Relation Matters: Relational Context-Aware Fully Convolutional Network for Semantic Segmentation of High Resolution Aerial Images

Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*—Most current semantic segmentation approaches fall back on deep Convolutional Neural Networks (CNNs). However, their use of convolution operations with local receptive fields causes failures in modeling contextual spatial relations. Prior works have sought to address this issue by using graphical models or spatial propagation modules in networks. But such models often fail to capture long-range spatial relationships between entities, which leads to spatially fragmented predictions. Moreover, recent works have demonstrated that channel-wise information also acts a pivotal part in CNNs. In this work, we introduce two simple yet effective network units, the spatial relation module and the channel relation module, to learn and reason about global relationships between any two spatial positions or feature maps, and then produce relation-augmented feature representations. The spatial and channel relation modules are general and extensible, and can be used in a plug-and-play fashion with the existing Fully Convolutional Network (FCN) framework. We evaluate relation module-equipped networks on semantic segmentation tasks using two aerial image datasets, namely ISPRS Vaihingen and Potsdam datasets, which fundamentally depend on long-range spatial relational reasoning. The networks achieve very competitive results, a mean F_1 score of 88.54% on the Vaihingen dataset and a mean F_1 score of 88.01% on the Potsdam dataset, bringing significant improvements over baselines.**

***Index Terms*—Relation network, fully convolutional network, semantic segmentation, high resolution aerial imagery.**

I. Introduction

THE widely availability of aeroplanes and Unmanned Aerial Vehicles (UAVs) has generated huge volume of high resolution aerial images. Automatic parsing of such images is a task of primary importance for a plethora of applications, to name a few, urban and traffic monitoring [1], [2], [3], [4], [5], [6], [7], map updates [8], [9], and disaster relief operations [10], [11]. One crucial step towards understanding an aerial image is to perform semantic segmentation.

This work is jointly supported by the China Scholarship Council, the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No [ERC-2016-StG-714087], Acronym: *So2Sat*), and Helmholtz Association under the framework of the Young Investigators Group "SiPEO" (VH-NG-1018, www.sipeo.bgu.tum.de), Helmholtz Artificial Intelligence Cooperation Unit (HAICU) - Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation - Big Data Fusion for Urban Research". The first two authors contributed equally to this work. *(Corresponding author: Xiao Xiang Zhu.)*

L. Mou, Y. Hua, and X. X. Zhu 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; yuansheng.hua@dlr.de; xiaoxiang.zhu@dlr.de).

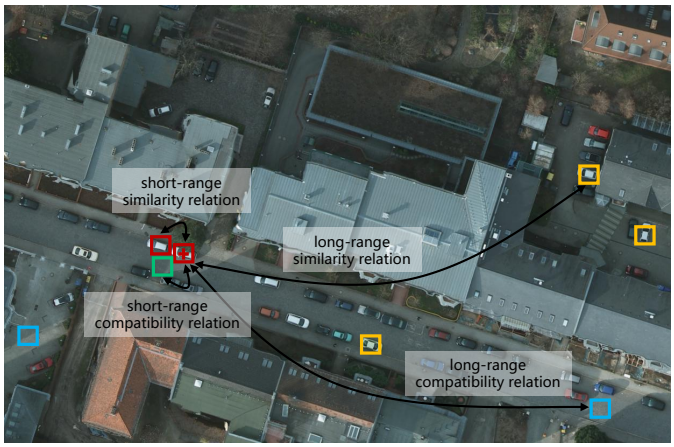


Fig. 1: Illustration of long-range spatial relations in an aerial image. Appearance similarity or semantic compatibility between patches within a local region (red–red and red–green) and patches in remote regions (red–yellow and red–blue) underlines our global relation modeling.

Semantic segmentation of an image refers to the challenging task of inferring every pixel in the image with the semantic category of the object to which it belongs. The emergence of deep convolutional neural networks (CNNs) [12], [13], [14], [15], [16], [17], [18] and vast training data has led to significant progress in this direction. However, although with more training data and with deeper and more complicated network architectures, there is a technical hurdle in the application of CNNs to semantic image segmentation—contextual information.

It has been well recognized in computer vision for years that contextual information, or *relation*, is capable of offering important cues for semantic segmentation tasks [19], [20], [21], [22], [23], [24]. For instance, given two regions in an image, their relation can be semantic similarity. In addition, spatial relations also involve compatibility and incompatibility relationships, e.g., a vehicle is likely to be driven or parked on pavements, and a piece of lawn is unlikely to appear on the roof of a building. Unfortunately, only convolution layers cannot model such spatial relations due to their local valid receptive field[1].

[1]Feature maps from deep CNNs like ResNet usually have large receptive fields due to deep architectures, whereas the study of [25] has shown that CNNs are apt to extract information mainly from much smaller regions in receptive fields, which are called valid receptive fields.

(a) (b) (c) (d)

Fig. 2: Illustration of different challenges in high resolution aerial images for semantic segmentation tasks: severe visual ambiguities caused by (a) shadows and (b) tree branches; big appearance variations within (c) roofs and (d) roads.

Nevertheless, under some circumstances, spatial relations are of paramount importance, particularly when a region in an image exhibits significant visual ambiguities. To address this issue, several attempts have been made to introduce spatial relations into networks by using either graphical models [26], [27], [28] or spatial propagation networks [29], [30], [31]. However, these methods seek to capture global spatial relations implicitly with a chain propagation way, whose effectiveness depends heavily on the learning effect of long-term memorization. Consequently, these models may not work well in some cases like aerial scenes (see Figure 6 and Figure 7), in which long-range spatial relations often exist (cf. Figure 1). Hence, explicit modeling of long-range relations may provide additional crucial information but still remains underexplored for semantic segmentation.

This work is inspired by the recent success of relation networks in visual question answering [32], object detection [33], and activity recognition in videos [34]. Being able to reason about relationships between entities is momentous for intelligent decision-making. A relation network is capable of inferring relationships between an individual entity (e.g., a patch in an image) and a set of other entities (e.g., all patches in the image) by agglomerating information. The relations vary at both long-range and short-range scales and are learned automatically, driven by tasks. Moreover, a relation network can model dependencies between entities, without making excessive assumptions on their feature distributions and locations.

In this work, our goal is to increase the representation capacity of a Fully Convolutional Network (FCN) for semantic segmentation in aerial scenes by using relation modules to describe relationships between observations in convolved images and produce relation-augmented feature representations. Given that convolutions operate by blending spatial and cross-channel information together, we capture relations in both spatial and channel domains. More specifically, two plug-and-play modules—a spatial relation module and a channel relation module—are appended on top of feature maps of an FCN to learn different aspects of relations and then generate spatial relation-augmented and channel relation-augmented features, respectively, for semantic segmentation. By doing so, relationships between any two spatial positions or feature maps can be modeled and used to further enhance feature representations. Furthermore, we study empirically two ways of integrating two relation modules—serial and parallel. This work's contributions are threefold:

- We propose a simple yet effective and interpretable relational context-aware network that enables spatial and channel relational reasoning. Learning such a relation network for semantic segmentation of aerial images has not been investigated yet to the best of our knowledge.
- A spatial relation module and a channel relation module are devised to explicitly model global relations, which are subsequently harnessed to produce spatial- and channel-augmented features.
- We validate the effectiveness of our relation modules through extensive ablation studies. Moreover, to figure out what the spatial relation module has learned, we study a pure spatial relation network in Section IV-F, which shows results beyond expected.

This paper is organized as follows. After the introductory Section I, Section II details relevant semantic segmentation methods and relation networks. Section III is dedicated to describe details of the proposed network. The experimental results are provided in Section IV. Finally, Section V concludes the paper.

II. Related Work

A. Semantic Segmentation of Aerial Imagery

In comparison with natural images in computer vision and hyper- and multi-spectral data in remote sensing, aerial images at high spatial resolution (GSD 5-30 cm) have pretty different characteristics, bringing challenges for semantic segmentation tasks. For example, intricate spatial details (e.g., roof-top pipes, tiny windows, and tiles, road markings, branches of trees, and windows of vehicles) result in big differences in visual appearance within an object category. Moreover, shadows of buildings lead to serious visual ambiguities. Figure 2 shows some of the challenges. Earlier studies [35], [36] have focused on extracting useful low-level, hand-crafted visual features and/or modeling mid-level semantic features on local portions of images (e.g., patches and superpixels); subsequently, a supervised classifier is employed to learn a mapping from the features to semantic categories.

