remote sensing

MDPI

Article

SGA-Net: Self-Constructing Graph Attention Neural Network for Semantic Segmentation of Remote Sensing Images

Wenjie Zi †, Wei Xiong †, Hao Chen *,†, Jun Li and Ning Jing

Department of Cognitive Communication, College of Electronic Science and Technology, National University of Defense Technology, Changsha 410000, China; ziwenjie@nudt.edu.cn (W.Z.); xiongwei@nudt.edu.cn (W.X.); junli@nudt.edu.cn (J.L.); ningjing@nudt.edu.cn (N.J.)

* Correspondence: hchen@nudt.edu.cn

† These authors contributed equally to this work.

Abstract: Semantic segmentation of remote sensing images is always a critical and challenging task. Graph neural networks, which can capture global contextual representations, can exploit long-range pixel dependency, thereby improving semantic segmentation performance. In this paper, a novel self-constructing graph attention neural network is proposed for such a purpose. Firstly, ResNet50 was employed as backbone of a feature extraction network to acquire feature maps of remote sensing images. Secondly, pixel-wise dependency graphs were constructed from the feature maps of images, and a graph attention network is designed to extract the correlations of pixels of the remote sensing images. Thirdly, the channel linear attention mechanism obtained the channel dependency of images, further improving the prediction of semantic segmentation. Lastly, we conducted comprehensive experiments and found that the proposed model consistently outperformed state-of-the-art methods on two widely used remote sensing image datasets.

Keywords: self-constructing graph; semantic segmentation; remote sensing

Citation: Zi, W.; Xiong, W.; Chen, H.; Li, J.; Jing, N. SGA-Net: Self-Constructing Graph Attention Neural Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2021**, *13*, 4201. https://doi.org/10.3390/rs13214201

Academic Editor: Filiberto Pla

Received: 5 September 2021
Accepted: 15 October 2021
Published: 20 October 2021

1. Introduction

Semantic segmentation of remote sensing images aims to assign each pixel in an image with a definite object category [1], which is an urgent issue in ground object interpretation [2]. It has become one of the most crucial methods for traffic monitoring [3], environmental protection [4], vehicle detection [5], and land use assessment [6]. Remote sensing images are usually composed of various objects, highly imbalanced ground, and intricate variations in color texture, which bring challenges to the semantic segmentation of remote sensing images. Before the time of deep learning to display the distribution of vegetation and land cover, the superpixel was often used as measure for drawing features from multi-spectral images. However, hand-crafted descriptors are challenging tthe flexibility of these indices.

The convolutional neural network (CNN) [7] is widely used for the semantic segmentation of images. To achieve a better performance, CNN-based models regularly use multi-scale and deep CNN architectures to acquire information from multi-scale receptive fields and derive local patterns as much as possible. Owing to the restriction of the convolutional kernel, CNN-based models can only capture the dependency of pixels from the limited receptive field rather than the entire image.

CNN-based models have no ability to model the global dependency of each two pixels. However, a graph includes the connection of two nodes, so a graph neural network-based (GNN-based) model can capture the long-range global spatial correlation of pixels. There is no doubt that the traditional form of an image can be converted to a graph structure [8]. In this way, the graph can model the spatial relationship of each two pixels. In contrast, CNN can only obtain information from the limited receptive field. The adjacency matrix of

GNNs can represent the global relationship of images, which can contain more information than CNN-based models. Hence, we adopted a GNN to carry out semantic segmentation.

Nevertheless, a GNN does not ultimately demonstrate a strong point and is seldom used for dense prediction tasks because of the lack of prior knowledge of the adjacency matrix. Previous attempts [9–11] used prior knowledge-based manually generated static graphs, which did not fit each image well. A graph obtained by a neural network, is called "A self-constructing graph". Compared with these methods, a self-constructing graph can adjust itself and reflect the features of each remote sensing image.

Attention mechanisms [12] are added within the convolutional frameworks to improve the semantic segmentation performance in remote sensing images. Every true color image has RGB channels, and the RGB channels of objects have a potential correlation, which can be used to get a better semantic segmentation. The convolutional block attention module (CBAM) [13] adopts two kinds of non-local attention modules to the top of the atrous convolutional neural network: channel attention and spatial attention, respectively. CBAM achieves a competitive segmentation performance in the corresponding dataset. The channel attention mechanism can acquire the correlation among channels, improving the performance of semantic segmentation in remote sensing images. Every pixel has several channels, and each has a different importance for different kinds of pixels. Our channel attention mechanism could model the channels correlation to a large extent, inhibiting or enhancing the corresponding channel in different tasks, respectively.

In this paper, we propose a self-constructing graph attention neural network (SGA-Net) to implement the semantic segmentation of remote sensing images to model global dependency and meticulous spatial relationships between long-range pixels. The main contributions of this paper are as follows:

- Incorporating GATs into self-constructing graphs enhances long-range dependencies between pixels.
- A channel linear attention mechanism to catch th correlation among channel outputs of the graph neural network and further improve performance of the proposed GNN-based model.
- Comprehensive experiments on two widely used datasets in which our framework outperformed the state-of-the-art approaches on the F1 score and mean IoU .

The rest of this paper is organized as follows, the related work is showed in Section 2. Section 3 presents that the details of our architecture SGA-Net. The experiments and corresponding analyses are showed in Section 4, and Section 5 presents the conclusion.

2. Related Work

2.1. Semantic Segmentation

The rise of convolutional neural networks (CNNs) marks a significant improvement in semantic segmentation. The fully convolutional network (FCN), which widely consists of the encoder–decoder module has dominated pixel-to-pixel semantic segmentation [14]. The FCN dominates semantic segmentation., and one with an encoder-decoder module can segment images at the pixel level by deconvolutional and upsampling layers, promoting the development of semantic segmentation. Compared with the FCN, the U-Net [15] applies multi-scale strategies to withdraw contextual patterns and perform semantic segmentation better. Owing to the use of multi-scale context patterns, U-Net can derive a better prediction result than the FCN. Segnet [16] proposes max-pooling indices to enhance location information, which can improve segmentation performance. Deeplab V1 [17] proposes atrous convolutions, which can enlarge the receptive field without increasing the number of parameters. Compared with Deeplab V1, Deeplab V2 [18] presents atrous spatial pyramid pooling (ASPP) modules that consist of atrous convolutions with different sampling rates. Because it uses information from a multi-scale rates receptive field, Deeplab V2 has better prediction than Deeplab V1. The above methods are all supervised models. FESTA [19] is a semi-supervised learning CNN-based model that encodes and regularizes image features and spatial relations. Compared to FESTA, our proposed method extracts

