GLOBAL MESSAGE PASSING IN NETWORKS VIA TASK-DRIVEN RANDOM WALKS FOR SEMANTIC SEGMENTATION OF REMOTE SENSING IMAGES

Lichao Mou[1], Yuansheng Hua[1, 2], Pu Jin[2], Xiao Xiang Zhu[1, 2]*

[1] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany
- (lichao.mou, yuansheng.hua, xiaoxiang.zhu)@dlr.de
[2] Signal Processing in Earth Observation, Technical University of Munich (TUM), Munich, Germany - pu.jin@tum.de

Commission II

KEY WORDS: Deep learning, global message passing, random walking, semantic segmentation, remote sensing

ABSTRACT:

The capability of globally modeling and reasoning about relations between image regions is crucial for complex scene understanding tasks such as semantic segmentation. Most current semantic segmentation methods fall back on deep convolutional neural networks (CNNs), while their use of convolutions with local receptive fields is typically inefficient at capturing long-range dependencies. Recent works on self-attention mechanisms and relational reasoning networks seek to address this issue by learning pairwise relations between each two entities and have showcased promising results. But such approaches have heavy computational and memory overheads, which is computationally infeasible for dense prediction tasks, particularly on large size images, i.e., aerial imagery. In this work, we propose an efficient method for global context modeling in which at each position, a sparse set of features, instead of all features, over the spatial domain are adaptively sampled and aggregated. We further devise a highly efficient instantiation of the proposed method, namely learning RANdom walK samplIng aNd feature aGgregation (RANKING). The proposed module is lightweight and general, which can be used in a plug-and-play fashion with the existing fully convolutional neural network (FCN) framework. To evaluate RANKING-equipped networks, we conduct experiments on two aerial scene parsing datasets, and the networks can achieve competitive results at significant low costs in terms of the computational and memory.

1. INTRODUCTION

Capturing and modeling both short- and long-range relations is of paramount importance for many vision tasks, to name a few, semantic segmentation (Fu et al., 2019, Liu et al., 2017, Bertasius et al., 2017), object detection (Shvets et al., 2019, Hu et al., 2018), action recognition (Wang et al., 2018), and visual question answering (VQA) (Santoro et al., 2017, Lobry et al., 2019). Being able to reason about such relations among different regions in an image/video is inherent to humans, but is not easy for convolutional neural networks (CNNs). Because an individual convolution layer can only learn features locally, and deep CNNs with large receptive fields have proven to be not efficient at modeling long-range dependencies (Luo et al., 2016, Zhou et al., 2015).

To address this issue, many efforts have been made to enhance the capacity of CNNs to capture long-term relations, such as dilated convolutions (Chen et al., 2015, Chen et al., 2018a, Chen et al., 2018b), introducing graphical models into networks (Chen et al., 2018a, Liu et al., 2015, Zheng et al., 2015), and constructing spatial propagation network modules (Bell et al., 2016, Liu et al., 2017). These approaches make an attempt at capturing global relations by means of a chain propagation way, which is implicitly global and whose effectiveness depends heavily on the learning effect of long-term memorization.

Recent advances in self-attention mechanisms (Vaswani et al., 2017, Wang et al., 2018, Hu et al., 2018) and relational reasoning networks (Santoro et al., 2017) have shown promising results in explicitly modeling global context. In essence, these methods somehow learn pairwise relations between each two entities (i.e., feature-map vectors and pixels) and then make use of them for feature aggregation or augmentation. By doing so, a fully connected relationship graph is learned to explicitly represent global context, which, however, leads to a quadratic inference complexity with respect to the number of entities and a high GPU memory overhead. This is computationally infeasible for dense prediction tasks, particularly on large size images.

The aforementioned methods imply that even for entity pairs whose relations actually do not matter, these models have to learn to infer their relationships, which is usually unnecessary. Hence, taking advantage of only relations that should be considered for global reasoning is conceptually interesting and helps in reducing significant computational and memory costs, but still remains under explored.

In this work, our goal is to explicitly model global context *with low computational and memory overheads* in a fully convolutional network (FCN) for aerial scene parsing by considering a sparse set of important short- and long-range relations instead of all. More specifically, a plug-and-play network module, RANKING (learning RANdom walK samplIng aNd feature aGgregation) is devised and appended on top of an FCN to adaptively sample informative feature-map vectors and then aggregate them in order to produce better segmentation results. This work is inspired by GraphSAGE (Hamilton et al., 2017), an inductive representation learning framework in natural language processing (NLP). But unlike the latter that assumes a static graph where neighbors for each node are fixed, our module is capable of dynamically learning a global sampling.

Contributions. This work's contributions are threefold.