Recent efforts employ deep CNNs and have made a great leap towards end-to-end aerial image parsing [37], [38], [39], [40], [41] [2] and classification [42], [43], [44]. In [45], authors use an FCN trained on the ImageNet dataset as a pre-trained model, which is then fine-tuned on high resolution aerial images for semantic segmentation tasks. To make use of both color images and Digital Surface Model (DSM) data as input, while respecting their different statistical properties, the authors of [46] employ a late fusion approach with two structurally identical, parallel FCNs. In [47], the authors focus on small object (e.g., car) segmentation through quantifying the uncertainty at a pixel level for FCNs. By doing so, they can obtain high overall accuracy and, at the same time, still achieve good accuracy for small objects. Recently, authors in [48] introduce a MultiLayer Perceptron (MLP) on the top of a

[2]This paper is an extension of [39].

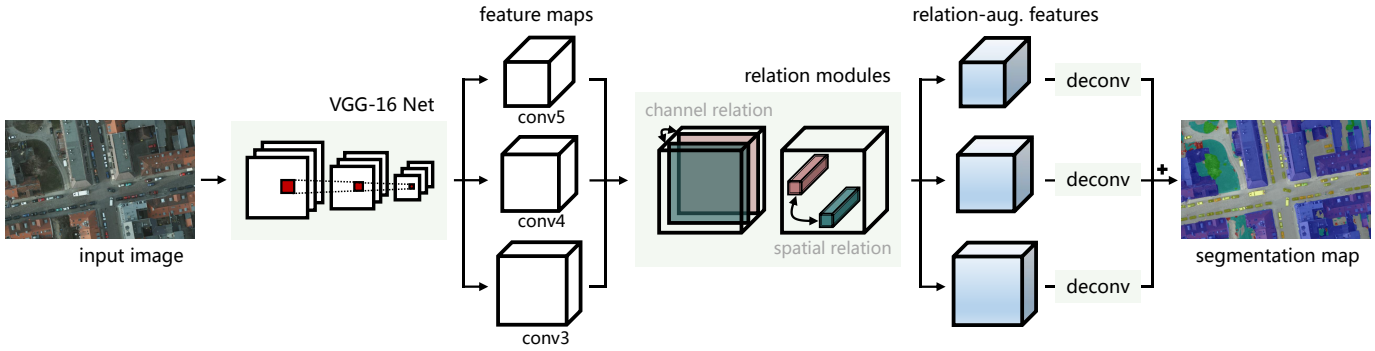


Fig. 3: An overview of the relation module-equipped fully convolutional network.

base FCN to learn how to effectively combine intermediate features to offer a better segmentation result. In [49], the authors investigate the use of another network architecture, SegNet [**?**], [50], for semantic segmentation of aerial images. In addition, they use a residual correction to perform data fusion from heterogeneous data (i.e., optical image and DSM). Later, in [51], they systematically study different network architectures for semantic segmentation of multimodal remote sensing data and, more specifically, they find that late fusion makes it possible to recover errors streaming from ambiguous data while early fusion allows for better joint feature learning but at the cost of higher sensitivity to missing data. In [52], a SegNet architecture is compared with a standard CNN performing patch classification for semantic segmentation purposes. The authors of [**?**] propose a segmentation network called Rotation Equivariant vector field Network (RotEqNet), which is able to be equivariant to rotation by encoding rotation in the network. By doing so, this network can be faced with an easier problem, as it does have to learn particular convolutional kernels that can cope with various rotated versions in the same object category. In [53], the authors propose a two-step framework that first trains a CNN to produce multi-scale edge likelihood maps from color-infrared and height data. Then, the object boundaries generated with each source are regarded as an additional channel and added to each source, and an FCN or SegNet is trained for semantic segmentation purposes. The intuition behind this work is that using predicted boundaries helps to achieve sharper segmentation maps. Saliency detection aims to segment primary objects with fine-grained boundaries from images using useful visual cues, e.g., color, texture, and location [54], which may be beneficial for semantic segmentation tasks. In addition, there are naturally the domain shift and small sample problems in remote sensing data parsing tasks. In these directions, semantic segmentation of aerial images may benefit from studies in [55], [56].

Moreover, there are numerous contests aiming at semantic segmentation from overhead imagery recently, e.g., Kaggle[3], SpaceNet[4], and DeepGlobal[5].

[3]https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection
[4]https://spacenetchallenge.github.io/
[5]http://deepglobe.org/challenge.html

B. Context-Aware Modeling

There are many graphical model-based methods being used to improve the performance of semantic segmentation [26], [27], [28], [57], [58]. For example, the work in [26] makes use of a CRF as post-processing to refine the final segmentation results. [27] and [28] further make the CRF module differentiable and integrate it as a joint-trained part within networks. Moreover, low-level visual cues, e.g., object contours, have also been considered structure information [59], [60]. These approaches, however, are sensitive to visual appearance changes and expensive due to iterative inference procedures required.

Learning spatial propagation with networks for semantic segmentation have attracted high interests in recent years [29], [30], [31], [61], [62], [63], [64]. In [61], the authors try to predict entities of an affinity matrix directly by learning a CNN, which presents a good performance on image segmentation, while the affinity is followed by a nondifferentiable, independent solver of spectral embedding and cannot be used for end-to-end predictions. The authors of [30] train a CNN model to learn a task-dependent affinity matrix by converting the modeling of affinity to learning a local linear spatial propagation, yielding a simple, yet effective approach for the enhancement of segmentation results. Several recent works [62], [63], [64] focus on the extension of this work. In [29], [31], spatial relations are modeled and reinforced via interlayer propagation. [31] proposes an Inside-Outside Net (ION) where four independent recurrent networks that move in four directions are used to pass information along rows or columns. [29] utilizes four slice-by-slice convolutions within feature maps, enabling message passings between neighboring rows and columns in a layer. The spatial propagation of these methods is serial in nature, and thus each position could only receive information from its neighbors.

Recently, a relational reasoning network has been proposed in [32] for visual question answering with super-human performance. Later, [34] proposes a temporal relation network to enable multi-scale temporal relational reasoning in neural networks for videos. In [33], the authors propose an object relation module, which allows modeling relationships among sets of objects, for object detection tasks. Our work is motivated by the success of these works, but we focus on modeling spatial and channel relations in a CNN for semantic

segmentation.

Unlike graphical model-based [26], [27], [28] and spatial propagation network-based methods [29], [30], [31], [61], [62], [63], [64], we explicitly take spatial relations and channel relations into account, so that semantic image segmentation could benefit from short- and long-range relational reasoning.

III. Our Approach

Unlike graphical model-based and spatial propagation network-based methods, we explicitly take spatial relations and channel relations into account, so that semantic image segmentation could benefit from short- and long-range relational reasoning. In this section, an overview of the proposed relational context-aware network is given to present a comprehensive picture. Afterwards, two key components, the spatial relation module and the channel relation module, are introduced, respectively. Finally, we describe the strategy of integrating these modules for semantic segmentation.

A. Overview

As illustrated in Figure 3, the proposed network takes VGG-16 [65] as a backbone to extract multi-level features. Outputs of *conv3*, *conv4*, and *conv5* are fed into the channel and spatial relation modules (see Figure 4) for generating relation-augmented features. These features are subsequently fed into convolutional layers with 1×1 filters to squash the number of channels to the number of categories. Finally, the convolved feature maps are upsampled to a desired full resolution and element-wise added to generate final segmentation maps.