long-range spatial dependency and channels correlation to perform segmentation, and our proposed method is a GNN-based model. There are also models of non-grid convolutions for semantic segmentation. Deformable convolution [20] adds 2D offsets to the regular grid sampling locations in the standard convolution, which enhances the geometric transformation modeling capability of CNN. Deformable convolution is still limited in capturing long-range structured relationships. DGMN [21] obtains long-range structured relationships by constructing a dynamic graph. Our proposed model also adopts the idea of a dynamic graph to obtain global long-range correction of remote sensing images. HG-CNNs [22] is a heterogeneous grid convolutional neural network that constructs a data-adaptive graph structure from the convolutional layer by microclustering and assembling features into the graph. Our proposed model also constructs a data-adaptive graph, but the graph structure is extracted by convolutional operation from the high-level feature map.

2.2. Graph Neural Network

Recently, the GNN has become popular due to its success in many fields, such as natural language processing [23], social networks [24], reinforcement learning [25], computer vision [26]. There are lots of natural datasets of graph structures, recommender systems [27], protein networks [28] and knowledge graphs [29]. More and more GNN variants are produced and applied to various fields. In the beginning, only datasets in the form of graphs [10,30] were entered into graph neural networks. However, in a GNN neatly arranged matrix forms like remote sensing images can be extracted and transformed into diffferent kinds of graph structures [8]: convolutional networks, auto-encoders, attention networks (GATs) and isomorphism networks [31]. A GAT [32] and GCN are crucial branches of a GNN. Gao et al. [33] performed action recognition by using structured prior knowledge in the form of knowledge graphs. Yan et al. [34] completed skeleton-based action recognition with spatial-temporal graph convolutional networks (STGCNs) that auto-learn spatial and temporal patterns. Wang et al. [35] proposed a graph-based, language-guided attention mechanism that can clearly reveal inter-object properties and relationships with flexibility. GNN-based models (ASTGCN) [36] are used to predict traffic flow. Liu et al. [8] adopted a GCN to conduct experiences of semantic segmentation in remote sensing images, and the GCN adjacency matrix is built by neural networks. A GCN can simultaneously perform end-to-end learning of node feature information and structure information. In comparison, a GAT proposes a weighted summation of neighboring node features using an attention mechanism. The weights of neighboring node features entirely depend on the node features and are independent of the graph structure. GraphSAGE [37] solves the GCN and GAT memory explosion problem by neighbori sampling for the large-scale graph. GNN-based models are used in a variety of applications.

2.3. Attention Mechanisms

With the publication of the paper in [12], attention mechanisms became more and more popular and attractive. Fu et al. [38] propose a dual attention network (DANet) that can adaptively learn local and global dependency to conduct semantic segmentation. Huang et al. [39] propose channelized axial attention (CAA) to integrate channel and axial attention seamlessly. CAA is similar to DANet in double-attention mechanisms, and these models have a competitive result in the corresponding dataset. CAA pays attention to channel and axial attention, DANet focuses on local and global attention. Compared with multi-attention mechanism, Tao et al. [40] propose a multi-scale attention mechanism that improves the accuracy of semantic segmentation. Transformer [12] is used to solve natural language processing, which is entirely based on the multi-head self-attention mechanism. Dosovitskiy et al. [41] adopt a transformer into the task of image classification, achieving excellent prediction results in many small- and medium-image recognition benchmarks.

3. Methods

In this section, we introduce the details of the model SGA-Net. An overview of the framework is presented in Figure 1 and consists of a feature maps extraction network, self-constructing graph attention network and a channel linear attention mechanism. The four SGA-Nets are shared weights. First, ResNet50 was employed as the backbone of the feature extraction network to acquire feature maps of remote sensing images, and X was denoted as the feature maps. Second, to ensure geometric consistency, feature maps were rotated by several degrees—90, 180 and 270. In addition, X_{90}, X_{180} and X_{270} indicated the feature maps multi-views, where the index was the degree rotation. Third, multi-view feature maps were used to obtain self-constructing graphs A_0, A_1, A_2 and A_3 by a convolution neural network, separately. Fourth, these self-constructing graphs were fed into a neural network based on a GAT to extract the long-range dependency of pixels. Fifth, This network is called the self-constructing graph attention network and the outputs were used for inputs into channel linear attention, the ouputs of which were added to predict the final results. The adjacency matrix A is a high-level feature map of the corresponding remote sensing image feature map, and the projected remote sensing features maps in a specific dimension are defined as nodes. Therefore, the features maps X are defined as the features of nodes. A_{ij} indicating the weight of the edge between node i and node j. We focused on the SGA-Net below.

3.1. Self-Constructing Graph Attention Network

The self-constructing graph is an undirected graph that shows the spatial similarity relationship of feature maps in remote images. The self-constructing graph is extracted by a neural network, instead of prior knowledge. Every image is unique; thus, models based on a self-constructing graph can be fitted for each remote sensing image very well.

The input image is denoted as I, where $I \in \mathbb{R}^{C \times H \times W}$, H and W present the hight and width of corresponding image respectively, and C denotes the number of channels. The high-level feature maps is used as X, where $X \in \mathbb{R}^{H^{'} \times W^{'} \times C^{'}}$, $H^{'}$, $W^{'}$ and $C^{'}$ indicate that the number of height, width and channels, respectively. Next, we applied a convolutional neural network and dropout layer to extract the latent embedding space S of every remote sensing image, where $S \in \mathbb{R}^{N \times E}$, $N = H^{'} \times W^{'}$, where E is the number of the classification.

