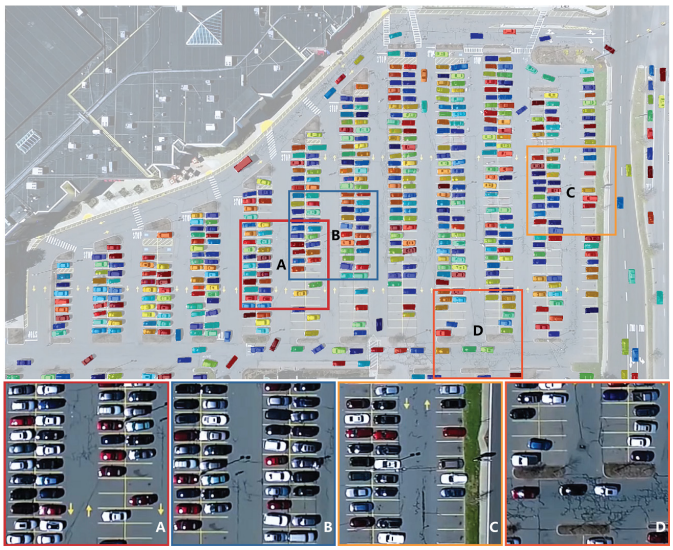


Fig. 6. Frame@1s from the proposed Busy Parking Lot UAV Video data set for vehicle instance segmentation. (Bottom) Four zoomed-in areas.

convergence than the standard stochastic gradient descent with momentum [45] or Adam [46]. We fixed almost all of the parameters of Nesterov Aadam as recommended in [43]: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e{-}08$, and a schedule decay of 0.004, making use of a fairly small learning rate of $2e{-}04$. All weights in the newly added layers are initialized with a Glorot uniform initializer [47] that draws samples from a uniform distribution. In our experiments, we note that the pixelwise F1 score of the network is less sensitive to the parameter λ and the instance-level performance is relatively sensitive to λ. Based on the sensitivity analysis (see Fig. 5), we set it as 0.1.

The networks are trained on the training set of the ISPRS Potsdam data set to predict instance segmentation maps. The training set has only 931 unique 256×256 patches. We make use of the data augmentation technique to increase the number of training samples. The RGB patches and the corresponding pixelwise ground truth are transformed by horizontally and vertically flipping three quarters of the patches. By doing so, the number of training samples increases to 14 896. To monitor overfitting during training, we randomly select 10% of the training samples as the validation set, i.e., splitting the training set into 13 406 training and 1490 validation pairs. We train the network for 50 epochs and make use of early stopping to avoid overfitting. Moreover, we use fairly small mini-batches of eight image pairs because, in a sense, every pixel is a training sample. We train our network on a single NVIDIA GeForce GTX TITAN with 12 GB of GPU memory, which takes about 2 h.

C. Qualitative Evaluation

Some vehicle instance segmentation results are shown in Fig. 7 (test set of the ISPRS Potsdam data set) and Fig. 9 (the Busy Parking Lot data set), respectively, in order to qualitatively illustrate the efficacy of our model. First, we compare various CNN variants used for FCN architecture to determine which one is the best suited for our task. In Fig. 7, we qualitatively investigate the accuracy of the predicted instance segmentation maps using FCN architecture with leading CNN variants, namely, VGG-FCN [33], Inception-FCN [48], Xception-FCN [49], and ResFCN on the ISPRS Potsdam data set. We implement VGG-FCN, Inception-FCN, and Xception-FCN by fusing the output feature maps of the last three convolutional blocks as we do for ResFCN (see Section II-B). From the segmentation results, we can see an improvement in quality from VGG-FCN to ResFCN. Moreover, on the Busy Parking Lot data set, the ResFCN also demonstrates a fairly strong ability to generalize to an "unseen" scene outside the training data set (see Fig. 9). However, there are some vehicles that cannot be separated in