B. Spatial Relation Module

In order to capture global spatial relations, we employ a spatial relation module (cf. Figure 4), where the spatial relation is defined as a composite function with the following equation:

$$\mathrm{SR}(\boldsymbol{x}_i, \boldsymbol{x}_j) = f_{\phi_s}(g_{\theta_s}(\boldsymbol{x}_i, \boldsymbol{x}_j)) \,. \tag{1}$$

Denote by $\boldsymbol{X} \in \mathbb{R}^{C \times H \times W}$ a random variable representing a set of feature maps. $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are two feature-map vectors and identified by spatial positions indices i and j. The size of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is $C \times 1 \times 1$. To model a compact relationship between these two feature-map vectors, we make use of a dot production operation as g_{θ_s} instead of a multilayer perceptron (MLP), and the latter is commonly used in relational reasoning modules [32], [34]. Particularly, g_{θ_s} is defined as follows:

$$g_{\theta_s}(\boldsymbol{x}_i, \boldsymbol{x}_j) = u_s(\boldsymbol{x}_i)^T v_s(\boldsymbol{x}_j) \,, \tag{2}$$

where $u_s(\boldsymbol{x}_i) = \boldsymbol{W}_{u_s}\boldsymbol{x}_i$ and $v_s(\boldsymbol{x}_j) = \boldsymbol{W}_{v_s}\boldsymbol{x}_j$. $\boldsymbol{W}_{u_s}$ and $\boldsymbol{W}_{v_s}$ are weight matrices and can be learned during the training phase. Considering computational efficiency, we realize Eq. (2) in matrix format with the following steps:

1) Feature maps $\boldsymbol{X}$ are fed into two convolutional layers with 1×1 filters to generate $u_s(\boldsymbol{X})$ and $v_s(\boldsymbol{X})$, respectively.
2) Then $u_s(\boldsymbol{X})$ and $v_s(\boldsymbol{X})$ are reshaped (and transposed) into $HW \times C$ and $C \times HW$, correspondingly.

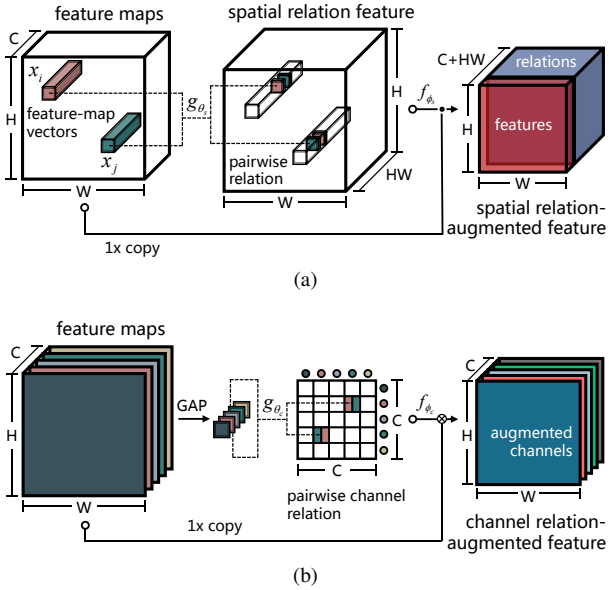


Fig 4: Diagrams of (a) spatial relation module and (b) channel relation module.

3) Eventually, the matrix multiplication of $u_s(\boldsymbol{X})$ and $v_s(\boldsymbol{X})$ is conducted to produce a $HW \times HW$ matrix, which is further reshaped to form a spatial relation feature of size $HW \times H \times W$.

It is worth nothing that the spatial relation feature is not further synthesized (e.g., summed up), as fine-grained contextual characteristics are essential in semantic segmentation tasks. Afterwards, we select the ReLU function as f_{ϕ_s} to eliminate negative spatial relations.

However, relying barely on spatial relations leads to a partial judgment. Therefore, we further blend the spatial relation feature and original feature maps $\boldsymbol{X}$ as follows:

$$\boldsymbol{X}_s = [\boldsymbol{X}, \mathrm{SR}(\boldsymbol{X})] \,. \tag{3}$$

Here we simply use a concatenation operation, i.e., $[\cdot, \cdot]$, to enhance original features with spatial relations. By doing so, output features are abundant in global spatial relations, while high-level semantic features are also preserved.

C. Channel Relation Module

Although the spatial relation module is capable of capturing global contextual dependencies for identifying various objects, misdiagnoses happen when objects share similar distribution patterns but vary in channel dimensionality. In addition, a recent work [66] has shown the benefit of enhancing channel encoding in a CNN for image classification tasks. Therefore, we propose a channel relation module to model channel relations, which can be used to enhance feature discriminabilities in the channel domain. Similar to the spatial relation module, we define the channel relation as a composite function with the following equation:

$$\mathrm{CR}(\boldsymbol{X}_p, \boldsymbol{X}_q) = f_{\phi_c}(g_{\theta_c}(\boldsymbol{X}_p, \boldsymbol{X}_q)) \,, \tag{4}$$

where the input is a set of feature maps $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_C\}$, and $\boldsymbol{X}_p$ as well as $\boldsymbol{X}_q$ represents the

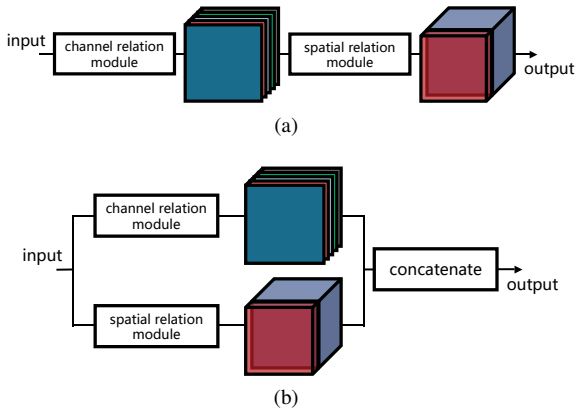


Fig. 5: Two integration manners: (a) serial and (b) parallel.

p-th and the q-th channels of $\boldsymbol{X}$. Dot production is employed to be g_{θ_c}, defined as

$$g_{\theta_c}(\boldsymbol{X}_p, \boldsymbol{X}_q) = u_c(\mathrm{GAP}(\boldsymbol{X}_p))^T v_c(\mathrm{GAP}(\boldsymbol{X}_q)), \quad (5)$$

for capturing global relationships between feature map pairs, where $\mathrm{GAP}(\cdot)$ denotes the global average pooling function. Notably, considering that the preservation of spatial structural information distracts the analysis of channel interdependencies, we adopt averages of $\boldsymbol{X}_p$ and $\boldsymbol{X}_q$ as channel descriptors before performing dot production. More specifically, we feed feature maps into a global average pooling layer for generating a set of channel descriptors of size $C \times 1 \times 1$, and then exploit two convolutional layers with 1×1 filters to produce $u_c(\boldsymbol{X})$ and $v_c(\boldsymbol{X})$, respectively. Afterwards, an outer production is performed to generate a $C \times C$ channel relation feature, where the element located at (p, q) indicates $g_{\theta_c}(\boldsymbol{X}_p, \boldsymbol{X}_q)$.

Furthermore, we emphasize class-relevant channel relations as well as suppress irrelevant channel dependencies by adopting a softmax function as f_{ϕ_c}, formulated as

$$f_{\phi_c}(g_{\theta_c}(\boldsymbol{X}_p, \boldsymbol{X}_q)) = \frac{\exp(g_{\theta_c}(\boldsymbol{X}_p, \boldsymbol{X}_q))}{\sum_{q=1}^{C} \exp(g_{\theta_c}(\boldsymbol{X}_p, \boldsymbol{X}_q))}, \quad (6)$$

where we take $\boldsymbol{X}_p$ as an example. Consequently, a discriminative channel relation map $\mathrm{CR}(\boldsymbol{X})$ can be obtained, where each element represents the corresponding pairwise channel relation.

To integrate $\mathrm{CR}(\boldsymbol{X})$ and original feature maps $\boldsymbol{X}$, we reshape $\boldsymbol{X}$ into a matrix of $C \times HW$ and employ a matrix multiplication as follows:

$$\boldsymbol{X}_c = \boldsymbol{X}^T \mathrm{CR}(\boldsymbol{X}). \quad (7)$$

With this design, the input features are enhanced with channel relations and embedded with not only initial discriminative channel properties but also global inter-channel correlations. Eventually, $\boldsymbol{X}_c$ is reshaped to $C \times H \times W$ and fed into subsequent procedures.

Figure 4 shows the diagram of our channel relation module.

TABLE I: Ablation Study on the Vaihingen Dataset.