As we can see from Figure 2, which shows the latent embedding space S of buildings, cars, roads, trees and grass, respectively. S of buildings indicated that they are brighter than other objects: the higher the gray value, the greater the spatial similarity. In general, the same kind of features have the greatest spatial similarity relationship. The adjacency matrix was defined as $A = \mathrm{ReLU}(\mathrm{matmul}(S, S^T))$, which highlighted and enhanced the differences between the target class and other categories. Since it does not arise from prior knowledge, but directly from the output of neural network the adjacency matrix is called the "self-constructing adjacency matrix ", which captures the distributions of the features in remote sensing images. Our model followed the convention of the variational auto-encoder [42] to learn the mean matrix M and the standard deviation matrix D, where $M \in \mathbb{R}^{N \times E}$ and $D \in \mathbb{R}^{N \times E}$, and E denotes the number of the classification. The details of the mean matrix M and logarithm of the standard deviation matrix D are as follows:

$$\begin{aligned} M' &= \text{ Flatten } \left(\mathrm{Conv}_{3\times 3,\ \mathrm{padding}\ =1}(X)\right) \\ M &= \mathrm{Dropout}(\mathrm{p}=0.2)(M') \end{aligned} \tag{1}$$

$$\begin{aligned} D' &= \text{ Flatten } (\mathrm{Conv}_{1\times 1})(X) \\ \log(D) &= \mathrm{Dropout}(\mathrm{p}=0.2)(D') \end{aligned} \tag{2}$$

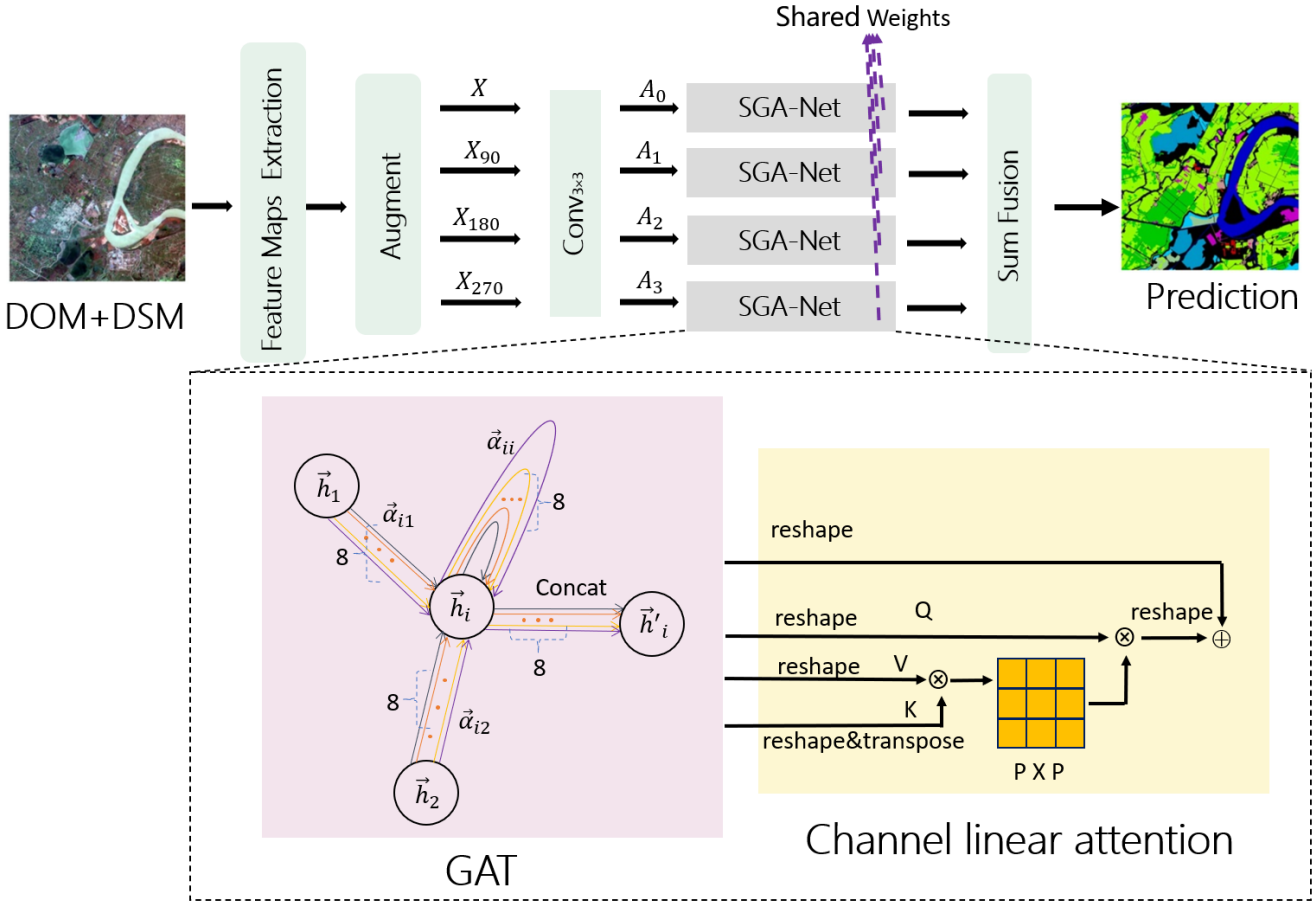

Figure 1. In the flow chart of our model for semantic segmentation, ResNet50 was selected as the feature maps extraction network of our model; $Conv_{3\times3}$ means the convolution operation with kernel size 3; SGA-Net denotes the self-constructing graph attention network and channel linear attention mechanism; GAT is graph attention network, and Q, K, V of channel linear attention mechanism indicate query, key and value, respectively. X denotes the feature input, X_{90}, X_{180} and X_{270} indicate the feature maps multi-views, where the index is the rotation degree, and A_0, A_1, A_2 and A_3 present the adjacency matrix of the self-constructing graph of corresponding feature maps. $\vec{h}_i$ means initial feature vector of each node, where i $\in$ [1, 3]; $\vec{\alpha}$ represents the correlation coefficient; Concat denotes a concatenating operation; P indicates the number of channels, and $\vec{h}_i'$ indicates the output of self-constructing graph attention neural network.

The latent embedding space $S = M + \log(D) \cdot \alpha$, where $\alpha \in \mathbb{R}^{N\times E}$ is an auxiliary noise variable that obeys standard normal distribution ($\alpha \sim \mathcal{N}_{N\times E}(\mathbf{0}, \mathbf{I})$). The adjacency matrix A was generated by an inner product operation between the transpose of the latent space embedding S^T and itself S, where $A \in \mathbb{R}^{N\times N}$ and A_{ij} denotes the spatial similarity relationship between node i and j.

$$A = \mathrm{ReLU}(\mathrm{matmul}(\mathrm{S}, \mathrm{S}^T)) \tag{3}$$

A therefore can indicate the spatial similarity relation of each two nodes of the latent embedding space S. However, the CNN receptive field was restricted by the kernel size, and the CNN did not have the ability to present a spatial similarity relation between each two nodes. A in our model is not traditional binary but weighted and undirected.