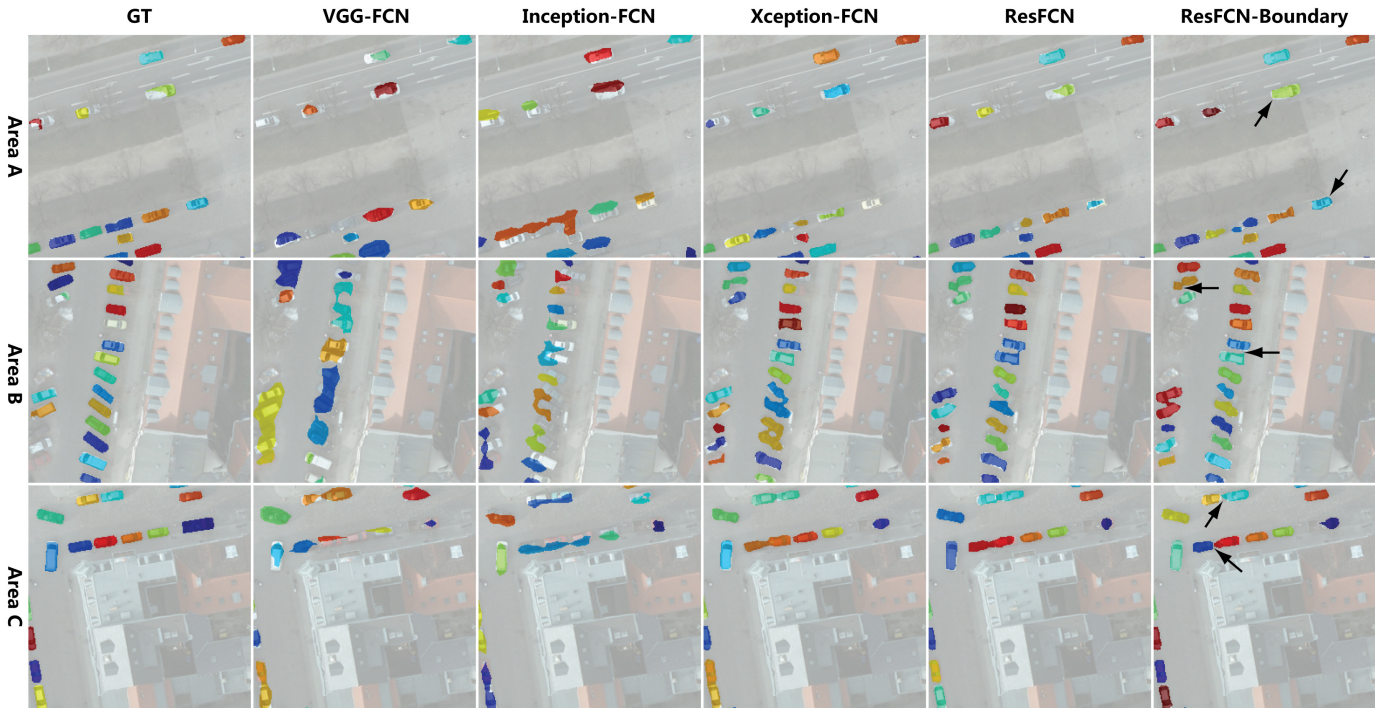


Fig. 7. Instance segmentation results of the ISPRS Potsdam data set. (From left to right) Ground truth, VGG-FCN, Inception-FCN, Xception-FCN, ResFCN, and B-ResFCN (different colors denote the individual vehicle objects). The three areas are derived from Fig. 4.

TABLE III
PIXEL-LEVEL OAS AND F1-SCORES FOR THE CAR CLASS ON THE ISPRS POTSDAM DATA SET

Model	OA	OA (eroded)	F1 score	F1 score (eroded)
ResFCN	99.79	99.89	93.43	95.66
B-ResFCN	99.79	99.89	93.44	95.87

both segmentation results produced using the aforementioned networks due to the extremely close vehicle distance. The situation is further deteriorated when the imagery suffers from the effects of shadow, as the cases shown in the zoomed-in areas of Fig. 9. On the other hand, to identify the role of the semantic boundary component of the proposed unified multitask learning network architecture, we also performed an ablation study to compare the performance of networks relying on the prediction of vehicles. In comparison with the ResFCN, the semantic B-ResFCN is able to separate those "touching" cars clearly, which qualitatively highlights the superiority of a semantic boundary-aware network by exploring the complementary information under a unified multitask learning network architecture. Fig. 8 shows a couple of example segmentations using the proposed B-ResFCN on several frames of the Busy Parking Lot data set.

