A Relation-Augmented Fully Convolutional Network for Semantic Segmentation in Aerial Scenes

Lichao Mou[1,2*], Yuansheng Hua[1,2*], Xiao Xiang Zhu[1,2]
[1] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Germany
[2] Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Germany
{lichao.mou, yuansheng.hua, xiaoxiang.zhu}@dlr.de

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, which fundamentally depend on long-range spatial relational reasoning. The networks achieve very competitive results, bringing signicant improvements over baselines.

1. Introduction

Semantic segmentation of an image involves a problem 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) [19, 33, 12, 16, 1, 40] and massive amounts of labeled data has brought significant progress in this direction. However, although with more complicated and deeper networks and more labeled samples, there is a technical hurdle in the application of CNNs to semantic image segmentation—contextual information.

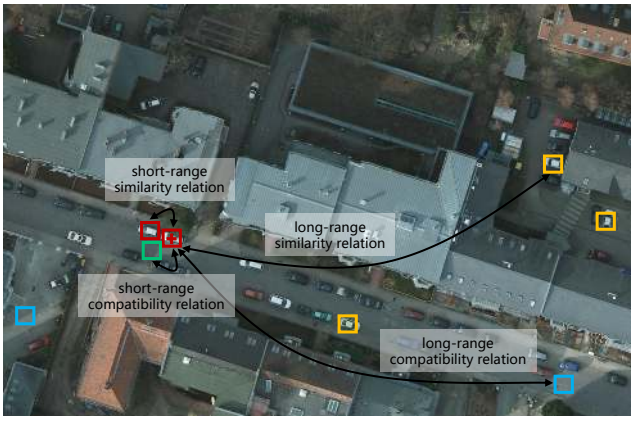

Figure 1: Illustration of long-range spatial relations in an aerial image. Appearance similarity or semantic compatibility between patches within a local region (red–red and red–green) and patches in remote regions (red–yellow and red–blue) underlines our global relation modeling.

It has been well recognized in the computer vision community for years that contextual information, or *relation*, is capable of offering important cues for semantic segmentation tasks [11, 39]. For instance, spatial relations can be considered semantic similarity relationships among regions in an image. In addition, spatial relations also involve compatibility and incompatibility relationships, *i.e.*, 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].

Nevertheless, under some circumstances, spatial rela-

*Equal contribution

[1]Feature maps from deep CNNs like ResNet usually have large receptive fields due to deep architectures, whereas the study of [43] has shown that CNNs are apt to extract information mainly from smaller regions in receptive fields, which are called valid receptive fields.

tions 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 or spatial propagation networks. 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 5 and Figure 6), 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 [31], object detection [13], and activity recognition in videos [42]. 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: describing relationships between observations in convolved images and producing 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.

Contributions. This work's contributions are threefold.

- We propose a simple yet effective and interpretable relation-augmented network that enables spatial and channel relational reasoning in networks for semantic segmentation on aerial imagery.
- 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.

2. Related Work

Semantic segmentation of aerial imagery. Earlier studies [35] have focused on extracting useful low-level, handcrafted visual features and/or modeling mid-level semantic features on local portions of images ([17, 26, 38, 27, 28, 44, 15] employ deep CNNs and have made a great leap towards end-to-end aerial image parsing. In addition, there are numerous contests aiming at semantic segmentation from overhead imagery recently, *e.g.*, Kaggle[2], SpaceNet[3], and DeepGlobal[4].

Graphical models. There are many graphical model-based methods being employed to achieve better semantic segmentation results. For example, the work in [5] makes use of a CRF as post-processing to improve the performance of semantic segmentation. [41] and [22] 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 [3, 4]. These approaches, however, are sensitive to changes in appearance and expensive due to iterative inference processes required.