The calculation of the SGA-Net was the same as for all kinds of attention mechanisms. The first step was computing the attention coefficient, and the last was aggregating the sum of weighted features [12]. For node i, the similarity coefficient between its neighbour nodes j and itself was calculated, where $i \in \mathbb{N}$ and $j \in \mathbb{N}$. The details of the similarity coefficient are as follows:

$$e_{ij} = \mathbf{a}([U \cdot \vec{h}_i, U \cdot \vec{h}_j]) \tag{4}$$

where U is the learnable weight matrix, $\vec{h}_i$ indicates the node feature of node i, $h = (\vec{h}_1, \vec{h}_2, \cdots, \vec{h}_N)$, $\vec{h}_i \in \mathbb{R}^{N \times F}$, where F denotes the number of features in each node and $\vec{h} = X$, and **a** indicates the operation of self-attention, which is inner product, and the self-constructing adjacency matrix A is set as a mask. Thus, $e_{ij} \in \mathbb{R}^{N \times N}$. Next, we computed the attention coefficient $\vec{\alpha}_{ij}$ as follows:

$$\vec{\alpha}_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(e_{ij}\right)\right)}{\sum_{k \in N} \exp(\text{LeakyReLU}(e_{ik}))} \tag{5}$$

We applied an 8-head graph attention network to enhance the predictive capability of the model and make it more stable iduring training to improve the framework performance.

$$\vec{h}'_i = \|_{l=1}^{L} \sigma\left(\sum_{j \in \mathcal{N}_i} \vec{\alpha}_{ij}^{k} U^k \vec{h}_j\right) \tag{6}$$

where $\|$ indicates the operation of concatenating, and L is the number of attention, *sigma* is the activate function sigmoid, and $\mathcal{N}_i$ indicates some neighborhood nodes of the node i in the graph, and $\vec{\alpha}_{ij}^{k}$ is the normalized attention coefficients computed by the kth attention mechanism $\mathbf{a}^{(k)}$, and the $U^{(k)}$ indicates the kth corresponding input weight matrix. Specifically, $L = 8$ and we use an 8-head graph attention network in the work.

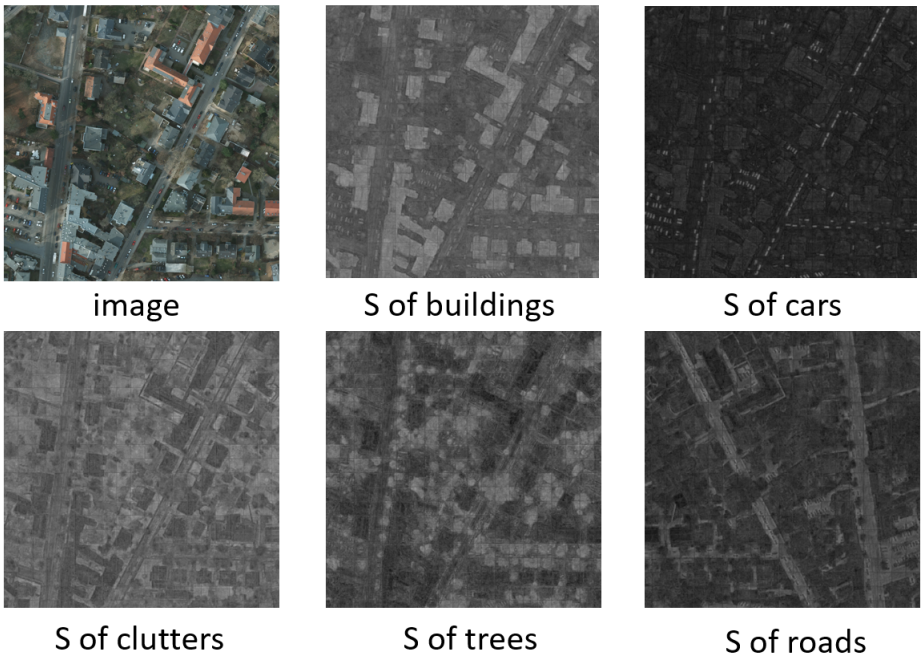

Figure 2. Latent embedding space of buildings, cars, roads, trees and low-vegetation present the latent embedding space of these categories separately.

3.2. Channel Linear Attention

Each channel of the high level features could be regarded as the special response of a category, and different responses have intrinsic independencies. The channels of each category had their own distinctive feature and correlations. Exploiting the inter-correlations among channels of images can improve the performance of specific semantic features. Therefore, we adopted a channel attention module to explore correlations among channels.

Suppose the query matrix is Q, the key matrix is K and the value matrix is V. In addition, all of Q, K and $V \in \mathbb{R}^{K \times P}$, where $P = H \times W$, and these are learnable parameters.

In addition, suppose the output of SGA-Net is $\vec{H}$, where $\vec{H} \in \mathbb{R}^{K \times P}$. The detail of the channel linear attention is as follows:

$$D(Q, K, V) = \vec{H} + \frac{V + \left(\frac{Q}{\|Q\|_2}\right)\left(\left(\frac{K}{\|K\|_2}\right)^T V\right)}{N + \left(\frac{Q}{\|Q\|_2}\right)\left(\frac{K}{\|K\|_2}\right)^T} \tag{7}$$

where N denotes the number of nodes. $D(Q, K, V) \in \mathbb{R}^{K \times P}$. The equation highlights the input of a GAT, and emphasizes the importance of the K, Q and V at the same time. The channel linear attention can model the importance of different channels in a different task.