D. Quantitative Evaluation

To verify the effectiveness of networks used, we reported the pixel-level OAs and F1 scores of the car class on our test set of the ISPRS Potsdam data set in Table III and compared with the state-of-the-art methods. These metrics are calculated on a full reference and an alternative ground truth obtained by eroding the boundaries of objects by a circular disk of 3 pixel radius. The current state-of-the-art CASIA2 (in the leaderboard http://www2.isprs.org/potsdam-2d-semantic-labeling.html) obtains the F1 score of 96.2% for the vehicle segmentation on the held-out test set (which is different from the validation set we use) using IRRG. Our B-ResFCN is competitive with the F1 score of 95.87% obtained by using the RGB information only on our own test set. This indicates that the trained network can be though as a good, competitive model for the follow-up experiments. Note that the pixelwise OA and F1 score can only evaluate the segmentation performance at a pixel level instead of instance level. Therefore, they are actually not suitable for our task.

To quantitatively evaluate the performance of different approaches for vehicle segmentation at the instance level, the evaluation criteria we use are instance-level F1 score, precision, recall, and Dice similarity coefficient. The first three criteria consider the performance of vehicle detection, and the last validates the performance of the instance-level segmentation.

1) Detection: For the vehicle detection evaluation, the metric instance-level F1 score[5] is employed, which is the harmonic mean of instance-level precision P and recall R, defined as

$$F1 = \frac{2PR}{P+R}, \quad P = \frac{N_{\text{tp}}}{N_{\text{tp}} + N_{\text{fp}}}, \quad R = \frac{N_{\text{tp}}}{N_{\text{tp}} + N_{\text{fn}}} \tag{7}$$

where N_{tp}, N_{fp}, and N_{fn} are the number of true positives, false positives, and false negatives, respectively. Here, the ground truth for each segmented vehicle is the object in the manually

[5]Note that the instance-level F1 score is different from the pixelwise F1 score used by the ISPRS semantic labeling evaluation (http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html).