Spatial propagation networks. Learning spatial propagation with networks for semantic segmentation have attracted high interests in recent years. In [25], the authors try to predict entities of an affinity matrix directly by learning a CNN, which presents a good segmentation performance, while the affinity is followed by a nondifferentiable solver for spectral embedding, which results in the fact that the whole model cannot be trained end-to-end. The authors of [20] train a CNN model to learn a task-dependent affinity matrix by converting the modeling of affinity to learning a local linear spatial propagation. Several recent works [18, 21, 6] focus on the extension of this work. In [2, 29], spatial relations are modeled and reinforced via interlayer propagation. [2] 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.

[2] https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection

[3] https://spacenetchallenge.github.io/

[4] http://deepglobe.org/challenge.html

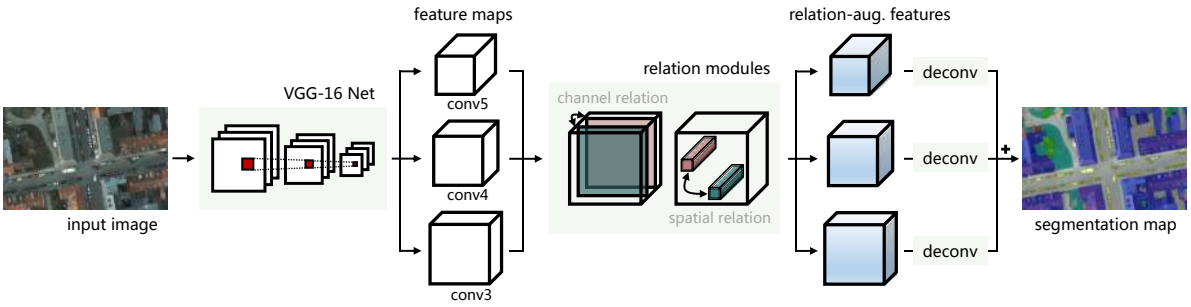

Figure 2: An overview of the relation module-equipped fully convolutional network.

Relation networks. Recently, the authors of [31] propose a relational reasoning network for the problem of visual question answering, and this network achieves a super-human performance. Later, [42] proposes a temporal relation network to enable multi-scale temporal relational reasoning in networks for video classification tasks. In [13], 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 recent success of these works, but we focus on modeling spatial and channel relations in a CNN for semantic segmentation.

Unlike graphical model-based [9, 37] and spatial propagation network-based methods, we explicitly take spatial relations and channel relations into account, so that semantic image segmentation could benet from short- and long-range relational reasoning.

3. Our Approach

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.

3.1. Overview

As illustrated in Fig. 2, the proposed network takes VGG-16 [34] 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 2) for generating relation-augmented features. These features are subsequently fed into respective 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.

3.2. Spatial Relation Module

In order to capture global spatial relations, we employ a spatial relation module, 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 an embedding dot production as g_{θ_s} instead of a multilayer perceptron (MLP), and the latter is commonly used in relational reasoning modules [31, 42]. 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.
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.

3.3. 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 [14] 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 p-th and the q-th channels of $\boldsymbol{X}$. Embedding dot production is employed to be g_{θ_c}, defined as

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

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 inter-dependencies, 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

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

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:

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

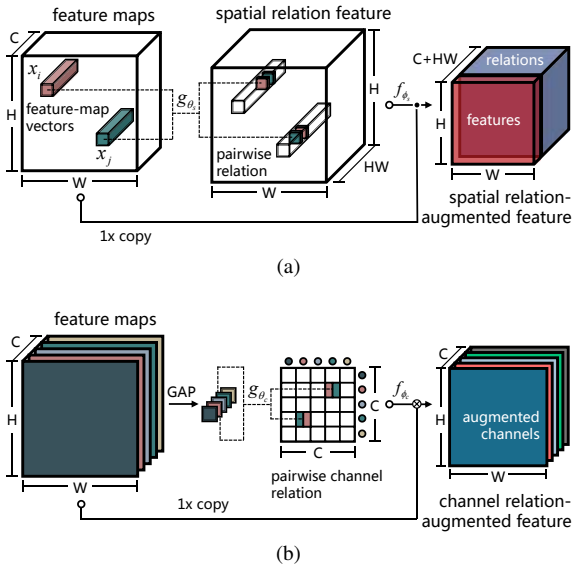

Fig 3: Diagrams of (a) spatial relation module and (b) channel relation module.