Model Name	crm	srm	mean F_1	OA
Baseline FCN[67]			83.74	86.51
RA-FCN-crm	✓		87.24	88.38
RA-FCN-srm		✓	88.36	89.03
P-RA-FCN	✓	✓	88.50	89.18
S-RA-FCN	✓	✓	**88.54**	**89.23**

[1] RA-FCN indicates the proposed relational context-aware FCN.
[2] crm indicates the channel relation module.
[3] srm indicates the spatial relation module.
[4] P-RA-FCN indicates that crm and srm are appended on top of the backbone in parallel.
[5] S-RA-FCN indicates that crm is followed by srm.

D. Integration of Relation Modules

In order to jointly enjoy benefits from spatial and channel relation modules, we further aggregate features $\boldsymbol{X}_s$ and $\boldsymbol{X}_c$ to generate spatial and channel relation-augmented features. As shown in Figure 5, we investigate two integration patterns, namely serial integration and parallel integration, to blend $\boldsymbol{X}_s$ and $\boldsymbol{X}_c$. For the former, we append the spatial relation module to the channel relation module and infer $\boldsymbol{X}_s$ from $\boldsymbol{X}_c$ instead of $\boldsymbol{X}$, as presented in Eq. (1) and Eq. (7). For the latter, spatial relation-augmented features and channel relation-augmented features are obtained simultaneously and then aggregated by performing concatenation. Influences of different strategies are discussed in Section IV-B.

IV. Experiments

To verify the effectiveness of long-range relation modeling in our network, aerial image datasets are used in experiments. This is because aerial images are taken from nadir view, and the spatial distribution/relation of objects in these images is diverse and complicated, as shown in Figure 1. Thus, we perform experiments on two aerial image semantic segmentation datasets, i.e., ISPRS Vaihingen and Potsdam datasets, and results are discussed in subsequent sections.

A. Experimental Setup

1) Datasets: The Vaihingen dataset[6] is composed of 33 aerial images collected over a 1.38 km^2 area of the city, Vaihingen, with a spatial resolution of 9 cm. The average size of each image is 2494×2064 pixels, and each of them has three bands, corresponding to near infrared (NIR), red (R), and green (G) wavelengths. Notably, DSMs, which indicate the height of all object surfaces in an image, are also provided as complementary data. Among these images, 16 of them are manually annotated with pixel-wise labels, and each pixel is classified into one of six land cover classes. Following the setup in [48], [52], [45], we select 11 images for training, and the remaining five images (image IDs: 11, 15, 28, 30, 34) are used to test our model.

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

The Potsdam dataset[7] consists of 38 high resolution aerial images, which covers an area of 3.42 km^2, and each aerial image is captured in four channels (NIR, R, G, and blue (B)). The size of all images is 6000×6000 pixels, which are annotated with pixels-level labels of six classes as the Vaihingen dataset. The spatial resolution is 5 cm, and coregistered DSMs are available as well. Compared to the Vaihingen dataset, this dataset is more challenging owing to its finer spatial resolution (5 cm/pixel vs. 9 cm/pixel) and wider area of coverage. To train and evaluate networks, we utilize 10 images for training and build the test set with the remaining images (image IDs: 02_12, 03_12, 04_12, 05_12, 06_12, 07_12).

Vaihingen and Potsdam datasets are public datasets provided by ISPRS-Commission III. As reported in [68], images were captured using digital aerial cameras carried out by the German Association of Photogrammetry and Remote Sensing (DGPF) [69] and mosaicked with Trimble INPHO OrthoVista. Void areas in DSMs were filled with a variant of nonlinear diffusion [70]. In our experiments, we directly train and evaluate models on these pre-processed images and DSMs.

2) Implementation: The proposed network is initialized with separate strategies with respect to two dominant components: the feature extraction module is initialized with CNNs pre-trained on ImageNet dataset [71], while convolutional layers in relation modules are initialized with a Glorot uniform initializer. Notably, weights in the feature extraction module are trainable and fine-tuned during the training phase.

Regarding the used optimizer, we choose Nestrov Adam [72] and set parameters of the optimizer as recommended: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\mathrm{e}{-08}$. The learning rate is initialized as $2\mathrm{e}{-04}$ and decayed by 0.1 when validation loss is saturated. The loss of our network is simply defined as categorical cross-entropy. We implement the network on TensorFlow and train it on one NVIDIA Tesla P100 16GB GPU for 250k iterations. The size of the training batch is 5, and we stop training when the validation loss fails to decrease.

3) Evaluation Metric: To evaluate the performance of networks, we calculate F_1 score with the following formula:

$$F_1 = (1+\beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}, \quad \beta = 1, \tag{8}$$

for each category. In this equation, *precision* and *recall* are calculated as:

$$precision = \frac{TP}{TP+FP}, \; recall = \frac{TP}{TP+FN}, \tag{9}$$

where TP, FP, and FN represent true positive, false positive, and false negative, respectively. Furthermore, mean F_1 score is computed by averaging all F_1 scores to assess models impartially. Notably, a large F_1 score suggests a better result. Besides, overall accuracy (OA) is also calculated for a comprehensive comparison with different models.

B. An Ablation Study for Relation Modules

In our network, spatial and channel relation modules are employed to explore global relations in both spatial and channel domains. To validate the effectiveness of these modules, we perform ablation experiments (cf. Table I). Particularly, instead of being utilized simultaneously, spatial and channel relation modules are embedded on top of the backbone (i.e., VGG-16), respectively. Besides, we also discuss different integration strategies (i.e., parallel and serial) of relation modules in Table I.

The ablation experiments are conducted on the Vaihingen dataset. As can be seen in Table I, relation modules bring a significant improvement as compared to the baseline FCN (VGG-16), and various integration schemes lead to a slight influence on the performance of our network. In detailed, the use of only the channel relation module yields a result of 87.24% in the mean F_1 score, which brings a 3.50% improvement. Meanwhile, RA-FCN with only the spatial relation module outperforms the baseline by a 4.62% gain in the mean F_1 score.

Moreover, by taking advantage of spatial relation-augmented and channel relation-augmented features simultaneously, the performance of our network is further boosted up. The parallel integration of relation modules brings increments of 1.26% and 0.14% in the mean F_1 score with respect to RA-FCN-crm and RA-FCN-srm. Besides, a serial aggregation strategy is discussed, and results demonstrate that it behaves superiorly as compared to other models. To be more specific, such design achieves the highest mean F_1 score, 88.54%, as well as the highest overall accuracy, 89.23%. To conclude, spatial- and channel-augmented features extracted from relation modules carry out not only high-level semantics but also global relations in spatial and channel dimensionalities, which reinforces the performance of a network for semantic segmentation in aerial scenes.

C. Comparing with Existing Works

For a comprehensive evaluation, we compare our model with six existing methods, including FCN [67], FCN with fully connected CRF (FCN-dCRF) [26], spatial propagation CNN (SCNN) [29], FCN with atrous convolution (Dilated FCN) [26], FCN with feature rearrangement (FCN-FR) [48], and RotEqNet [73].

Numerical results on the Vaihingen dataset are shown in Table II. It is demonstrated that RA-FCN outperforms other methods in terms of both mean F_1 score and overall accuracy. Specifically, comparisons with FCN-dCRF and SCNN, where RA-FCN-srm obtains increments of 4.98% and 3.69% in mean F_1 score, respectively, validate the high performance of the spatial relation module in our network. Besides, compared to FCN-FR, RA-FCN reaches improvements of 1.96% and 1.57% in mean F_1 score and overall accuracy, which indicates the effectiveness of integrating the spatial relation module and channel relation module. In comparison with FCN-FR, although our model achieves lower performance in identifying impervious surfaces and buildings, it reaches improvements of 1.96% and 1.57% in mean F_1 score and overall accuracy, which demonstrates the effectiveness of integrating the spatial and channel relation module in semantic segmentation of aerial images. Besides, compared to dilated FCN and RotEqNet,

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

TABLE II: Experimental Results on the Vaihingen Dataset.