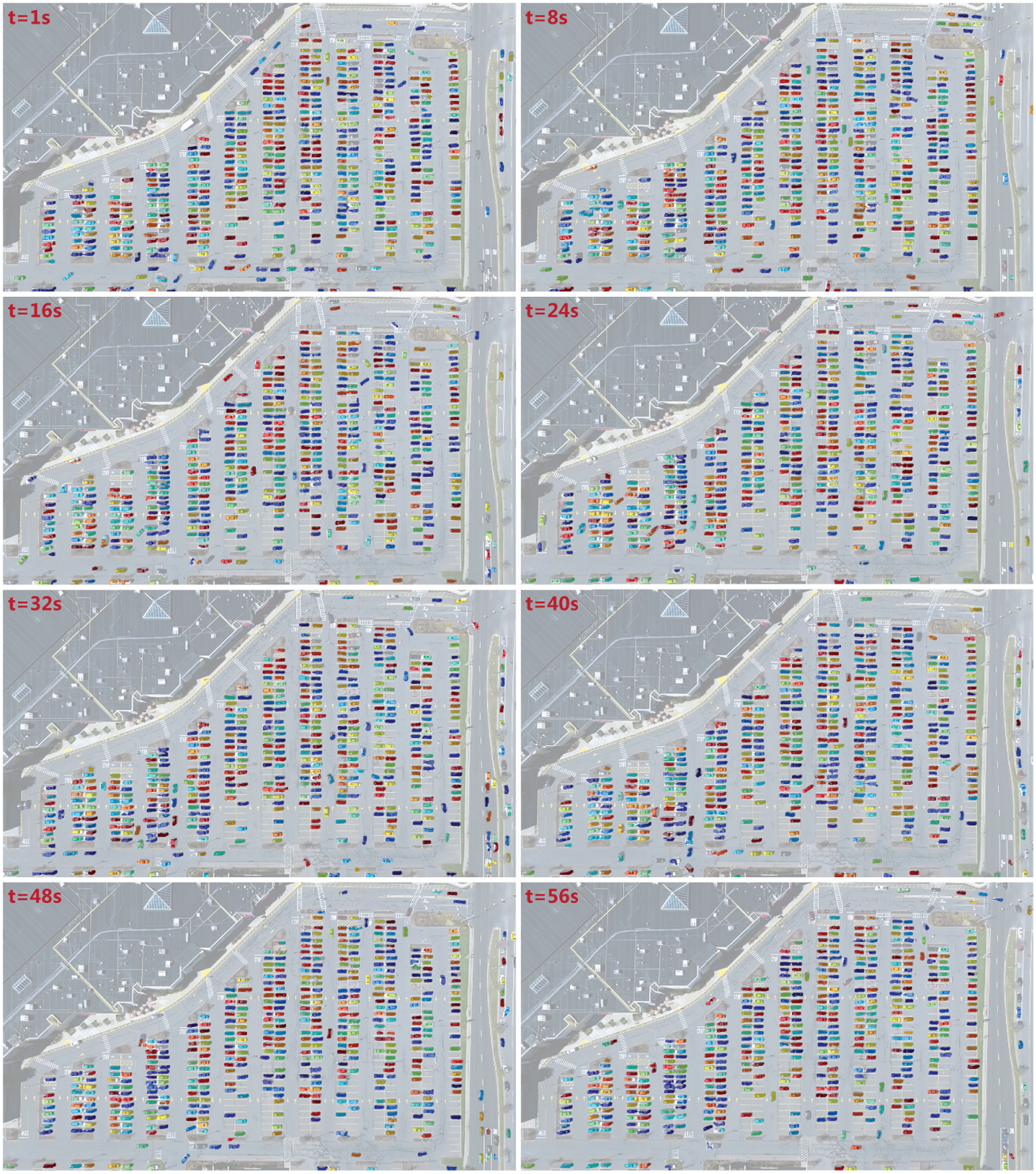


Fig. 8. Example segmentations using the proposed B-ResFCN in several frames of the Busy Parking Lot data set.

labeled segmentation mask that has a maximum overlap with the segmented vehicle. When calculating N_{tp} and N_{fp}, a segmented vehicle that intersects with at least 50% of its ground truth is considered as a true positive; otherwise, it is regarded as a false positive. For N_{fn}, a false negative indicates a ground-truth object that has less than 50% of its area overlapped by its corresponding segmented vehicle or has no corresponding segmented vehicle.

The detection results of different networks on the ISPRS Potsdam data set and the Busy Parking Lot scene are shown in Tables IV and V, respectively. Among the networks without a semantic boundary component, the ResFCN surpasses all other models (VGG-FCN, Inception-FCN, and Xception-FCN), highlighting the strength of residual learning-based FCN architecture with the multilevel contextual feature representations in our task. The network with the semantic boundary