* Corresponding author

https://doi.org/10.5194/isprs-annals-V-2-2020-533-2020

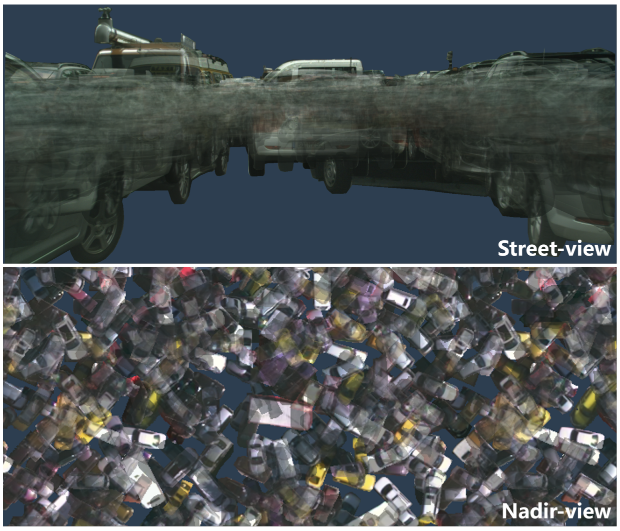

Figure 1. Illustration of different distributions of the same object (taking vehicle as an example) in street-view (top) and nadir-view (bottom) images. It can be seen that in overhead images there are more long-range relations (see appearance similarities among white cars). [Statistics based on Cityscapes and ISPRS Vaihingen datasets.]

- We propose a simple yet efficient approach for global context modeling in networks with low computational and memory costs by adaptively sampling feature-map vectors and then aggregating them at each position.
- We devise RANKING, a highly efficient instantiation of the proposed method, that implements the sampling by learnable random walks and the feature aggregation via an averaging operation at zero parameters.
- We validate the effectiveness of our network module through extensive ablation studies.

2. RELATED WORK

Self-attention mechanism. Self-attention mechanism has by now been successfully applied in a wide range of NLP tasks, e.g., machine translation (Vaswani et al., 2017), due to its superior ability in modeling long-range dependencies. A recent trend in NLP is replacing recurrent neural networks (RNNs) by self-attention models, thereby allowing more efficient learning and parallelized implementations. From NLP to computer vision, (Wang et al., 2018) extends self-attention for NLP to a more general form of non-local operations, which computes the response at a position by attending to all positions and taking their weighted sum in an embedding space based on a learned affinity matrix. In (Hu et al., 2018), the authors exploit the self-attention mechanism to model relations among sets of objects in object detection. There are also works that make use of the self-attention mechanism for semantic segmentation tasks (Fu et al., 2019, Huang et al., 2019, Yuan, Wang, arXiv:1809.00916). In addition, several variants of the original non-local module (Wang et al., 2018) can be found in (Yue et al., 2018, Chen et al., 2019, Zhang et al., arXiv:1908.06955).

Relational reasoning networks. Recently, (Santoro et al., 2017) proposes a relation network to solve problems that involve spatial relational reasoning by learning the potential relations between all feature-map vector pairs, and this network achieves a super-human performance in VQA tasks. Later, in (Zhou et al., 2018), the authors introduce a temporal relation network module that explicitly learns multi-scale temporal dependencies among video frames for video classification problems. Besides spatial and temporal relations in images and videos, the authors of (Duan et al., 2019) present a structural relation network to reason about structural dependencies of local regions in 3D point clouds. In (Cadène et al., 2019), a multimodal relational network is proposed to represent interactions between a question and image regions and model region relations with pairwise combinations for VQA.

Aerial scene parsing. There is a long tradition of leveraging computer vision techniques for aerial scene parsing. Earlier works (Liu, Liu, 2014, Blaschke et al., 2004, Predoehl et al., 2013) mainly lie on exploring effective visual features and semantic modeling approaches. Recently, deep CNNs have been widely explored in this field and taken a giant leap (Marcos et al., 2018a, Li et al., 2019, Sun et al., 2019, Azimi et al., 2019, Wang et al., 2017, Kellenberger et al., 2019, Cheng et al., 2019, Marcos et al., 2018a, Mou et al., 2019). Furthermore, there are numerous challenges being aimed at semantic segmentation of overhead images, e.g., Deep Global[1], and SpaceNet[2].

3. OUR APPROACH

3.1 Problem Formulation

Let $\mathbf{F} = \{\mathbf{f}(\mathbf{p})\}$ with $\mathbf{f}(\mathbf{p}) \in \mathbb{R}^{C\times1\times1}$ interpret C feature maps, where each vector is identified by a spatial position index $\mathbf{p} = (x, y)$. Our goal is to learn a set of refined feature-map vectors $\mathbf{Z} = \{\mathbf{z}(\mathbf{p})\}$ by globally adaptively aggregating a sparse set of vectors at different locations. To this end, in this work, we propose a network module, RANKING, to learn random walk sampling and aggregate sampled features (cf. Figure 2).

3.2 Learning Random Walk Sampling

We consider learning a random walk with t steps operating across grids. The position $\mathbf{p}_\tau = (x_\tau, y_\tau)$ at step τ $(0 \le \tau < t)$ can be traced to position $\mathbf{p}_{\tau+1} = (x_{\tau+1}, y_{\tau+1})$ at the next step $(\tau + 1)$ with a motion vector $\overrightarrow{\omega} = (u_\tau, v_\tau)$ using the following equation:

$$\mathbf{p}_{\tau+1} = \mathbf{p}_\tau + \overrightarrow{\omega}(\mathbf{p}_\tau) = (x_\tau, y_\tau) + \overrightarrow{\omega}|_{(x_\tau, y_\tau)}\,. \tag{1}$$