3.3. Loss Function

There is no doubt that A_{ii} ought to be greater than 0 and close to 1; hence, we introduced a diagonal log regularization term to improve the prediction which was defined as:

$$\gamma = \sqrt{1 + \frac{n}{\sum_{i=1}^{n} A_{ii} + \epsilon}} \tag{8}$$

$$\mathcal{L}_{dl} = -\frac{\gamma}{n^2} \sum_{i=1}^{n} \log\left(|A_{ii}|_{[0,1]} + \epsilon\right) \tag{9}$$

where the subscript [0, 1] indicates that A_{ii} is clamped to [0, 1], and ϵ is a fixed and small positive tiny parameter and ($\epsilon = 10^{-5}$). We adopted the Kullback–Leibler divergence, which measures the difference between the distribution of latent variables and the unit Gaussian distribution [42] to be the part of loss function, and the details of Kullback–Leibler divergence were as follows:

$$\mathcal{L}_{kl} = -\frac{1}{2NK} \sum_{i=1}^{N} \sum_{j=1}^{K} \left(1 + \log\left(D_{ij}\right)^2 - M_{ij}^2 - \left(D_{ij}\right)^2\right) \tag{10}$$

where D is the standard deviation matrix. In addition, we adopted an adaptive multi-class weighting (ACW) loss function [26] to address the highly imbalanced distribution of the classes. The detail of $\mathcal{L}_{acw}$ is as follows:

$$\mathcal{L}_{acw} = \frac{1}{|Y|} \sum_{i \in Y} \sum_{j \in C} \tilde{w}_{ij} \cdot p_{ij} - \log\left(\text{MEAN}\{d_j \mid j \in C\}\right) \tag{11}$$

where Y includes all the labeled pixels and d_j denotes the dice coefficient:

$$d_j = \frac{2 \sum_{i \in Y} y_{ij} \tilde{y}_{ij}}{\sum_{i \in Y} y_{ij} + \sum_{i \in Y} \tilde{y}_{ij}} \tag{12}$$

where $y_{i,j}$ and $\tilde{y}_{i,y}$ denote the ijth ground truth and prediction of class j respectively. p_{ij} is positive and negative balanced factor of node i and node j and its detail as follows:

$$p = (y - \tilde{y})^2 - \log\left(\frac{1 - ((y - \tilde{y})^2)}{1 + (y - \tilde{y})^2}\right) \tag{13}$$

$\tilde{w}_{ij}$ is a weight about the frequency of all categories, and the detail of it as follows:

$$\tilde{w}_{ij} = \frac{w_j^t}{\sum_{j \in C}\left(w_j^t\right)} \cdot \left(1 + y_{ij} + \tilde{y}_{ij}\right) \tag{14}$$

$$w_j^t = \frac{\text{MEDIAN}\left(\left\{f_j^t \mid j \in C\right\}\right)}{f_j^t + \epsilon} \tag{15}$$

$$f_j^t = \frac{\hat{f}_j^t + (t-1) \cdot f_j^{t-1}}{t} \tag{16}$$

where ϵ is a fixed parameter and $\epsilon = 10^{-5}$; C indicates the number of class; t is the iteration number; f_j^t represents the pixel sum of class j at the tth training step, which can be computed as $\frac{\mathrm{SUM}(y_j)}{\sum_{j \in C} \mathrm{SUM}(y_j)}$, and when $t = 0$, $f_j^t = 0$.

For refining the final prediction result, we adopted the sum of three kinds of loss function as the final loss function in our framework, which are $\mathcal{L}_{kl}$, $\mathcal{L}_{dl}$, and $\mathcal{L}_{acw}$ respectively. The loss function can be formulated as below:

$$Loss = \mathcal{L}_{kl} + \mathcal{L}_{dl} + \mathcal{L}_{acw} \tag{17}$$

4. Experiments

4.1. Datasets

We used two public benchmark the ISPRS 2D semantic labeling contest datasets as our datasets. The ISPRS datasets consisted of aerial images in two German cities: Potsdam and Vaihingen. They are labeled with six common land cover classes:impervious surfaces, buildings, low vegetation, trees, cars and clutter.

- Potsdam: The Potsdam datasets (https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam/, accessed on 3 September 2021) comprised 38 tiles of a ground resolution of 5 cm with size 6000 × 6000 pixels. Moreover, these tiles consisted of four channel images—Red-Green-Blue-Infrared (RGB-IR)—and the dataset contained both digital surface model (DSM) and normalized digital surface model (nDSM) data. Of these tiles, 14 were used as hold-out test images: 2 were used as validation images, and 12 were used as training data. Furthermore, to compare with other models fairly, we only used RGB images as experience data in this paper.
- Vaihingen: The Vaihingen dataset(https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/, accessed on 3 September 2021) consists of 33 tiles of varying size with a ground resolution of 9cm, of which 17 tiles are used as hold-out test images, 2 tiles are used as validation set, and the rest tiles are taken as training set. In addition, these tiles contain Infrared-Red-Green (IRRG) 3-channel images. In addition, the dataset includes DSM and nDSM. To compare other works fairly, we only apply 3-channel IRRG data in these frameworks in this paper.

4.2. Evaluation Metrics

To acquire reasonable and impartial results, we adopted the mean Intersection over Union (mIoU), the F1 score (F1) and accuracy (Acc) to evaluate performance, all of which are widely applied in semantic segmentation. In addition, based on the accumulated confusion matrix, these evaluation indicators were computed as:

$$\mathrm{mIoU} = \frac{1}{N} \sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k + FN_k}, \tag{18}$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}, \tag{19}$$

$$\mathrm{Acc} = \frac{\sum_{k=1}^{N} TP_k + TN_k}{\sum_{k=1}^{N} TP_k + FP_k + TN_k + FN_k} \tag{20}$$

where TP_k, FP_k, TN_k, and FN_k are the true positive, false positive, true negative, and false negatives, respectively, and k indicates the number of object index. Acc was computed for all categories except for clutter.

4.3. Experimental Setting

We achieved the proposed SGA-Net as well as all baselines working with PyTorch on a Linux cluster. Models were trained in a single Nvidia GeForce RTX 3090 with a batch size of 5. We applied AMSGrad [43] with adam as the optimizer with weight decay 2×10^{-5}. The weight decay was used in all learnable parameters except batch-norm and bias parameters. Polynomial learning rate (LR) decay was $\left(1 - \frac{cur-iter}{max-iter}\right)^{0.9}$ with the maximum iterations of 10^8, and learning rate decay set to 0.9. The learning rate of the bias parameters is $2 \times$ LR. The initial learning rate was set to $\frac{1.5 \times 10^{-4}}{\sqrt{3}}$. We sampled the patches of size 512×512 as input, and set the node size of graph to 1024×1024.