Model Name	Imp. surf.	Build.	Low veg.	Tree	Car	mean F_1	OA
FCN[68]	88.67	92.83	76.32	86.67	74.21	83.74	86.51
FCN-dCRF[27]	88.80	92.99	76.58	86.78	71.75	83.38	86.65
SCNN[29]	88.21	91.80	77.17	87.23	78.60	84.40	86.43
Dilated FCN[26]	90.19	94.49	77.69	87.24	76.77	85.28	87.70
FCN-FR*[48]	**91.69**	**95.24**	79.44	88.12	78.42	86.58	88.92
RotEqNet*[53]	89.50	94.80	77.50	86.50	72.60	84.18	87.50
RA-FCN-srm	91.01	94.86	80.01	88.74	87.16	88.36	89.03
P-RA-FCN	91.46	95.02	80.40	88.56	**87.08**	88.50	89.18
S-RA-FCN	91.47	94.97	**80.63**	**88.57**	87.05	**88.54**	**89.23**

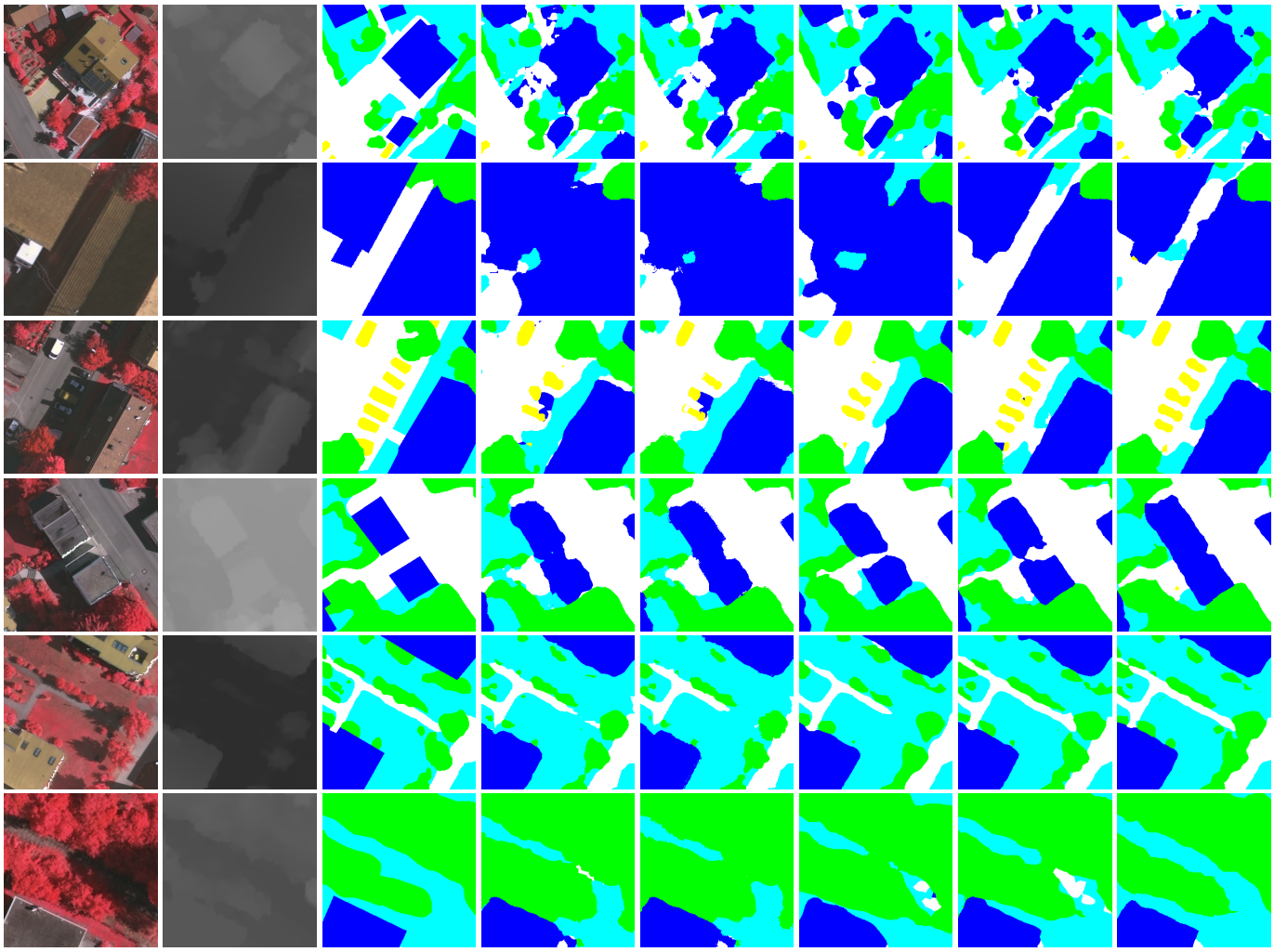


Fig. 6: Examples of segmentation results on the Vaihingen dataset. From left to right: image, nDSM, ground truth, FCN, FCN-dCRF, SCNN, RA-FCN-srm, and S-RA-FCN. Legend—white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars.

RA-FCN obtains increments of 3.26% and 4.36% in mean F_1 score, respectively. Furthermore, per-class F_1 scores are calculated to assess the performance of recognizing different objects. It is noteworthy that our method remarkably surpasses other competitors in identifying scattered cars for its capacity of capturing long-range spatial relation.

D. Qualitative Results

Figure 6 shows a few examples of segmentation results. The second row demonstrates that networks with local receptive fields or relying on fully connected CRFs and spatial propagation modules fail to recognize impervious surfaces between two buildings, whereas our models make relatively

TABLE III: Numerical Results on the Potsdam Dataset.

Model Name	Imp. surf.	Build.	Low veg.	Tree	Car	Clutter	mean F_1	OA
FCN[67]	88.61	93.29	83.29	79.83	93.02	69.77	84.63	85.59
FCN-dCRF[26]	88.62	93.29	83.29	79.83	93.03	69.79	84.64	85.60
SCNN[29]	88.37	92.32	83.68	80.94	91.17	68.86	84.22	85.57
Dilated FCN [26]	86.52	90.78	83.01	78.41	90.42	68.67	82.94	84.14
FCN-FR* [48]	89.31	94.37	84.83	81.10	93.56	76.54	86.62	87.02
RA-FCN-srm	90.48	93.74	85.67	83.10	94.34	74.02	86.89	87.61
P-RA-FCN	90.92	94.20	86.64	83.00	94.44	**77.88**	87.85	88.30
S-RA-FCN	**91.33**	**94.70**	**86.81**	**83.47**	**94.52**	77.27	**88.01**	**88.59**