The final sampling position $\mathbf{p}_t$ can be calculated by iteratively applying Eq. (1). In a traditional random walk, the motion vector $\overrightarrow{\omega}$ can be arbitrarily long and in any direction. Here we would like to learn a data- and task-driven random walk sampling and define a learnable $\overrightarrow{\omega}$ as follows:

$$\overrightarrow{\omega} = (u_\tau(\mathbf{f}(\mathbf{p}_\tau)), v_\tau(\mathbf{f}(\mathbf{p}_\tau)))\,. \tag{2}$$

With $u_\tau(\cdot)$ and $v_\tau(\cdot)$, the next position can be predicted, conditioned on the feature of the current position. For simplicity and more efficient computation, we consider them in the form of a linear embedding, i.e.,

$$u_\tau(\mathbf{f}(\mathbf{p}_\tau)) = \mathbf{w}_{u_\tau}^T \mathbf{f}(\mathbf{p}_\tau)\,, \tag{3}$$

[1] https://www.deepglobal.org/challenge.html
[2] https://spacenetchallenge.github.io/

https://doi.org/10.5194/isprs-annals-V-2-2020-533-2020 | © Authors 2020. CC BY 4.0 License.

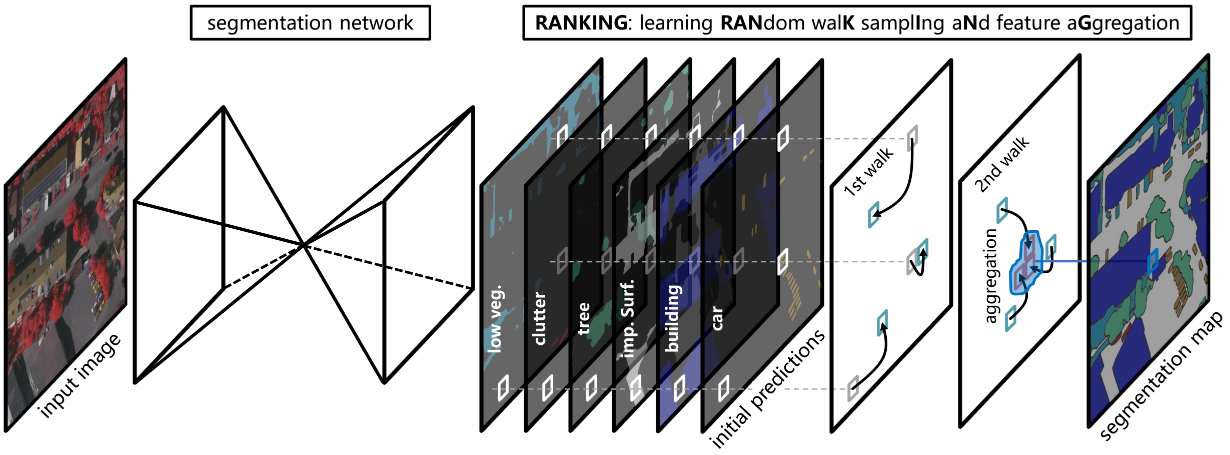

Figure 2. An overview of a RANKING-equipped fully convolutional network for aerial scene parsing tasks. The number of steps in the sampling procedure is 2 in this case.

$$v_\tau(\mathbf{f}(\mathbf{p}_\tau)) = \mathbf{w}_{v_\tau}^T \mathbf{f}(\mathbf{p}_\tau)\,, \qquad (4)$$

where $\mathbf{w}_{u_\tau}$ and $\mathbf{w}_{v_\tau}$ are learnable weight vectors and can be implemented as 1×1 convolutions.

Since outputs of $u_\tau(\cdot)$ and $v_\tau(\cdot)$ are typically real values, the sampling position $\mathbf{p}_\tau$ becomes fractional. It can be seen from Eq. (2) that the estimation of $\vec{\omega}$ is associated with the feature-map vector at the position $\mathbf{p}_\tau$. Hence we make use of an interpolation algorithm with a sampling kernel $\mathcal{K}$ to generate $\mathbf{f}(\mathbf{p}_\tau)$ as follows:

$$\mathbf{f}(\mathbf{p}_\tau) = \sum_{\mathbf{q} \in \mathcal{N}(\mathbf{p}_\tau)} \mathcal{K}(\mathbf{q}, \mathbf{p}_\tau)\mathbf{f}(\mathbf{q})\,, \qquad (5)$$

where $\mathcal{N}(\mathbf{p}_\tau)$ indicates four nearest neighbors of the position $\mathbf{p}_\tau$ on the grid. We do not dive deeper into various choices of $\mathcal{K}$ and utilize bilinear interpolation as default.

3.3 Feature Aggregation

The goal of this stage is to aggregate sampled features and generate new feature representations that can facilitate the subsequent classification/segmentation tasks.

We first revisit the feature aggregation in self-attention models, which is considered the following equation:

$$\mathbf{z}_\mathbf{p} = \sum_{\forall \mathbf{q}} \frac{1}{\mathcal{C}} w_{\mathbf{pq}} \mathbf{f}(\mathbf{q})\,. \qquad (6)$$

Here $\mathbf{p}$ is a query position whose response $\mathbf{z}_\mathbf{p}$ is to be calculated and $\mathbf{q}$ indicates all possible positions. $w_{\mathbf{pq}}$ represents the relationship between $\mathbf{p}$ and $\mathbf{q}$. Moreover, $\mathcal{C}$ is a normalization constant. Eq. (6) is a weighted sum of all feature-map vectors, but learning pairwise relations $\mathbf{W} = \{w_{\mathbf{pq}}\}$ is computationally expensive.