4.4. Baselines and Comparison

Our model was compared with several works as follows:

- **DDCM** [44]: This is a CNN-based model that consists of dense dilated convolutions merged with varying dilation rates. It can enlarge the receptive fields effectively. Moreover, this model can obtain fused global and local context information to raise the discriminative capability for the surroundings.
- **MSCG-Net** [26]: This method is a self-constructing graph convolutional network that applies neural networks to build graphs from the input of high-level features instead of prior knowledge. In addition, it is a GNN-based model. The feature maps extraction network of our entire framework was similar to a MSCG-Net, but our model used a self-constructing graph to input a GAT, and its outputs were input channel linear attention.
- **DANet** [45]: This framework includes the position and the channel attention mechanisms. The position attention mechanism can learn the spatial relationship of features, and the channel attention mechanism can obtain the channel dependency of images. It is an attention-based method.
- **DUNet** [46]: The model uses redundancy in the label space of semantic segmentation and can recover the pixel-level prediction from low-resolution results of CNNs. It is a CNN-based model.
- **DeeplabV3** [47]: This method captures multi-scale backgrounds by multi-scale cascading or parallel dilated convolution, which can improve the prediction of semantic segmentation. In addition, it is a CNN-based framework.

4.4.1. Prediction on Potsdam Dataset

We compared our model with five baselines on the Potsdam dataset. Table 1 presents the evaluation metrics of prediction in semantic segmentation. Obviously, Table 1 shows that the proposed SGA-Net outperformed the other models.

The SGA-Net was 3.4% higher than the MSCG-Net in mean F1 score, because a self-constructing graph attention network can acquire long-range global spatial dependency of images and channel linear attention to obtain a correlation among all channels. In addition, the proposed framework outperformed other model, which showed that the self-constructing graph had the ability to extract the spatial dependency of images well. In fact, we applied a self-constructing graph, obtained by neural network rather than prior knowledge, to a GAT. Our model performed better than DANet for prediction in all categories, indicating that a self-constructing graph attention neural network can dig the global long-range spatial correlation of nodes for the channel linear attention. Moreover, the multiviews of feature maps in remote sensing images can ensure the geometric consistency of spatial patterns. The reasons for the 3% improvement in average F1 score and 2.6% improvement in mIoU of SGA-Net over Deeplab V3 were that the self-constructing graph neural network obtaied the spatial similarity of each two nodes, and the channel linear attention mechanism captured the correlation among the channel outputs of the graph neural network. The GAT modeled the dependencies between each two nodes, thereby increasing information entropy about spatial correlation. The channel linear attention

mechanism enhanced or inhibited the corresponding channel in different tasks. Furthermore, multi-views also can get more information about initial images, which has the ability to support predicting remote sensing images.

Table 1. The experimental results on the Potsdam dataset (bold: best; underline: runner-up).

Method	Road Surf	Buildings	Low Veg.	Trees	Cars	Mean F1	Acc	mIoU
MSCG-Net (GNN-based)	0.907	0.926	0.851	0.872	0.911	0.893	0.959	0.807
DANet (Attention-based)	0.907	0.922	0.853	0.868	0.919	0.894	0.959	0.807
Deeplab V3 (CNN-based)	0.905	0.924	0.850	0.870	0.939	0.897	0.958	0.806
DUNet (CNN-based)	0.907	0.925	0.853	0.869	0.935	0.898	0.959	0.808
DDCM (CNN-based)	0.901	0.924	0.871	0.890	0.932	0.904	0.961	0.808
SGA-Net (GNN-based)	**0.927**	**0.958**	**0.886**	**0.896**	**0.968**	**0.927**	**0.964**	**0.832**

Figure 3 shows the ground truth and predictions of all methods in tile5_15, and trhat the SGA-Net overmatched all baselines in the Potsdam dataset. The figure shows the overall predicting capability of our method in remote sensing images. For example, our model predicted surfaces better than that of MSCG-Net, while the proposed model outperformed all baselines in predicting buildings. The above phenomena illustrated that our framework modeled regularly shaped grounds well. Figure 4 is the result of predicting details from all baselines and the SGA-Net. The black boxes highlight the difference of results among ground truth, baselines and the SGA-Net. The first row shows that the proposed framework did much better predicting buildings compared to the other models, demonstrating that the SGA-Net can model global spatial dependency and channel correlation of remote sensing images.

The second row shows that the SGA-Net outperformed all baselines in predicting trees and buildings, which indicates that the SGA-Net can extract channel correlation in images well. The third row shows that the SGA-Net surpassed the other frameworks in predicting surfaces and low-vegetation. In addition, the last row shows that our model was superior to the other models for predicting trees and low-vegetation. The above phenomena illustrate that self-constructing graph attention network can capture long-range global spatial dependency of images, and the channel linear attention mechanism can acquire a correlation of images among channels. In addition, multiviews feature maps can ensure geometric consistency, improving the performance of predicting semantic segmentation in remote sensing images.

In conclusion, Figure 4 shows that the SGA-Net had a better performance predicting buildings, trees, low-vegetation, cars and surfaces in detail, demonstrating SGA-Net has powerful prediction in the semantic segmentation of remote sensing images.

4.4.2. Prediction on Vaihingen Dataset

We compared our framework with these five baselines on Vaihingen dataset, Table 2 presents the evaluation metrics of prediction in all models. The result showed that the mean F1 score of the SGA-Net was higher than that of the other methods, indicating the powerful ability of prediction in remote sensing images.

To be specific, the F1 score of our model for road surfaces, buildings and cars exceeded all baselines, and accuracy was higher than in other models. Because the SGA-Net contains a self-constructing graph attention neural network and a channel linear attention mechanism, the framework can model the spatial dependency and channel correlation of remote sensing images. Furthermore, because the self-constructing graph attention neural network has the ability to obtain a long-range global spatial correlation of the regular grounds, the predicting result of buildings and cars from the SGA-Net surpassed all baselines. The

reason for bad performance on low-vegetation and trees is that the two kinds of grounds are surrounded by many others, leading to poor extraction of spatial dependency by the self-constructing graph. The similarity of tree colors to low-vegetation and the fact that the SGA-Net captures long-range dependencies results in a segmentation performance for trees that is slightly worse than some other methods. The distribution of low-vegetation is more scattered than other objects, and the proposed model cannot extract a very complex spatial relationship of low-vegetation, leading to a poorer performance than DDCM in semantic segmentation.

Table 2. The experimental results on the Vaihingen dataset (bold: best; underlined: runner-up).