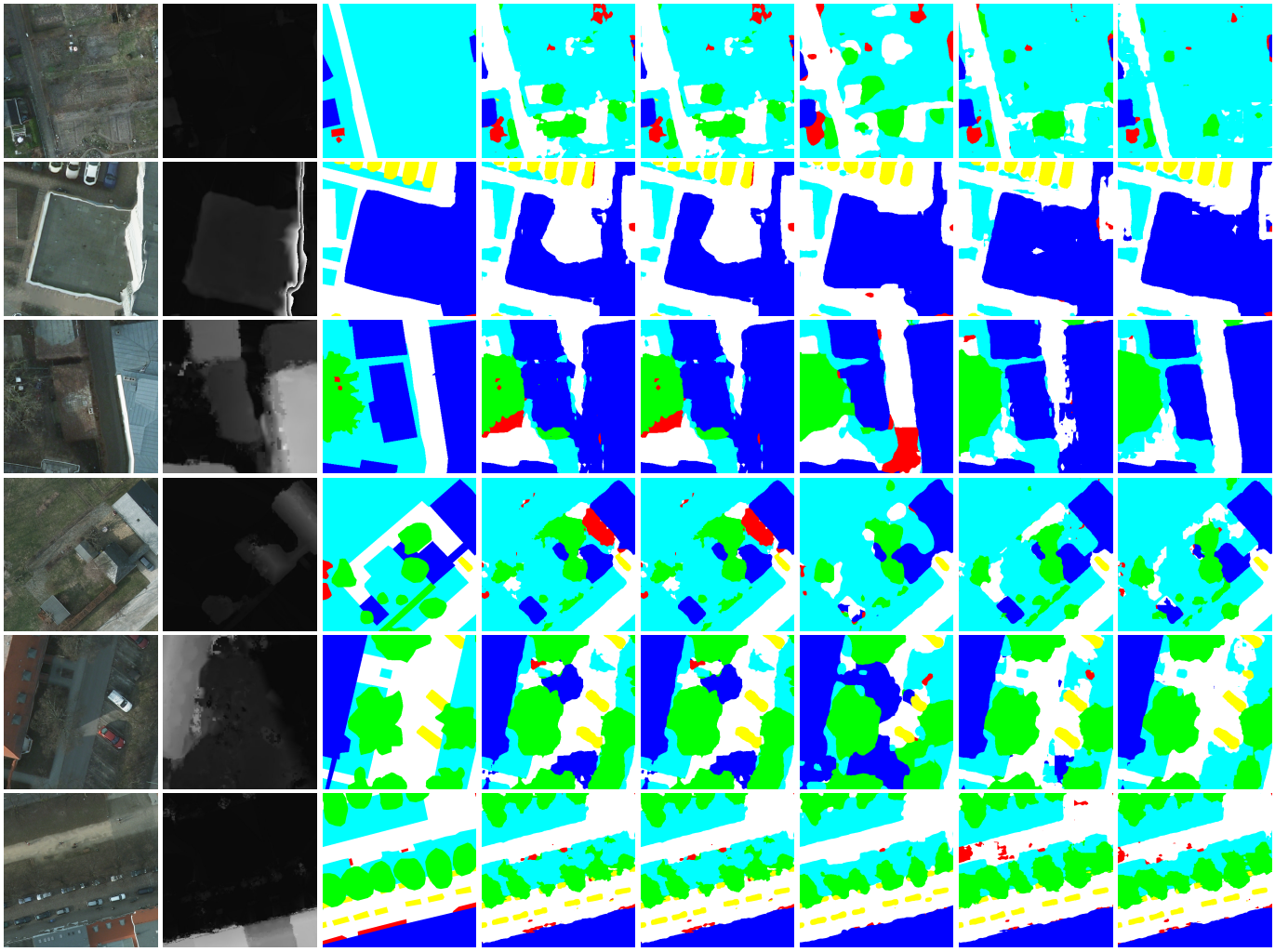


Fig. 7: Examples of segmentation results on the Potsdam dataset. From left to right: image, nDSM, ground truth, FCN, FCN-dCRF, SCNN, RA-FCN-srm, and S-RA-FCN. Legend—white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars, red: clutter/background.

accurate predictions. This is mainly because in this scene, the appearance of impervious surfaces is highly similar to that of the right building, which leads to a misjudgment of rival models. Thanks to the spatial relation module, RA-FCN-srm or RA-FCN is able to effectively capture useful visual cues from more remote regions in the image for an accurate inference. Besides, examples in the third row illustrate that RA-FCN is capable of identifying dispersively distributed objects as expected.

E. Results on the Potsdam Dataset

In order to further validate the effectiveness of our network, we conduct experiments on the Potsdam dataset, and numerical results are shown in Table III. The spatial relation module

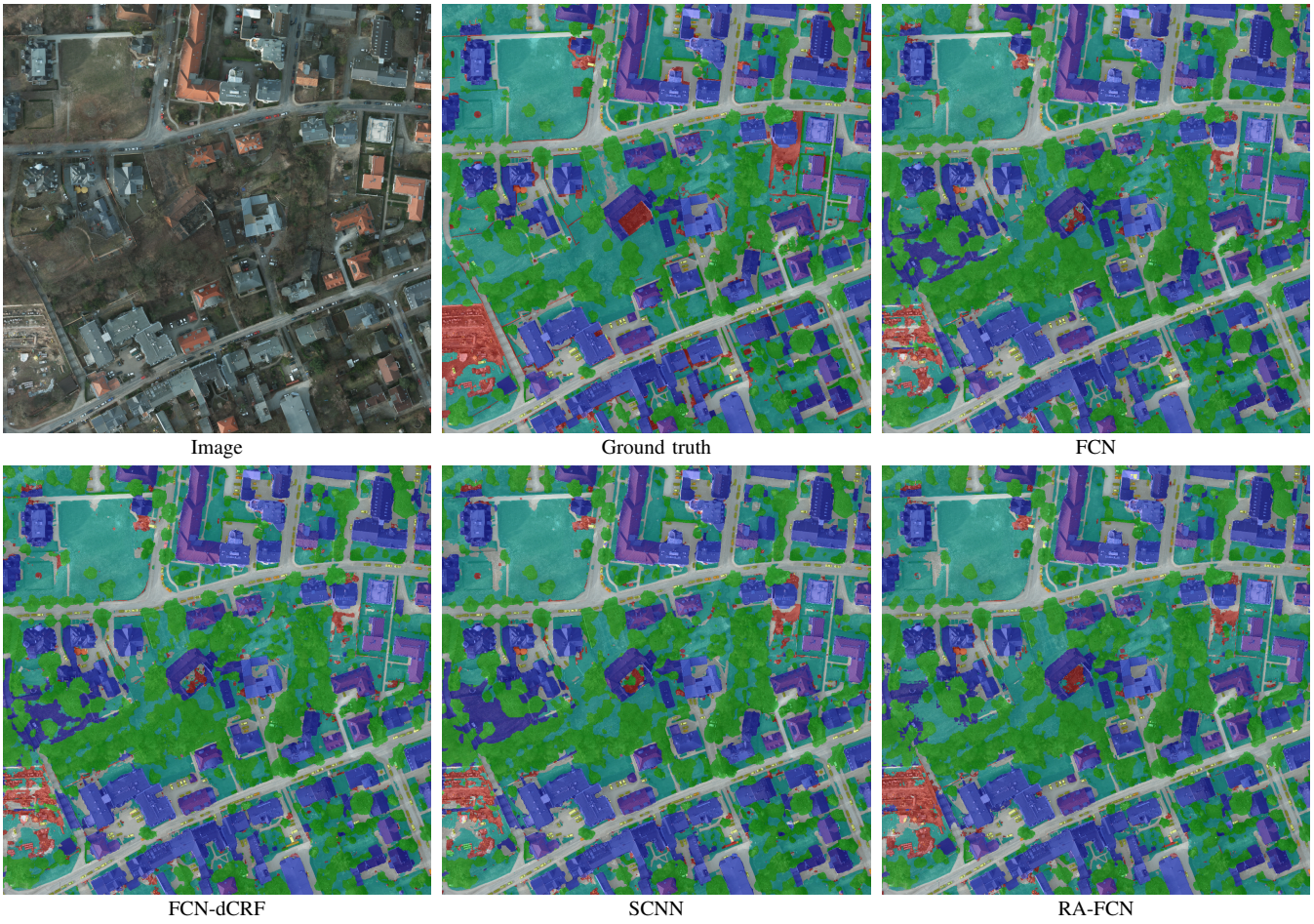


Fig. 8: Full prediction for the tile ID 4_10 of the Potsdam dataset. Legend—white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars, red: clutter/background. Zoom in for details.

contributes to improvements of 2.25% and 2.67% in the mean F_1 score with respect to FCN-dCRF and SCNN, and the serial integration of both relation modules brings increments of 1.39% and 1.54% in the mean F_1 score and overall accuracy, respectively. Compared to dilated FCN, our network increases mean F_1 score and overall accuracy by 5.07% and 4.45%, respectively, which illustrates the importance of capturing long-range relations. Moreover, RA-FCN surpasses FCN-FR in recognizing all six land cover classes in the Potsdam dataset and gains improvements of 1.39% and 1.57% in mean F_1 score and overall accuracy, respectively.

Moreover, qualitative results are presented in Figure 7. As shown in the first row, although low vegetation regions comprise intricate local contextual information and are liable to be misidentified, RA-FCN obtains more accurate results in comparison with other methods due to its remarkable capacity of exploiting global relations to solve visual ambiguities. The fourth row illustrates that outliers, i.e., the misclassified part of the building, can be eliminated by RA-FCN, while it is not easy for other competitors. To provide a thorough view of the performance of our network, we also exhibit a large-scale aerial scene as well as semantic segmentation results in Figure 8. As we can see, RA-FCN shows higher performance in recognizing clutter (see bottom left and central red regions) compared to other competitors. It is noteworthy that identifying clutter is challenging owing to its complicated structures and textures, and thus, leveraging global spatial and channel relations can yield a better classification accuracy. Besides, SCNN tends to confuse low vegetation with buildings (see left blue regions), while for our network, such mistakes are alleviated.