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.

3.4. 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 Fig. 4, 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 4.2.

4. 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.

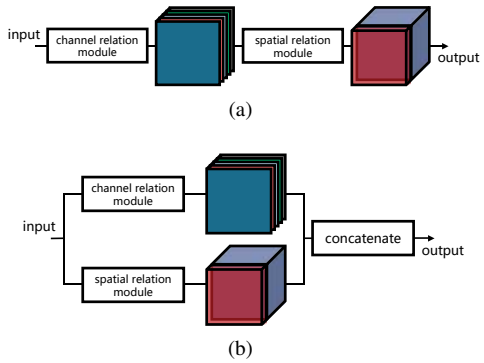

Fig 4: Two integration manners: (a) serial and (b) parallel.

4.1. Experimental Setup

Datasets. The Vaihingen dataset[5] 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 [24, 36, 32, 27], we select 11 images for training, and the remaining five images (image IDs: 11, 15, 28, 30, 34) are used to test our model.

The Potsdam dataset[6] 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. To train and evaluate networks, we utilize 10 images for training and build the test set with the remaining images (image IDs: 02_11, 02_12, 04_10, 05_11, 06_07, 07_08, 07_10), which follows the setup in [24, 32].

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 [7], 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 [8] and set parameters of the optimizer as recommended: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\text{e}{-08}$. The learning rate is initialized as $2\text{e}{-04}$ 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.

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

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

Table 1: Ablation Study on the Vaihingen Dataset.

Model Name	crm	srm	mean F_1	OA
Baseline FCN [23]			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 relation-augmented 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.

Evaluation metric. To evaluate the performance of networks, we calculate F_1 score with the following formula:

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

for each category. 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, mean IoU (mIoU) and overall accuracy (OA) that indicates overall pixel accuracy, are also calculated for a comprehensive comparison with different models.

4.2. 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 1). 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 1.

The ablation experiments are conducted on the Vaihingen dataset. As can be seen in Table 1, relation modules bring a significant improvement as compared to the baseline FCN (VGG-16), and various integration schemes lead

Table 2: Experimental Results on the Vaihingen Dataset

Model Name	Imp. surf.	Build.	Low veg.	Tree	Car	mean F_1	mIoU	OA
SVL-boosting+CRF* [10]	86.10	90.90	77.60	84.90	59.90	79.90	-	84.70
RF+dCRF* [30]	86.90	92.00	78.3	86.90	29.00	74.60	-	85.90
CNN-FPL* [36]	-	-	-	-	-	83.58	-	87.83
FCN [23]	88.67	92.83	76.32	86.67	74.21	83.74	72.69	86.51
FCN-dCRF [5]	88.80	92.99	76.58	86.78	71.75	83.38	72.28	86.65
SCNN [29]	88.21	91.80	77.17	87.23	78.60	84.40	73.73	86.43
Dilated FCN [5]	90.19	94.49	77.69	87.24	76.77	85.28	-	87.70
FCN-FR* [24]	**91.69**	**95.24**	79.44	88.12	78.42	86.58	-	88.92
PSPNet (VGG16) [40]	89.92	94.36	78.19	87.12	72.97	84.51	73.97	87.62
RotEqNet* [27]	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	79.48	89.03
P-RA-FCN	91.46	95.02	80.40	88.56	**87.08**	88.50	79.72	89.18
S-RA-FCN	91.47	94.97	**80.63**	**88.57**	87.05	**88.54**	**79.76**	**89.23**