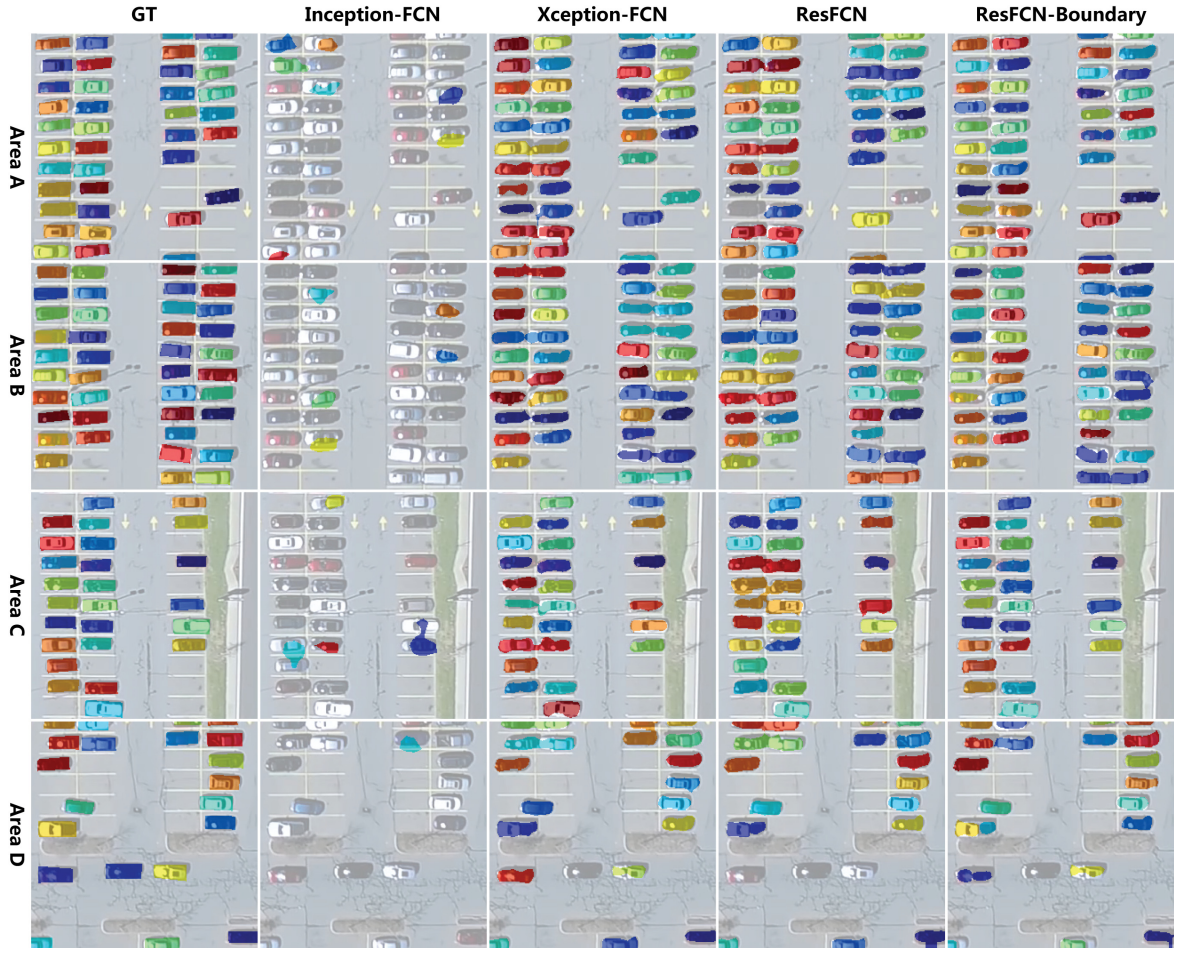


Fig. 9. Instance segmentation maps of the Busy Parking Lot data set. (From left to right) Ground truth, Inception-FCN, Xception-FCN, ResFCN, and B-ResFCN (different colors denote the individual vehicles). The four areas are derived from Fig. 6.

TABLE IV

DETECTION RESULTS OF DIFFERENT NETWORKS ON THE ISPRS POTSDAM SEMANTIC LABELING DATA SET (INSTANCE-LEVEL F1 SCORE, PRECISION, AND RECALL)

Model	2_12			5_12			7_7			7_8			7_9		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
VGG-FCN	66.04	70.00	62.50	57.00	61.45	53.14	59.21	61.95	56.70	57.21	66.84	50.00	61.31	65.91	57.31
B-VGG-FCN	70.27	68.42	72.22	69.85	67.42	72.47	71.03	68.47	73.79	67.96	66.86	69.09	66.47	60.96	73.08
Inception-FCN	51.91	55.45	48.80	31.65	37.42	27.42	40.00	43.41	37.08	27.79	31.70	24.74	40.87	45.02	37.42
B-Inception-FCN	55.15	50.61	60.58	46.14	47.42	44.92	53.81	52.91	54.75	43.47	42.45	44.54	50.74	47.49	54.47
Xception-FCN	96.92	98.21	95.65	83.55	81.11	86.14	93.33	94.59	92.11	92.05	93.10	91.01	93.92	96.59	91.40
B-Xception-FCN	97.00	**100**	94.17	88.40	**88.60**	88.19	93.65	96.47	91.00	93.58	97.54	89.94	94.63	97.50	91.92
ResFCN	97.93	**100**	95.93	83.88	80.84	87.15	94.72	96.86	92.67	**95.62**	**97.93**	**93.42**	95.25	96.23	**94.30**
B-ResFCN	**98.31**	**100**	**96.67**	**88.57**	87.08	**90.11**	**96.43**	**97.12**	**95.74**	95.19	97.88	92.64	**95.76**	**97.83**	93.77

TABLE V

DETECTION RESULTS OF DIFFERENT METHODS ON THE PROPOSED BUSY PARKING LOT UAV VIDEO DATA SET (INSTANCE-LEVEL F1 SCORE, PRECISION, AND RECALL)