In order to reduce the computational overhead, in this work, we perform the feature aggregation at zero parameters as follows:

$$\mathbf{z}_\mathbf{p} = \sum_{\mathbf{q} \in \mathcal{V}(\mathbf{p})} \frac{1}{|\mathcal{V}(\mathbf{p})|} \mathbf{f}(\mathbf{q})\,, \qquad (7)$$

where $\mathcal{V}(\mathbf{p})$ is a set of sampled positions, conditioned on the position $\mathbf{p}$. As compared to Eq. (6), Eq. (7) has two changes: 1) $\forall \mathbf{q} \rightarrow \mathbf{q} \in \mathcal{V}(\mathbf{p})$; 2) $\frac{1}{\mathcal{C}} w_{\mathbf{pq}} \rightarrow \frac{1}{|\mathcal{V}(\mathbf{p})|}$. By doing so, we achieve the feature aggregation in an efficient way.

This stage is actually flexible, and we believe that alternative versions, e.g., LSTM (Hochreiter, Schmidhuber, 1997), are possible and may improve results.

3.4 Implementation

The proposed network module can be easily incorporated into a large variety of existing backbone CNN architectures in a plug-and-play fashion. To make the proposed method fully comparable with others, we choose VGG-16 as the backbone for aerial scene parsing tasks. Outputs of *conv3*, *conv4*, and *conv5* are fed into respective 1×1 convolutional layers to squash the number of channels to the number of categories, and then the convolved feature maps are upsampled to a desired full resolution and element-wise added to generate initial segmentation maps. These seed predictions are subsequently refined by the proposed RANKING module.

4. EXPERIMENTS

To demonstrate the effectiveness of our RANKING module, we conduct experiments on two aerial scene parsing datasets, i.e., ISPRS Vaihingen and Potsdam datasets. In our experiments, we perform ablation studies on the Vaihingen dataset and compare our network with existing methods on both datasets. Moreover, we visualize trajectories of the adaptive random walk sampling to provide an insight view into our RANKING module. Notably, image samples in the two datasets are collected from nadir view, and thus the spatial distribution of objects in these images is diverse and complex (see Figure 1).

4.1 Experimental Setup

Datasets. The Vaihingen dataset[3] is an aerial image semantic segmentation dataset, which consists of 33 aerial images covering a 1.38 km^2 area of the city of Vaihingen. The spatial resolution of each image is 9 cm, and their average size is 2494×2064 pixels. Three bands, including near infrared (NIR), red (R), and green (G) wavelengths, and digital surface models (DSMs) are available for each aerial image. Besides, pixel-wise annotations

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

of only 16 images are provided, and most existing works (Maggiori et al., 2017, Volpi, Tuia, 2017, Sherrah, 2016, Marcos et al., 2018b) select 11 images to train their models. The remaining five images (image IDs: 11,15, 28, 30, 34) are used to test their models. In this work, we follow this train-test split in our experiments.

The Potsdam dataset[4] is more challenging owing to its increasing number of samples, enlarged image size, and finer spatial resolution. Specifically, 38 images with a size of 6000 × 6000 pixels are gathered, and the spatial resolution of them is 5 cm. In addition, each aerial image covers an area of of 3.42 km^2, and four bands (NIR, R, G, and blue (B)) are collected for these images. DSMs with the same spatial resolution is provided as well. In our experiments, we follow the setup in (Maggiori et al., 2017) and train our network with 17 images. The remaining samples (image IDs: 02_11, 02_12, 04_10, 05_11, 06_07, 07_08, 07_10) are used to test our model.

Initialization and training strategies. We adopt different initialization strategies with respect to each component of our network: the backbone is initialized with corresponding pre-trained CNNs, and convolutional filters in the RANKING module are initialized using a normal distribution with zero mean and a modest standard deviation 0.01. This is based on an assumption that regarding each pixel, its original neighbours should be the most relevant and could provide sufficient semantics for predicting it.

We implement our network on TensorFlow and select Nestrov Adam (Dozat, 2015) as the optimizer. Parameters of the optimizer is set as recommended: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\mathrm{e}{-08}$. The initial learning rate is $2\mathrm{e}{-04}$ and decayed by 0.1 once the validation loss is saturated. The batch size is 5, and we define the loss function of the network as categorical cross-entropy. During the training phase, all weights are learnable, and each model is trained on one NVIDIA TeslaP100 16GB GPU.

Evaluation metrics. To measure the performance of networks for aerial scene parsing comprehensively, we first calculate per-class F_1 scores and then average them to obtain mean F_1 score. Here, a large F_1 score indicates a better result. Besides, mean IoU (mIoU) and overall accuracy (OA) are calculated as well.