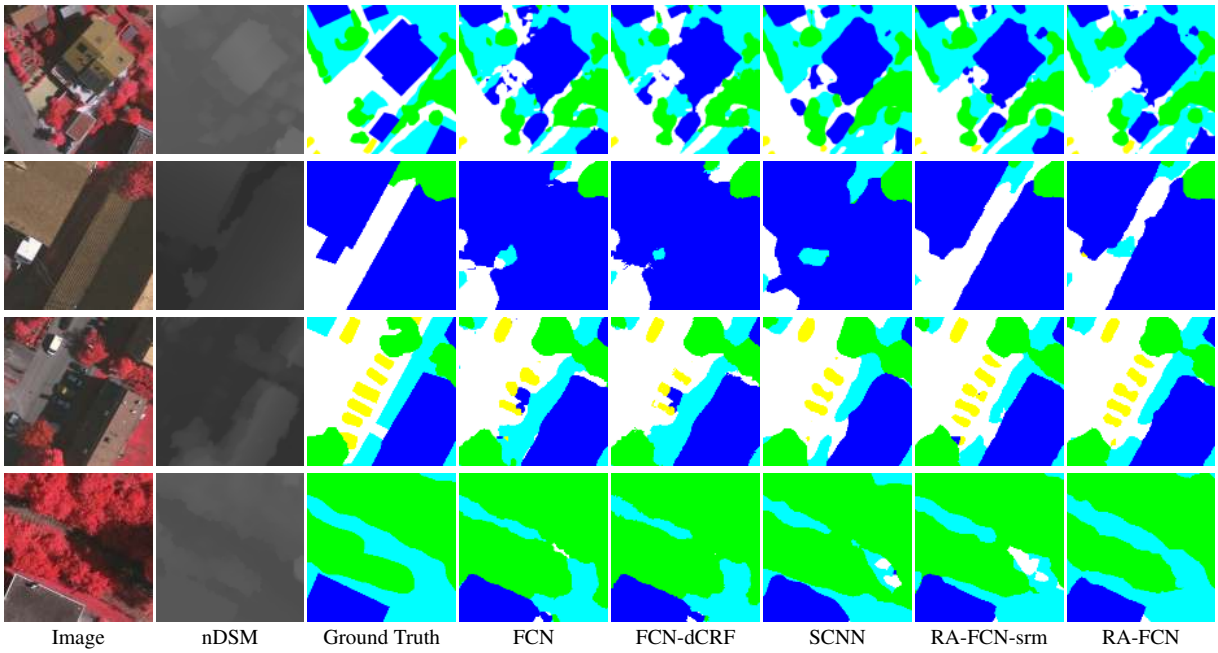

Figure 5: Examples of segmentation results on the Vaihingen dataset. Legend—white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars.

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. In addition, we note that squeeze-and-excitation module [14] can also model dependencies between channels. However, in our experiments, the proposed channel relation module performs better.

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

Table 3: Numerical Results on the Potsdam Dataset

Model Name	Imp. surf.	Build.	Low veg.	Tree	Car	Clutter	mean F_1	mIoU	OA
FCN [23]	88.61	93.29	83.29	79.83	93.02	69.77	84.63	78.34	85.59
FCN-dCRF [5]	88.62	93.29	83.29	79.83	93.03	69.79	84.64	78.35	85.60
SCNN [29]	88.37	92.32	83.68	80.94	91.17	68.86	84.22	77.72	85.57
Dilated FCN* [5]	86.52	90.78	83.01	78.41	90.42	68.67	82.94	-	84.14
FCN-FR* [24]	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	81.23	87.61
P-RA-FCN	90.92	94.20	86.64	83.00	94.44	**77.88**	87.85	81.85	88.30
S-RA-FCN	**91.33**	**94.70**	**86.81**	**83.47**	**94.52**	77.27	**88.01**	**82.38**	**88.59**