Model	Frame@1s			Frame@15s			Frame@30s			Frame@45s			Frame@59s		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Inception-FCN	15.48	60.00	8.89	15.67	51.09	9.25	13.92	43.43	8.29	11.56	41.98	6.71	7.75	39.29	4.30
B-Inception-FCN	17.74	62.50	10.34	19.84	58.72	11.94	18.71	51.69	11.42	17.84	55.34	10.63	10.63	51.67	5.93
Xception-FCN	87.25	86.82	87.69	87.27	85.28	89.36	86.58	84.14	89.16	87.10	84.82	89.50	75.65	74.12	77.25
B-Xception-FCN	91.43	89.72	**93.20**	90.15	86.80	**93.78**	90.12	87.69	92.70	90.35	87.64	**93.22**	88.30	84.24	92.77
ResFCN	88.73	89.71	87.77	89.43	89.76	89.10	90.43	91.38	89.50	88.81	88.69	88.92	87.10	90.23	84.17
B-ResFCN	**93.29**	**95.16**	91.50	**92.55**	**91.52**	93.61	**93.62**	**94.02**	**93.22**	**93.06**	**94.33**	91.83	**94.54**	**95.28**	**93.81**

component—i.e., B-ResFCN—achieved the best results on most test images of the ISPRS Potsdam scene and surpassed the others by a significant margin on the Busy Parking Lot data set, demonstrating the effectiveness of the semantic boundary-aware multitask learning network in this instance segmentation problem. From Tables IV and V, we observe that all the

TABLE VI
SEGMENTATION RESULTS OF DIFFERENT METHODS ON THE BUSY PARKING LOT UAV VIDEO DATA SET (INSTANCE-LEVEL DICE SIMILARITY COEFFICIENT)

Model	Frame@1s	Frame@15s	Frame@30s	Frame@45s	Frame@59s
Inception-FCN	26.81	26.06	25.68	22.89	23.77
B-Inception-FCN	32.37	33.07	33.34	30.44	31.26
Xception-FCN	72.74	72.74	72.85	72.47	71.31
B-Xception-FCN	77.31	**77.50**	77.22	77.13	76.32
ResFCN	71.17	71.47	71.76	68.82	72.73
B-ResFCN	**78.84**	77.33	**79.13**	**77.83**	**79.39**

networks yield a fairly lower instance-level F1, precision, and recall on the Busy Parking Lot data set than on the ISPRS Potsdam data set. This mainly comes from the different difficulty levels of the two data sets. Specifically, high-density parking, strong light conditions, critical effects of shadow, and a slightly blurry image quality lead to the fact that networks have achieved a more inferior performance on the proposed data set than on the Potsdam scene.

2) Segmentation: The Dice similarity coefficient is often used to evaluate a segmentation performance. Given a set of pixels $\boldsymbol{V}$ denoted as a segmented vehicle and a set of pixels $\boldsymbol{G}$ annotated as a ground-truth object, the Dice similarity coefficient is defined as

$$D(\boldsymbol{V}, \boldsymbol{G}) = 2(|\boldsymbol{V} \cap \boldsymbol{G}|)/(|\boldsymbol{V}| + |\boldsymbol{G}|). \tag{8}$$

However, this is not suitable for segmentation evaluation on individual objects (i.e., instance segmentation). Instead, in this paper, an instance-level Dice similarity coefficient is defined and employed as

$$D_{\text{ins}}(\boldsymbol{V}, \boldsymbol{G}) = \frac{1}{2}\left[\sum_{i=1}^{N_V} \omega_i D(\boldsymbol{V}_i, \boldsymbol{G}_i) + \sum_{j=1}^{N_G} \tilde{\omega}_j D(\tilde{\boldsymbol{V}}_j, \tilde{\boldsymbol{G}}_j)\right] \tag{9}$$

where $\boldsymbol{V}_i$, $\boldsymbol{G}_i$, $\tilde{\boldsymbol{G}}_j$, and $\tilde{\boldsymbol{V}}_j$ are the ith segmented vehicle, the ground-truth object that maximally overlaps $\boldsymbol{V}_i$, the jth ground-truth object, and the segmented vehicle that maximally overlaps $\tilde{\boldsymbol{G}}_j$, respectively. N_V and N_G denote the total number of segmented vehicles and ground-truth objects, respectively. Furthermore, ω_i and $\tilde{\omega}_j$ are coefficients and can be calculated as

$$\omega_i = \frac{|\boldsymbol{V}_i|}{\sum_{k=1}^{N_V} |\boldsymbol{V}_k|}, \quad \tilde{\omega}_j = \frac{|\tilde{\boldsymbol{G}}_j|}{\sum_{k=1}^{N_G} |\tilde{\boldsymbol{G}}_k|}. \tag{10}$$