4.2 Ablation Studies

Effectiveness of RANKING module. In the ablation study, we first evaluate our RANKING module by comparing FCN+RANKING with a vanilla FCN. As can be seen in Table 1, by using the proposed RANKING module, our network can achieve improvements of at least 4.37% and 2.23% (see FCN+RANKING-3-1) in the mean F_1 score and OA, respectively, as compared to the baseline FCN. Furthermore, the maximum increment for the mean F_1 score can reach 4.76% (cf. FCN+RANKING-3-2). In general, introducing RANKING module into the baseline model brings a significant improvement.

Table 1. Ablation Study on the Vaihingen Dataset.

Model	p	# walks	m. F_1	OA
Baseline FCN	-	-	83.74	86.51
FCN+RANKING-3-1	3 × 3	1	88.24	88.80
FCN+RANKING-5-1	5 × 5	1	88.11	88.74
FCN+RANKING-7-1	7 × 7	1	88.38	**89.06**
FCN+RANKING-3-2	3 × 3	2	**88.50**	88.97
FCN+RANKING-5-2	5 × 5	2	88.49	88.96
FCN+RANKING-7-2	7 × 7	2	88.31	88.84
FCN+RANKING-3-3	3 × 3	3	88.20	88.86
FCN+RANKING-5-3	5 × 5	3	88.30	88.84
FCN+RANKING-7-3	7 × 7	3	88.39	88.97
FCN+RANKING-3-4	3 × 3	4	88.46	88.85
FCN+RANKING-5-4	5 × 5	4	88.12	88.83
FCN+RANKING-7-4	7 × 7	4	88.24	88.85

[1] FCN+RANKING-X-Y indicates a RANKING-equipped FCN with the following configuration: the aggregation operator size is X and the number of steps is Y.

[2] m. F_1 indicates the mean F_1 score.

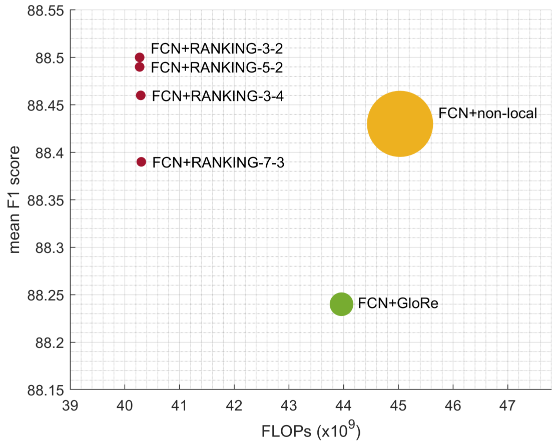

Figure 3. Ablation study on the Vaihingen dataset. Red circles indicate our proposed FCN+RANKING network, and different versions are considered here. Green and yellow circles correspond to FCN+GloRe and FCN+non-local, respectively. The size of circles refers to the memory consumption of each model. Notably, models located at the upper left corner are high-performance and computationally efficient.

Effect of the aggregation operator size. To explore the effect of the aggregation operator size, denoted as p, we evaluate our FCN+RANKING with various p and report results in Table 1. As shown here, when the number of steps is 1 and 3, FCN+RANKING with a large p, i.e., 7 × 7, achieves best F_1 scores. Besides, for the number of steps 2 and 4, FCN+RANKING with a 3 × 3 aggregation operator performs best. To conclude, varying the aggregation operator size brings modest effect to our module.

Effect of the number of steps. The number of steps is an essential property of our module as it determines how far the sampler goes. Hence, it is useful to analyze how different numbers of walks influence performance. In Table 1, we can observe that by randomly walking twice, FCN+RANKING can achieve the highest mean F_1 score with a small p. One possible explanation could be that the correlation between the center pixel and sampled pixels might be weak once the number of steps exceeds 2.

Computational and memory overheads. To measure the computational complexity, we report FLOPs (floating-point multiply-adds $\times 10^9$) and the number of parameters in Table 3. As shown in this table, our model requires comparable FLOPs but can achieve an increment of 4.76% in the mean F_1 score compared

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

Table 2. Experimental Results on the Vaihingen Dataset

Model	Imp. surf.	Build.	Low veg.	Tree	Car	mean F_1	mIoU	OA
FCN (Long et al., 2015)	88.67	92.83	76.32	86.67	74.21	83.74	72.69	86.51
CNN-FPL* (Volpi, Tuia, 2017)	-	-	-	-	-	83.58	-	87.83
RF+dCRF* (Quang et al., 2015)	86.90	92.00	78.3	86.90	29.00	74.60	-	85.90
SVL-boosting+CRF* (Gerke, 2015)	86.10	90.90	77.60	84.90	59.90	79.90	-	84.70
Dilated FCN (Chen et al., 2018a)	90.19	94.49	77.69	87.24	76.77	85.28	-	87.70
FCN-FR* (Maggiori et al., 2017)	**91.69**	**95.24**	79.44	88.12	78.42	86.58	-	88.92
PSPNet (VGG16) (Zhao et al., 2017)	89.92	94.36	78.19	87.12	72.97	84.51	73.97	87.62
RotEqNet* (Marcos et al., 2018b)	89.50	94.80	77.50	86.50	72.60	84.18	-	87.50
FCN-dCRF (Chen et al., 2018a)	88.80	92.99	76.58	86.78	71.75	83.38	72.28	86.65
SCNN (Pan et al., 2018)	88.21	91.80	77.17	87.23	78.60	84.40	73.73	86.43
FCN+non-local (Yue et al., 2018)	89.54	93.42	79.46	88.17	71.57	88.43	73.84	87.86
FCN+GloRe (Chen et al., 2019)	91.25	94.98	**80.15**	88.38	86.42	88.24	79.30	**89.02**
FCN+RANKING-3-2	91.16	94.97	79.80	**88.42**	**88.17**	**88.50**	**79.73**	88.97
FCN+RANKING-5-3	91.10	94.77	79.86	88.42	87.15	88.26	79.33	88.91