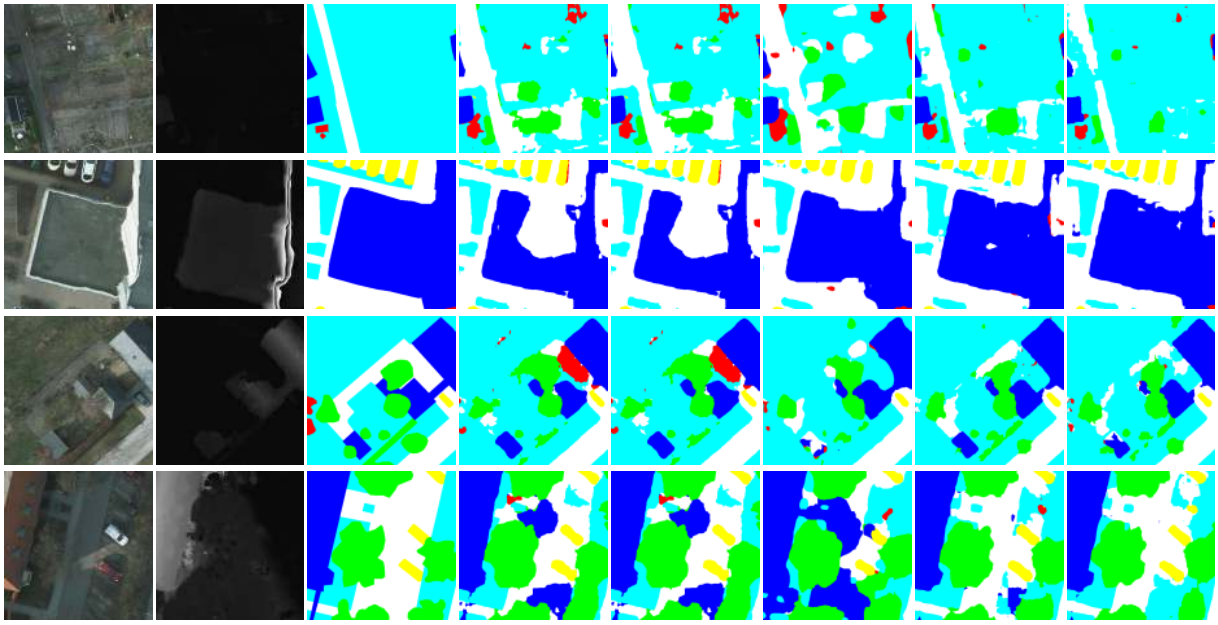

Figure 6: Examples of segmentation results on the Potsdam dataset. Legend—white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars, red: clutter/background.

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.

4.3. Comparing with Existing Works

For a comprehensive evaluation, we compare our model with six existing methods, including FCN [23], FCN with fully connected CRF (FCN-dCRF) [5], spatial propagation CNN (SCNN) [29], FCN with atrous convolution (Dilated FCN) [5], FCN with feature rearrangement (FCN-FR) [24], CNN with full patch labeling by learned upsampling (CNN-FPL) [36], RotEqNet [27], PSPNet with VGG16 as backbone [40], and several traditional methods [10, 30].

Numerical results on the Vaihingen dataset are shown in Table 2. It is demonstrated that RA-FCN outperforms other methods in terms of mean F_1 score, mean IoU, 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. 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.