Method	Road Surf	Buildings	Low Veg.	Trees	Cars	Mean F1	Acc	mIoU
MSCG-Net (GNN-based)	0.906	0.924	0.816	0.887	0.820	0.870	0.955	0.796
DANet (Attention-based)	0.905	0.934	**0.833**	0.887	0.761	0.859	0.955	0.797
Deeplab V3 (CNN-based)	0.911	0.927	0.819	0.886	0.818	0.872	0.956	0.800
DUNet (CNN-based)	0.910	0.927	0.817	0.887	0.843	0.877	0.955	0.801
DDCM (CNN-based)	0.927	0.953	**0.833**	**0.890**	0.883	0.898	0.963	**0.828**
SGA-Net (GNN-based)	**0.932**	**0.955**	0.826	0.884	**0.928**	**0.905**	**0.965**	0.826

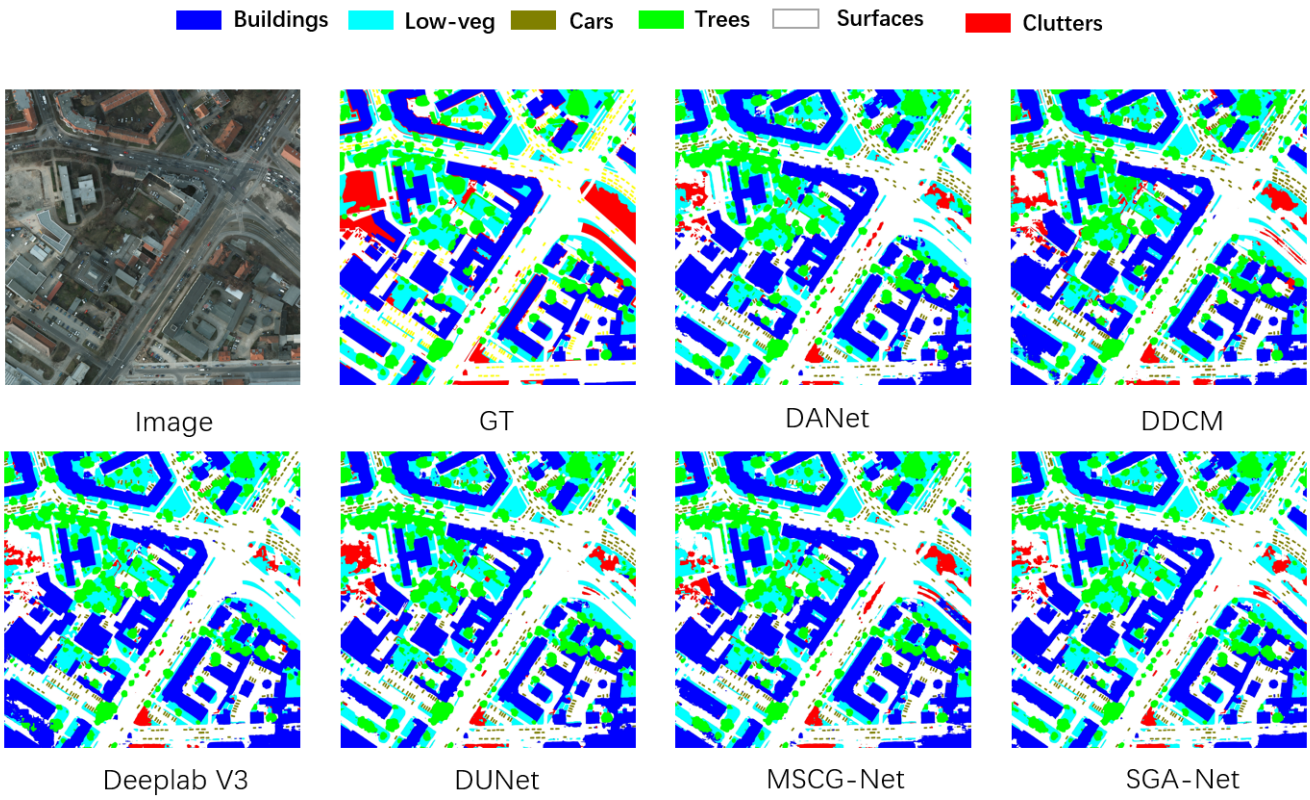

Figure 3. Visualization of tile5_15 in the Potsdam dataset.