to baseline FCN. In comparison with FCN+GloRe (Chen et al., 2019), our model achieves an improvement of 0.26% in the mean F_1 score with fewer FLOPs. Besides, although our model surpasses FCN+non-local by a marginal improvement, the computational complexity of our model is quite low.

To calculate the memory consumption, we take only the forward pass of one patch into consideration. As shown in Table 3, our network requires only 14% and 2% of memory consumption required by FCN+GloRe and FCN+non-local, respectively. Besides, comparisons between FCN+RANKING-3-2 and baseline FCN also demonstrate that our module is very lightweight and memory efficient. To conclude, the integration of the RANKING module can reinforce the performance of a network for aerial scene parsing at a very low computational cost.

Table 3. Comparison of computational costs.

Model	m.F_1	FLOPs	Mem.
Baseline FCN (Long et al., 2015)	83.74	40.23 G	0.23 GB
FCN+non-local (Yue et al., 2018)	88.43	45.03 G	13.03 GB
FCN+GloRe (Chen et al., 2019)	88.24	43.96 G	1.59 GB
FCN+RANKING-3-2	88.50	40.27 G	0.23 GB

[1] To calculate memory consumption, all tensors are considered and their type is defined as float32.

4.3 Comparison with Existing Works

In order to further evaluate our model, we compare FCN+RANKING with twelve existing models, including FCN (Long et al., 2015), RotEqNet (Marcos et al., 2018b), FCN with atrous convolution (DilatedFCN) (Chen et al., 2018a), spatial propagation CNN (SCNN) (Pan et al., 2018), FCN with fully connected CRF (FCN-dCRF) (Chen et al., 2018a), CNN with full patch labeling by learned upsampling (CNN-FPL) (Volpi, Tuia, 2017), PSPNet with VGG16 as back-bone (Zhao et al., 2017), FCN with feature rearrangement (FCN-FR) (Maggiori et al., 2017), FCN with a non-local block (embedded Gaussian version) (Yue et al., 2018), FCN with a GloRe unit (Chen et al., 2019), and two traditional machine learning methods (Gerke, 2015, Quang et al., 2015).

Quantitative results of all models on the Vaihingen dataset are exhibited in Table 2. It can be seen that our FCN+RANKING-3-2 achieves the highest mean F_1 score and mean IoU compared to other competitors. To be more specific, our model surpasses FCN-dCRF and SCNN by 4.98% and 3.69% in the mean F_1 score, respectively. By comparing FCN+RANKING with FCN+non-local and FCN+GloRe, we observe that although our method requires relatively few computational resources, it can still achieve marginal increments in both the mean F_1 score and OA. Besides, we note that our method surpasses other competitors in recognizing cars owing to its capacity of modeling global context.

4.4 Qualitative Results

Figure 4 (top two rows) shows a few examples of segmentation results. In the first row, only FCN+RANKING can successfully identify clutter in this complex scene. From the second row, it can be found that networks relying on either spatial propagation modules or memory-consuming reasoning modules could fail to recognize impervious surfaces in the shadow, while our model can predict more accurately.

4.5 Results on the Potsdam Dataset

In addition to comparisons on the Vaihingen dataset, we also compare our model with existing methods on the Potsdam dataset. Numerical results are shown in Table 4, and qualitative results are presented in Figure 4. We can observe that the integration of RANKING module contributes to increments of 2.17% and 2.01% in the mean F_1 score and overall accuracy, respectively, compared to FCN-dCRF. Besides, FCN+RANKING gains 0.57% higher mean F_1 score than FCN+non-local, while 0.53% better than FCN+GloRe.

Some qualitative results are present in Figure 4 (bottom two rows). As we can see in the third row, FCN, FCN-dCRF, and FCN+non-local tends to misclassify impervious surfaces into clutter, while our FCN+RANKING can make accurate predictions. Moreover, by referring to long-range dependencies, RANKING module is robust to visual ambiguities (e.g., a low vegetation-like roof in the forth row) and capable of correctly perceiving objects with complex appearances (e.g., a wave-shaped roof in the forth row).