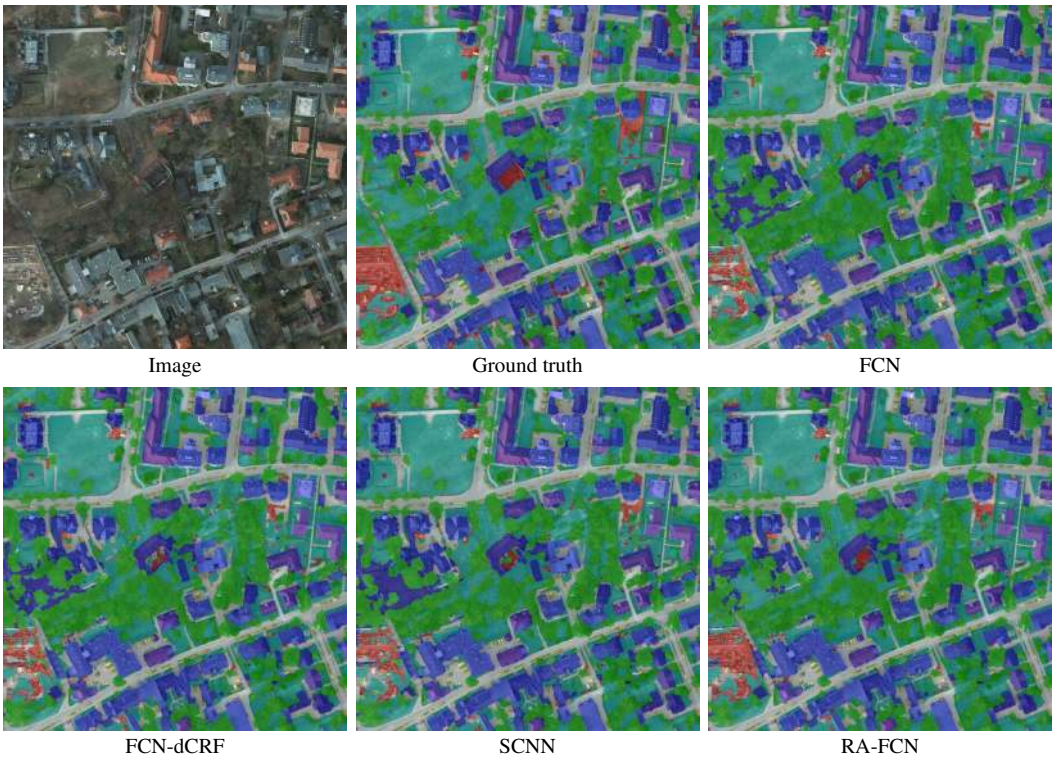

Fig 7: Example segmentation results of an image in the test set on Potsdam dataset ($90,000$ m^2). Legend—white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars, red: clutter/background. Zoom in for details.

4.4. Qualitative Results

Fig. 5 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 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.

4.5. 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 3. The spatial relation module 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, mean IoU, and overall accuracy, respectively.

Moreover, qualitative results are presented in Figure 6. 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 7.

5. 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 both relation modules have learned global relation information between objects and feature maps. However, our understanding of how these relation modules work for segmentation problems is preliminary and left as future works.

References

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

[2] S. Bell, C. Lawrence 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] T. Dozat. Incorporating Nesterov momentum into Adam. 2015.

[9] N. Friedman and D. Koller. Being Bayesian about network structure. a Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50(1-2):95–125, 2003.

[10] M. Gerke. *Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)*. 2015.

[11] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3):300–316, 2008.

[12] 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.

[13] 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.

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

[15] 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*, 149:188–199, 2019.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6054–6068, 2017.

[18] 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.

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

[20] S. Liu, S. De 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.

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

[22] 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.

[23] 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.

[24] 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*, 55(12):7092–7103, 2017.

[25] 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.

[26] D. Marcos, D. Tuia, B. Kellenberger, L. Zhang, M. Bai, R. Liao, and R. Urtasun. Learning deep structured active contours end-to-end. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[27] 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*, 145:96–107, 2018.

[28] 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*, 135:158–172, 2018.

[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] N. Quang, N. Thuy, D. Sang, and H. Binh. An efficient framework for pixel-wise building segmentation from aerial

images. In *International Symposium on Information and Communication Technology, ACM*, 2015.

[31] A. Santoro, D. Raposo, D. G.T. 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.

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

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

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

[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*, 53(1):280–295, 2015.

[36] M. Volpi and D. Tuia. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):881–893, 2017.

[37] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[38] S. Wang, M. Bai, G. Mattyus, H. Chen, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun. TorontoCity: Seeing the world with a million eyes. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[39] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012.

[40] 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.

[41] 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.

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

[43] 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.

[44] X. X. Zhu, D. Tuia, L. Mou, G. 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*, 5(4):8–36, 2017.