Tables VI and VII show the segmentation results of different approaches on the Potsdam scene and Busy Parking Lot data set, respectively. We can see that our B-ResFCN achieves the best performance on these two data sets. Compared with the ResFCN, there is a 1.16% increment in terms of the instance-level Dice similarity coefficient on the Potsdam data set and a 7.31% improvement on the Busy Parking Lot scene. From the figures in Tables VI and VII, we can see that the networks offer a more inferior performance on the Busy Parking Lot data set than on the Potsdam scene. This is also in line with our intention of proposing a more challenging benchmark data set for the vehicle instance segmentation problem. In addition, it is worth noting that basically all the networks with boundary components can offer better instance segmentations compared with those without boundary. This means that multitask learning is useful for different CNN variants in our task.

TABLE VII
SEGMENTATION RESULTS OF DIFFERENT METHODS ON THE ISPRS POTSDAM SEMANTIC LABELING DATA SET (INSTANCE-LEVEL DICE SIMILARITY COEFFICIENT)

Model	2_12	5_12	7_7	7_8	7_9
VGG-FCN	58.88	45.79	53.13	51.09	54.25
B-VGG-FCN	71.48	64.48	74.54	70.43	69.47
Inception-FCN	52.79	34.37	37.15	35.08	44.22
B-Inception-FCN	55.26	35.69	46.76	37.33	47.14
Xception-FCN	90.05	73.05	84.84	84.58	86.54
B-Xception-FCN	91.44	75.47	85.12	88.64	87.95
ResFCN	91.97	77.68	89.10	89.78	89.65
B-ResFCN	**93.80**	**77.72**	**90.61**	**91.19**	**90.66**

IV. CONCLUSION

In this paper, we propose a semantic boundary-aware unified multitask learning ResFCN in order to handle a novel problem (i.e., vehicle instance segmentation). In particular, the proposed network harnesses the multilevel contextual features learned from different residual blocks in a residual network architecture to produce better pixelwise likelihood maps. We theoretically analyze the reason behind this. Furthermore, our network creates two separate, yet identical branches to simultaneously predict the semantic segmentation masks of vehicles and semantic boundaries. The joint learning of these two problems is beneficial for separating "touching" vehicles which are often not correctly differentiated into instances. The network is validated using a large high-resolution aerial image data set, ISPRS Potsdam Semantic Labeling data set, and the proposed Busy Parking Lot UAV Video data set. To quantitatively evaluate the performance of different approaches for the vehicle instance segmentation, we advocate using an instance-level F1 score, precision, recall, and Dice similarity coefficient as evaluation criteria, instead of traditional pixelwise OA and F1 score for semantic segmentation. Both visual and quantitative analyses of the experimental results demonstrate the effectiveness of our approach.

ACKNOWLEDGMENT

The authors would like to thank the ISPRS for making the Potsdam data set available.

REFERENCES

[1] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2016.

[2] L. Mou *et al.*, "Multitemporal very high resolution from space: Outcome of the 2016 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3435–3447, Aug. 2017.

[3] N. Audebert, B. Le Saux, and S. Lefèvre, "Fusion of heterogeneous data in convolutional networks for urban semantic labeling," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, Mar. 2017, pp. 1–4.

[4] L. Mou and X. X. Zhu. (2018). "RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images." [Online]. Available: https://arxiv.org/abs/1805.02091

[5] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Graph-based registration, change detection, and classification in very high resolution multitemporal remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 7, pp. 2940–2951, Jul. 2016.

[6] D. Wen, X. Huang, L. Zhang, and J. A. Benediktsson, "A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 609–625, Jan. 2016.

[7] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," *Remote Sens. Environ.*, vol. 199, pp. 241–255, Sep. 2017.

[8] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sens.*, vol. 8, no. 6, p. 506, 2016.

[9] L. Mou and X. X. Zhu, "Spatiotemporal scene interpretation of space videos via deep neural network and tracklet analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1823–1826.