F. Discussion: Pure Spatial Relation Network?

We study a pure spatial relation network, which only uses learned spatial relations and does not exploit convolved feature maps at all to output final segmentation maps. More specifically, we remove the concatenation operation in the spatial relation module of RA-FCN-srm and directly employ $\mathrm{SR}(\boldsymbol{X})$ as the output. We will call the pure spatial relation network FCN-sr hereafter. Experiments are carried out on both the Vaihingen and Potsdam datasets, and quantitative results are reported in Table IV. Before experiments, we expect that without the help of appearance features produced by VGG-16, FCN-sr cannot achieve decent results. However, the performance of FCN-sr is quite better than expected.

TABLE IV: Segmentation Performance of FCNs and Pure Spatial Relation Networks on the Vaihingen (top) and Potsdam (bottom) Datasets.

Model	Imp. surf.	Build.	Low veg.	Tree	Car	Clutter	mean F_1	OA
FCN[67]	88.67	92.83	76.32	86.67	74.21	-	83.74	86.51
FCN-sr	87.94	92.62	76.09	86.56	33.83	-	75.41	86.02
FCN[67]	88.61	93.29	83.29	79.83	93.02	69.77	84.63	85.59
FCN-sr	89.01	93.45	84.20	81.00	91.83	72.29	85.30	86.35

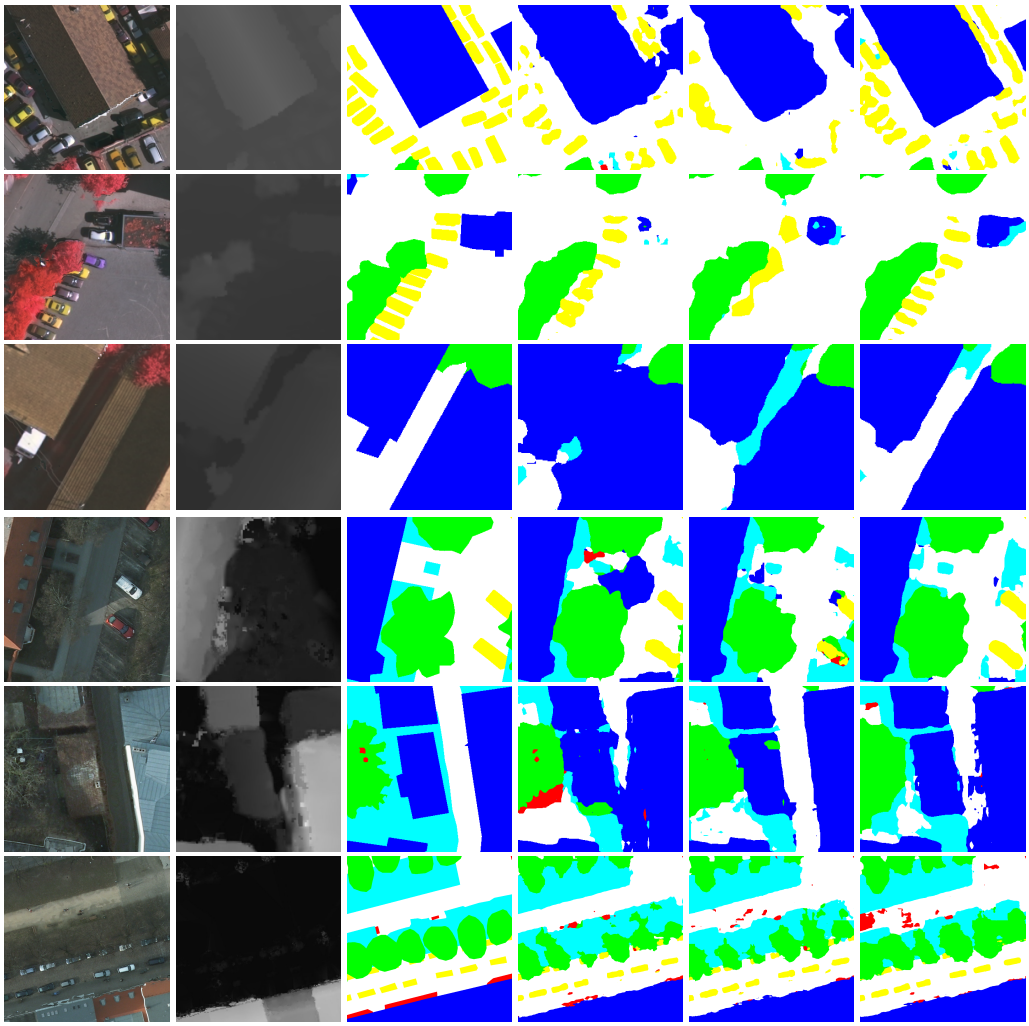


Fig. 9: Comparisons of FCN, pure spatial relation network, and spatial relation-augmented FCN on the Vaihingen dataset (top three rows) and the Potsdam dataset (bottom three rows). From left to right: image, nDSM, ground truth, FCN, FCN-sr, and RA-FCN-srm. Legend—white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars, red: clutter/background.

It is noteworthy that on the Vaihingen dataset, FCN-sr achieves a comparable OA but a rather low mean F_1 score that is 8.33% lower as compared to FCN. This dramatic decrement is mainly because of the misdiagnosis of cars, of which the F_1 is merely 33.83%, less than half of that achieved by FCN. For an intuitive illustration, qualitative results are shown in Figure 9. It is not difficult to find that contours of cars are obscure and conflated with their neighbors. In contrast, the use of appearance information (i.e., FCN) evidently improves the performance of segmenting cars, and combining both spatial relations and appearance information (i.e., RA-FCN-srm) is capable of obtaining a better result. Besides, we find that purely applying learned spatial relations exhibits a superior performance of differentiating building instances as compared to FCN, where only appearance features are used (cf. Figure 9).

On the Vaihingen dataset, FCN-sr achieves both higher mean F_1 score and OA in comparison with FCN on the

Potsdam dataset. We believe this is owing to the fact that the Potsdam dataset has a higher spatial resolution. Qualitative resluts are shown in Figure 9. From this figure, we note that the performance of FCN-sr in terms of inferring roads even surpasses that of RA-FCN-srm, particularly when nDSMs are inaccurate (cf. the 5th column).

V. Conclusion

In this paper, we have introduced two effective network modules, namely the spatial relation module and the channel relation module, to enable relational reasoning in networks for semantic segmentation in aerial scenes. The comprehensive ablation experiments on aerial datasets where long-range spatial relations exist suggest that spatial- and channel-augmented features extracted from relation modules carry out not only high-level semantics but also global relations in spatial and channel dimensionalities, which reinforces the performance of a network for semantic segmentation in aerial scenes. However, our understanding of how these relation modules work for segmentation problems is preliminary and left as future works.

Acknowledgment

The authors would like to thank the ISPRS for making the Vaihingen and Potsdam datasets available.

References

[1] M. Volpi and D. Tuia, "Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 144, no. October, pp. 48–60, 2018.

[2] R. Zhang, G. Li, M. Li, and L. Wang, "Fusion of images and point clouds for the semantic segmentation of large-scale 3d scenes based on deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 143, no. September, pp. 85–96, 2018.

[3] A. L. Fytsilis, A. Prokos, K. D. Koutroumbas, D. Michail, and C. C. Kontoes, "A methodology for near real-time change detection between Unmanned Aerial Vehicle and wide area satellite images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 119, no. September, pp. 165–186, 2016.

[4] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, no. December, pp. 182–196, 2018.

[5] P. Reinartz, M. Lachaise, E. Schmeer, T. Krauss, and H. Runge, "Traffic monitoring with serial images from airborne cameras," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 61, no. December, pp. 149–158, 2006.

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

[7] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X. X. Zhu, "R^3-net: A deep network for multi-oriented vehicle detection in aerial images and videos," *arXiv:1808.05560*, 2018.