5. CONCLUSION

In this paper, a computation- and memory-efficient network module, RANKING, is proposed for global context modeling by adaptively sample and aggregate a sparse set of features. As compared to existing self-attention modules, known as heavy computational and GPU memory overheads due to the calculation of dense pairwise relations, our module is lightweight but of high performance. Ablation studies have been carried out on two aerial scene parsing datasets to demonstrate the effectiveness of our module. The visualization of the adaptive random

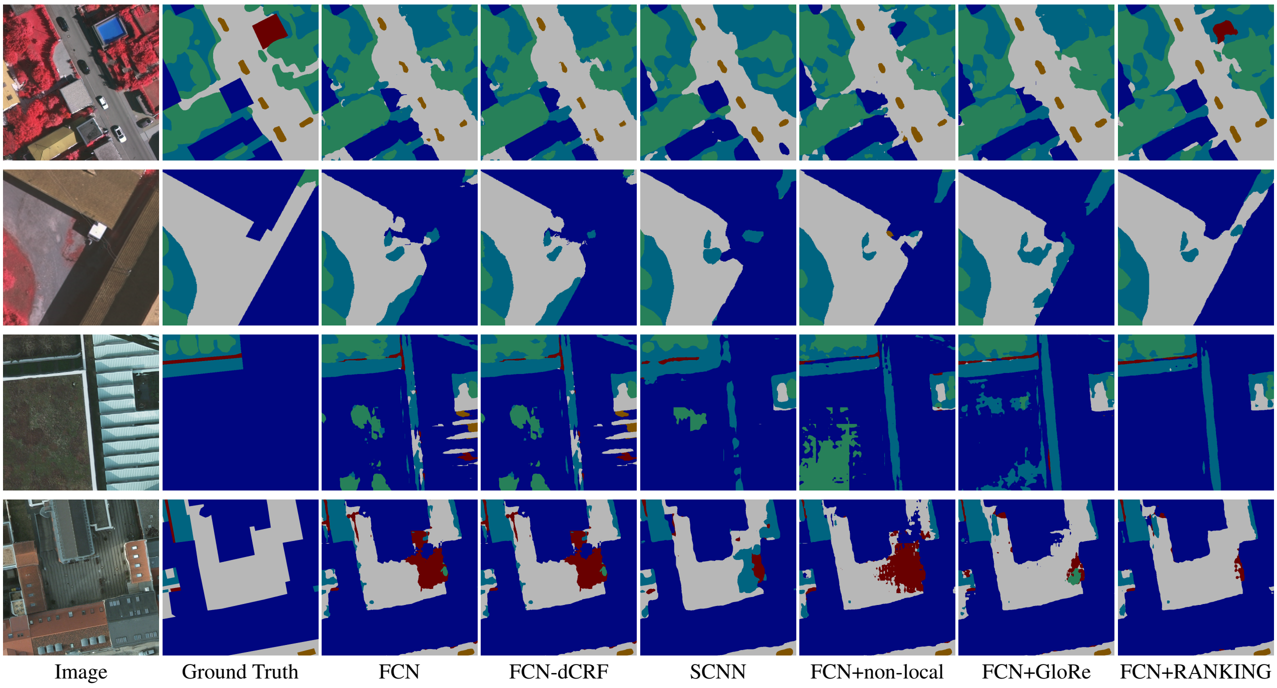

Figure 4. Examples of segmentation results on the Vaihingen (top two rows) and Potsdam (bottom two rows) dataset. Legend—gray: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, brown: cars, and red: clutter/background.

Table 4. Numerical Results on the Potsdam Dataset.

Model	Imp. surf.	Build.	Low veg.	Tree	Car	Clutter	mean F_1	mIoU	OA
FCN (Long et al., 2015)	88.61	93.29	83.29	79.83	93.02	69.77	84.63	78.34	85.59
Dilated FCN* (Chen et al., 2018a)	86.52	90.78	83.01	78.41	90.42	68.67	82.94	-	84.14
FCN-dCRF (Chen et al., 2018a)	88.62	93.29	83.29	79.83	93.03	69.79	84.64	**78.35**	85.60
FCN-FR* (Maggiori et al., 2017)	89.31	**94.37**	84.83	81.10	93.56	**76.54**	86.62	-	87.02
SCNN (Pan et al., 2018)	88.37	92.32	83.68	80.94	91.17	68.86	84.22	77.72	85.57
FCN+non-local (Yue et al., 2018)	89.90	93.64	85.11	81.44	93.96	73.40	86.24	76.51	87.01
FCN+GloRe (Chen et al., 2019)	**90.49**	93.89	85.23	**83.06**	94.60	72.99	86.71	77.27	87.45
FCN+RANKING-5-2	90.37	93.90	85.74	**83.06**	**94.60**	73.18	**86.81**	77.41	**87.61**
FCN+RANKING-3-2	90.26	93.88	**85.78**	83.04	94.58	72.91	86.74	77.32	87.56

walk sampling illustrates how the proposed RANKING module works. Furthermore, we evaluate our module by comparing a RANKING-equipped FCN with existing methods, and results suggest that our network can obtain competitive results while at low computational and memory overheads.