[10] G. Kopsiaftis and K. Karantzalos, "Vehicle detection and traffic density monitoring from very high resolution satellite video data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 1881–1884.

[11] W. Shao, W. Yang, G. Liu, and J. Liu, "Car detection from high-resolution aerial imagery using multiple features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2012, pp. 4379–4382.

[12] T. Moranduzzo and F. Melgani, "Automatic car counting method for unmanned aerial vehicle images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1635–1647, Mar. 2014.

[13] T. Moranduzzo and F. Melgani, "Detecting cars in UAV images with a catalog-based approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6356–6367, Oct. 2014.

[14] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1938–1942, Sep. 2015.

[15] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[16] L. Mou and X. X. Zhu. (2018). "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network." [Online]. Available: https://arxiv.org/abs/1802.10249

[17] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–spatial feature learning via deep residual conv–deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.

[18] L. Mou, L. Bruzzone, and X. X. Zhu. (2018). "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery." [Online]. Available: https://arxiv.org/abs/1803.02642

[19] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[20] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.

[21] N. Ammour, H. Alhichri, Y. Bazi, B. Benjdira, N. Alajlan, and M. Zuair, "Deep learning approach for car detection in UAV imagery," *Remote Sens.*, vol. 9, no. 4, p. 312, 2017.

[22] N. Audebert, B. Le Saux, and S. Lefèvre, "Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images," *Remote Sens.*, vol. 9, no. 4, p. 368, 2017.

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla. (2015). "SegNet: A deep convolutional encoder-decoder architecture for image segmentation." [Online]. Available: https://arxiv.org/abs/1511.00561

[24] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshop*, Jun. 2016.

[25] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2015, pp. 2650–2658.

[26] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 1–10.

[27] Z. Wu, C. Shen, and A. van den Hengel. (2016). "High-performance semantic segmentation using very deep fully convolutional networks." [Online]. Available: https://arxiv.org/abs/1604.04339

[28] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 239–248.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–9.

[30] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2016). "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1606.00915

[32] S. Zhang *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1–9.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. IEEE Int. Conf. Learn. Represent. (ICLR)*, Sep. 2015, pp. 1–14.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 630–645.

[35] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 1–9.

[36] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 519–534.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[38] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 818–833.

[39] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, p. 1.

[40] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother, "Instancecut: from edges to instances with multicut," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 1–10.

[41] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[42] F. Rottensteiner *et al.*, "The ISPRS benchmark on urban object classification and 3D building reconstruction," *ISPRS Ann. Photogramm., Remote Sensing Spatial Inf. Sci.*, vol. 1, no. 3, pp. 293–298, 2012.

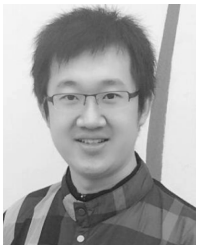
[43] T. Dozat, *Incorporating Nesterov Momentum into Adam*. Accessed: Jun. 26, 2018. [Online]. Available: http://cs229.stanford.edu/proj2015/054_report.pdf

[44] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 1–9.

[45] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[46] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[47] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–6.
[49] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 1–8.



Lichao Mou (S'16) received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, and the master's degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2015. He is currently pursuing the Ph.D. degree with the German Aerospace Center, Wessling, Germany, and the Technical University of Munich, Munich, Germany.

In 2015, he was with the Computer Vision Group, University of Freiburg, Breisgau, Germany, for six months. His research interests include remote sensing, computer vision, and machine/deep learning, especially remote sensing video analysis and deep networks with their applications in remote sensing.

Mr. Mou was a recipient of the First Place at the 2016 IEEE GRSS Data Fusion Contest and a finalist for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event.

Xiao Xiang Zhu (S'10–M'12–SM'14) received the M.Sc., Dr.Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, in 2009, Fudan University, Shanghai, China, in 2014, The University of Tokyo, Tokyo, Japan, in 2015, and the University of California at Los Angeles, Los Angeles, CA, USA, in 2016. She is currently the Professor for Signal Processing in Earth Observation (SiPEO), TUM, and the German Aerospace Center (DLR), Wessling, Germany, and the Head of the Department of EO Data Science, Earth Observation Center, DLR, and the Helmholtz Young Investigator Group (SiPEO), DLR, and TUM. Her research interests include remote sensing and earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.