[8] L. Matikainen, K. Karila, J. Hyyppä, P. Litkey, E. Puttonen, and E. Ahokas, "Object-based analysis of multispectral airborne laser scanner data for land cover classification and map updating," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 128, no. June, pp. 298–313, 2017.

[9] J. E. Vargas-Munoz, S. Lobry, A. X. Falcao, and D. Tuia, "Correcting rural building annotations in OpenStreetMap using convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 147, no. January, pp. 283–293, 2019.

[10] L. Sahar, S. Muthukumar, and S. P. French, "Using aerial imagery and GIS in automated building footprint extraction and shape recognition for earthquake risk assessment of urban inventories," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 9, pp. 3511–3520, 2010.

[11] A. Vetrivel, M. Gerke, N. Kerle, F. Nex, and G. Vosselman, "Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, no. June, pp. 45–59, 2018.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

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

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Q. Li, C. Qiu, L. Ma, M. Schmitt, Zhu, and X. X., "Mapping the land cover of Africa at 10 m resolution from multi-source remote sensing data with Google earth engine," *Remote Sensing*, vol. 12, no. 4, 2020.

[19] R. Socher, C. C. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *IEEE International Conference on Machine Learning (ICML)*, 2011.

[20] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[21] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, "Augmenting CRFs with Boltzmann machine shape priors for image labeling," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[22] H. Myeong and K. M. Lee, "Tensor-based high-order semantic relation transfer for semantic scene segmentation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[23] J. J. Corso, "Toward parts-based scene understanding with pixel-support parts-sparse pictorial structures," *Pattern Recognition Letters*, vol. 34, no. 7, pp. 762–769, 2013.

[24] Q. Li, Y. Shi, X. Huang, and Z. X. X., "Building footprint generation by integratingconvolution neural network with feature pairwise conditional random field (FPCRF)," 2020.

[25] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," in *IEEE International Conference on Learning Representation (ICLR)*, 2015.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *arXiv:1606.00915*, 2016.

[27] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[28] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[29] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial CNN for traffic scene understanding," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[30] S. Liu, S. D. Mello, J. Gu, G. Zhong, M.-H. Yang, and J. Kautz, "Learning affinity via spatial propagation networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[31] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[33] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] B. Zhou, A. Andonian, and A. Torralba, "Temporal relational reasoning in videos," in *European Conference on Computer Vision (ECCV)*, 2018.

[35] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 280–295, 2015.

[36] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 10, pp. 3747–3762, 2010.

[37] L. Mou and X. X. Zhu, "RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images," *arXiv:1805.02091*, 2018.

[38] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. Le Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia, "Processing of extremely high-resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest–Part A: 2-D contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 12, pp. 5547–5559, 2016.

[39] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] L. Mou, X. X. Zhu, M. Vakalopoulou, K. Karantzalos, N. Paragios, B. L. Saux, G. Moser, and D. Tuia, "Multitemporal very high resolution from space: Outcome of the 2016 IEEE GRSS Data Fusion Contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 8, pp. 3435–3447, 2017.

[41] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6699–6711, 2018.

[42] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[43] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 149, pp. 188–199, 2019.

[44] Y. Hua and L. Mou and X. X. Zhu, "Relation network for multi-label aerial image classification," *IEEE Transactions on Geoscience and Remote Sensing*, no. 99, pp. 1–15, 2020.

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

[46] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNNs," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016.

[47] 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 *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016.

[48] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7092–7103, 2017.

[49] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Asian Conference on Computer Vision (ACCV)*, 2016.

[50] V., A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[51] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, no. June, pp. 20–32, 2018.

[52] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2017.

[53] 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 Journal of Photogrammetry and Remote Sensing*, vol. 135, no. January, pp. 158–172, 2018.

[54] Q. Wang, Y. Yuan, and P. Yan, "Visual saliency by selective contrast," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1150–1155, 2013.

[55] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Transactions on Image Processing*, 2019.

[56] Z. Jiang, Q. Wang, and Y. Yuan, "Modeling with prejudice: Small-sample learning via adversary for semantic segmentation," *IEEE Access*, vol. 6, pp. 77 965–77 974, 2018.

[57] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[58] N. Friedman and D. Koller, "Being Bayesian about network structure. a Bayesian approach to structure discovery in Bayesian networks," *Machine Learning*, vol. 50, no. 1-2, pp. 95–125, 2003.

[59] G. Bertasius, J. Shi, and L. Torresani, "Semantic segmentation with boundary neural fields," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[60] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[61] M. Maire, T. Narihira, and S. X. Yu, "Affinity CNN: Learning pixel-centric pairwise relations for figure/ground embedding," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[62] T.-W. Ke, J.-J. Hwang, Z. Liu, and S. X. Yu, "Adaptive affinity fields for semantic segmentation," in *European Conference on Computer Vision (ECCV)*, 2018.

[63] S. Liu, G. Zhong, S. D. Mello, J. Gu, V. Jampani, M.-H. Yang, and J. Kautz, "Switchable temporal propagation network," in *European Conference on Computer Vision (ECCV)*, 2018.

[64] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *European Conference on Computer Vision (ECCV)*, 2018.

[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *IEEE International Conference on Learning Representation (ICLR)*, 2015.

[66] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[67] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[68] F. Rottensteiner, G. Sohn, M. Gerke, and J. Wegner, "ISPRS test project on urban classification and 3D building reconstruction," http://www2.isprs.org/tl_files/isprs/wg34/docs/ComplexScenes_revision_v4.pdf, 2013.

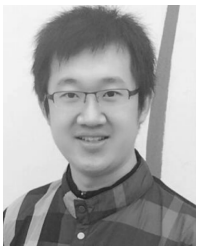
[69] M. Cramer, "The DGPF-Test on digital airborne camera evaluation – overview and test design," https://www.dgpf.de/pfg/2010/pfg2010_2_Cramer.pdf, 2010.

[70] S. Kosov, F. Rottensteiner, C. Heipke, J. Leitloff, and S. Hinz, "3D classification of crossroads from multiple aerial images using Markov Random Fields," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2012.

[71] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[72] T. Dozat, "Incorporating Nesterov momentum into Adam," 2015.

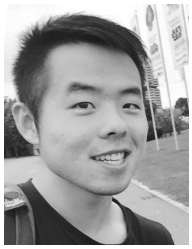
[73] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 96–107, 2018.



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 (UCAS), Beijing, China, in 2015. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Wessling, Germany, and also with the Technical University of Munich (TUM), Munich, Germany. In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany.

In 2019, he was a Visiting Researcher with the University of Cambridge, Cambridge, U.K. His research interests include remote sensing, computer vision, and machine learning, especially deep networks and their applications in remote sensing.

Mr. Mou was a recipient of the first place in the 2016 IEEE GRSS Data Fusion Contest and finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.



Yuansheng Hua (S'18) received the bachelor's degree in remote sensing science and technology from the Wuhan University, Wuhan, China, in 2014, and the master's degree in Earth Oriented Space Science and Technology (ESPACE) from the Technical University of Munich (TUM), Munich, Germany, in 2018. He also received the master's degree in State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) from the Wuhan University, Wuhan, China, in 2019. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Wessling, Germany and the Technical University of Munich (TUM), Munich, Germany.

In 2019, he was a visiting researcher with the Wageningen University & Research, Wageningen, Netherlands. His research interests include remote sensing, computer vision, and deep learning, especially their applications in remote sensing.

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

She is currently the Professor for Signal Processing in Earth Observation (www.sipeo.bgu.tum.de) at Technical University of Munich (TUM) and German Aerospace Center (DLR); the head of the department "EO Data Science" at DLR's Earth Observation Center; and the head of the Helmholtz Young Investigator Group "SiPEO" at DLR and TUM. Since 2019, Zhu is co-coordinating the Munich Data Science Research School (www.mu-ds.de). She is also leading the Helmholtz Artificial Intelligence Cooperation Unit (HAICU) – Research Field "Aeronautics, Space and Transport". Prof. Zhu was a guest scientist or visiting professor at the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan and University of California, Los Angeles, United States in 2009, 2014, 2015 and 2016, respectively. Her main research interests are 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 young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is an associate Editor of IEEE Transactions on Geoscience and Remote Sensing.