REFERENCES

Azimi, S., andL. Sommer, C. H., Schumann, A., Vig, E., 2019. Skyscapes fine-grained semantic understanding of aerial scenes. *IEEE International Conference on Computer Vision (ICCV)*. 2

Bell, S., Zitnick, C. L., Bala, K., Girshick, R. B., 2016. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1

Bertasius, G., Torresani, L., Yu, S. X., Shi, J., 2017. Convolutional random walk networks for semantic image segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1

Blaschke, T., Burnett, C., Pekkarinen, A., 2004. Image segmentation methods for object-based analysis and classification. *Remote Sensing Image Analysis: Including the spatial domain*, Springer, 211–236. 2

Cadène, R., Ben-younes, H., Cord, M., Thome, N., 2019. MUREL: multimodal relational reasoning for visual question answering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2015. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *International Conference on Learning Representations (ICLR)*. 1

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2018a. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848. 1, 5, 6

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. *European Conference on Computer Vision (ECCV)*. 1

Chen, Y., Rohrbach, M., Yan, Z., Yan, S., Feng, J., Kalantidis, Y., 2019. Graph-based global reasoning networks. *IEEE Con-*

ference on Computer Vision and Pattern Recognition (CVPR). 2, 5, 6

Cheng, D., Liao, R., Fidler, S., Urtasun, R., 2019. Darnet: Deep active ray network for building segmentation. *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Dozat, T., 2015. Incorporating Nesterov momentum into Adam. 4

Duan, Y., Zheng, Y., Lu, J., Zhou, J., Tian, Q., 2019. Structural relational reasoning of point clouds. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1, 2

Gerke, M., 2015. *Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)*. 5

Hamilton, W. L., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems (NIPS)*, 1024–1034. 1

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Computation*, 9(8), 1735–1780. 3

Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y., 2018. Relation networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1, 2

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W., 2019. CCNet: Criss-cross attention for semantic segmentation. *IEEE International Conference on Computer Vision (ICCV)*. 2

Kellenberger, B., Marcos, D., Tuia, D., 2019. When a few clicks make all the difference: Improving weakly-supervised wildlife detection in uav images. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2

Li, Z., Wegner, J., Lucchi, A., 2019. Topological map extraction from overhead images. *IEEE International Conference on Computer Vision (ICCV)*. 2

Liu, J., Liu, Y., 2014. Local regularity-driven city-scale facade detection from aerial images. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Liu, S., Mello, S. D., Gu, J., Zhong, G., Yang, M., Kautz, J., 2017. Learning affinity via spatial propagation networks. *Advances in Neural Information Processing Systems (NIPS)*. 1

Liu, Z., Li, X., Luo, P., Loy, C. C., Tang, X., 2015. Semantic image segmentation via deep parsing network. *IEEE International Conference on Computer Vision (ICCV)*. 1

Lobry, S., Murray, J., Marcos, D., Tuia, D., 2019. Visual question answering from remote sensing images. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. 1

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. 5, 6

Luo, W., Li, Y., Urtasun, R., Zemel, R. S., 2016. Understanding the effective receptive field in deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*. 1

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. High-resolution aerial image labeling with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12), 7092–7103. 4, 5, 6

Marcos, D., Tuia, D., Kellenberger, B., Bai, L. Z. M., Liao, R., Urtasun, R., 2018a. Learning deep structured active contours end-to-end. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Marcos, D., Volpi, M., Kellenberger, B., Tuia, D., 2018b. 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. 4, 5

Mou, L., Hua, Y., Zhu, X. X., 2019. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Pan, X., Shi, J., Luo, P., Wang, X., Tang, X., 2018. Spatial as deep: Spatial CNN for traffic scene understanding. *AAAI Conference on Artificial Intelligence (AAAI)*. 5, 6

Predoehl, A., Morris, S., Barnard, K., 2013. A statistical model for recreational trails in aerial images. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Quang, N., Thuy, N., Sang, D., Binh, H., 2015. An efficient framework for pixel-wise building segmentation from aerial images. *International Symposium on Information and Communication Technology, ACM*. 5

Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P. W., Lillicrap, T., 2017. A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems (NIPS)*. 1, 2

Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv:1606.02585*. 4

Shvets, M., Liu, W., Berg, A. C., 2019. Leveraging long-range temporal relationships between proposals for video object detection. *IEEE International Conference on Computer Vision (ICCV)*. 1

Sun, T., Di, Z., Che, P., Liu, C., Wang, Y., 2019. Leveraging crowdsourced gps data for road extraction from aerial imagery. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*. 1, 2

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2), 881–893. 4, 5

Wang, W., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Cheverie, J., Fidler, S., Urtasun, R., 2017. Torontocity: Seeing the world with a million eyes. *IEEE International Conference on Computer Vision (ICCV)*. 2

Wang, X., Girshick, R. B., Gupta, A., He, K., 2018. Non-local neural networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1, 2

Yuan, Y., Wang, J., arXiv:1809.00916. OCNet: Object context network for scene parsing. 2

Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., Xu, F., 2018. Compact generalized non-local network. *Advances in Neural Information Processing Systems (NeurIPS)*. 2, 5, 6

Zhang, L., Xu, D., Arnab, A., Torr, P. H., arXiv:1908.06955. Dynamic graph message passing networks. 2

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. 5

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P. H. S., 2015. Conditional random fields as recurrent neural networks. *IEEE International Conference on Computer Vision (ICCV)*. 1

Zhou, B., Andonian, A., Oliva, A., Torralba, A., 2018. Temporal relational reasoning in videos. *European Conference on Computer Vision (ECCV)*. 2

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2015. Object detectors emerge in deep scene CNNs. *International Conference on Learning Representations (ICLR)*. 1