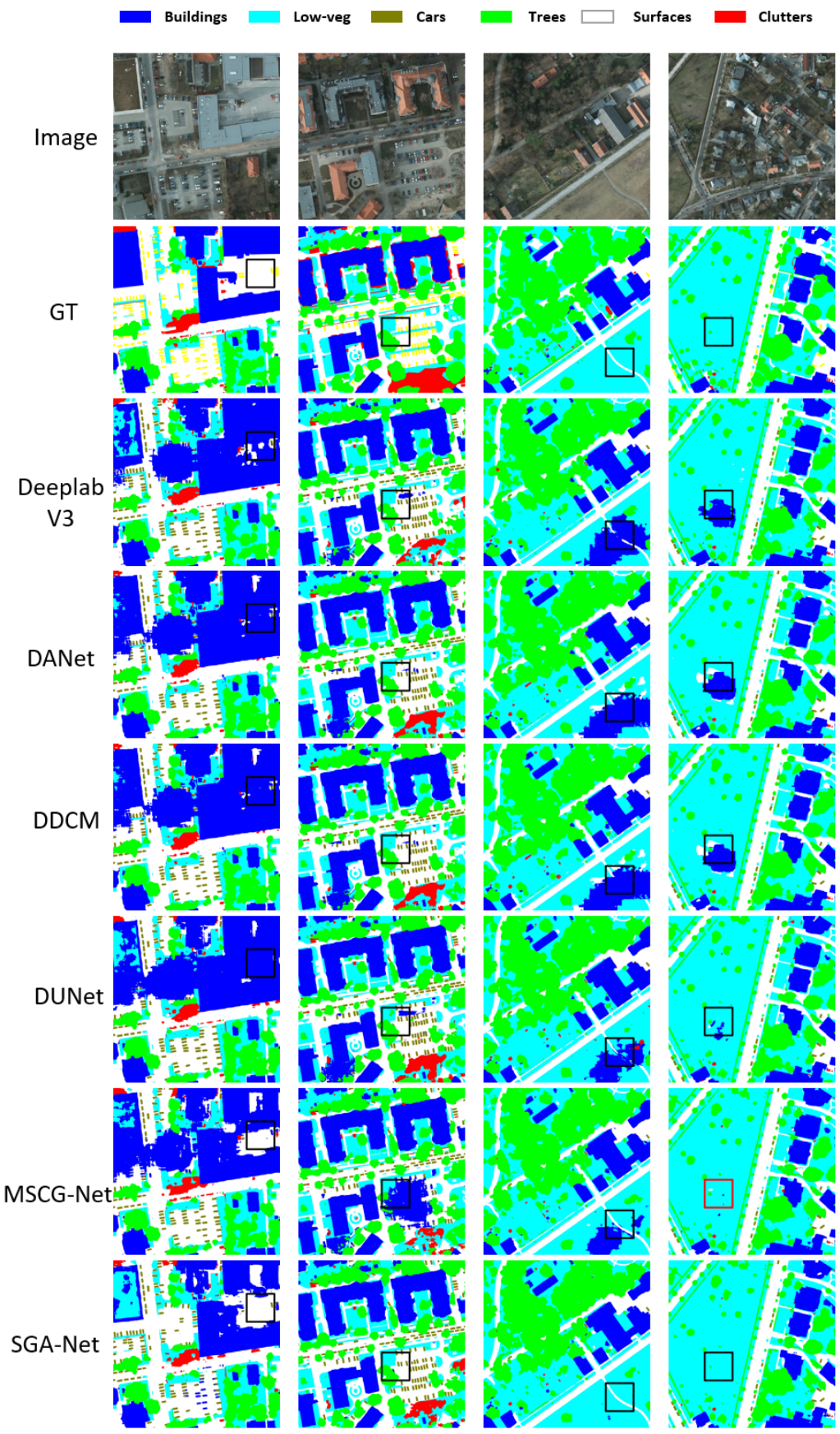

Figure 4. Visualization of prediction detail in the Potsdam dataset.

In addition, Figure 5 shows that the proposed model had a good overall prediction performance. In particular, this figure distinctly indicates that the predicting results of buildings and cars from the SGA-Net surpassed all models, showing that multi-views feature maps can enhance prediction capability, and a self-constructing graph can mine long-range spatial dependency for each image. Additionally, Figure 6 shows the details of the prediction results of the Vaihingen dataset. Because the self-constructing graph attention network can acquire the spatial dependency of each two nodes, the top three rows of Figure 6 indicate that the predictive buildings of the SGA-Net performed better than all baselines, and the last row shows that the predicting trees of our model were much better than other frameworks.

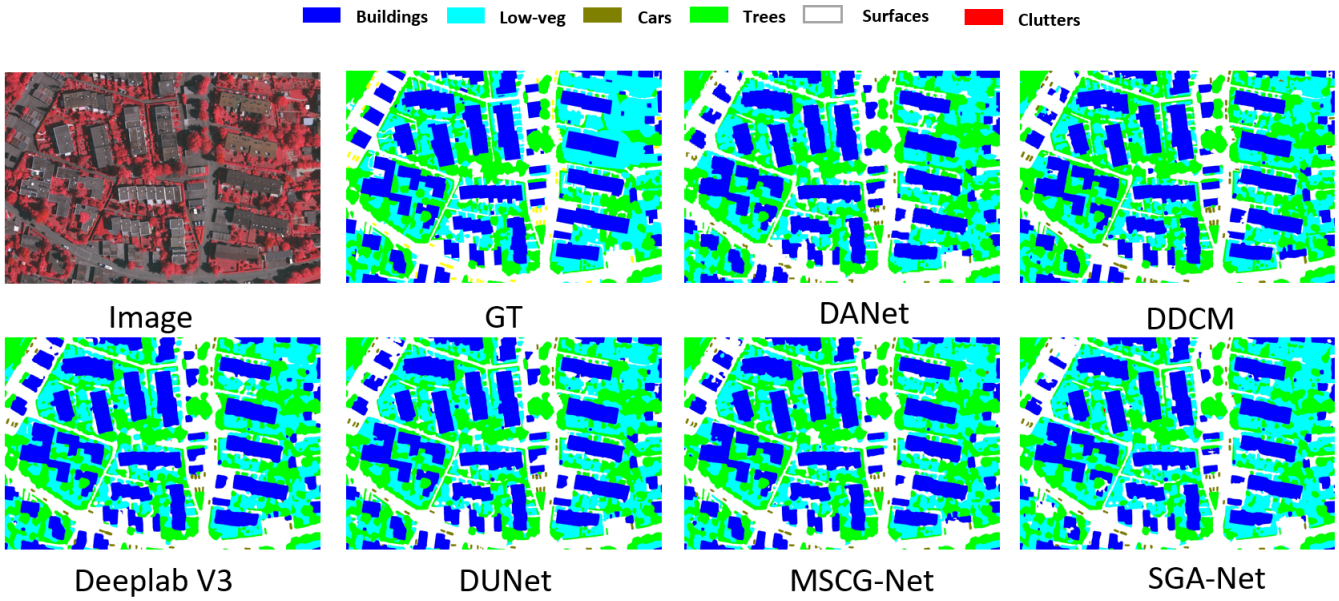

Figure 5. Visualization of tile35 in the Vaihingen dataset.

4.5. Ablation Studies

We conducted ample ablation experimentation to prove the effectiveness of the self-constructing graph neural network and channel linear attention mechanism (SGA-Net) in the proposed framework. Following the main experience as closely as possible, ResNet50 was selected as the baseline and feature extraction layers in our framework. To research the effectiveness of each model component further, we compared the SGA-Net with its variants as follows:

- ResNet50 [48]: a CNN-based neural network adopted as the feature extraction component of the proposed model.
- SGA-Net-ncl: To validate the effectiveness of the self-constructing graph neural network, we directly removed the channel linear attention mechanism from the framework.
- SGA-Net-one: To validate the effect of geometric consistency, we removed the branch roads of X_{90}, X_{180} and X_{270}.
- SGA-Net: our whole SGA-Net framework .

As can be seen from Table 3, the performance of the SGA-Net-ncl significantly overmatched the baseline of ResNet50, thereby showing how effectively a self-constructing graph can model the long-range global spatial correlation of images and get a competitive result. The SGA-Net outperformed ResNet50 and SGA-Net-ncl in two datasets, which shows that channel linear attention has ability to derive a correlation among channel outputs of a graph neural network, and further improve performance of the proposed model. The SGA-Net surpassed SGA-Net-one in predicting remote sensing images, showing that the rotation of images can keep geometric consistency, which improves image prediction performance.

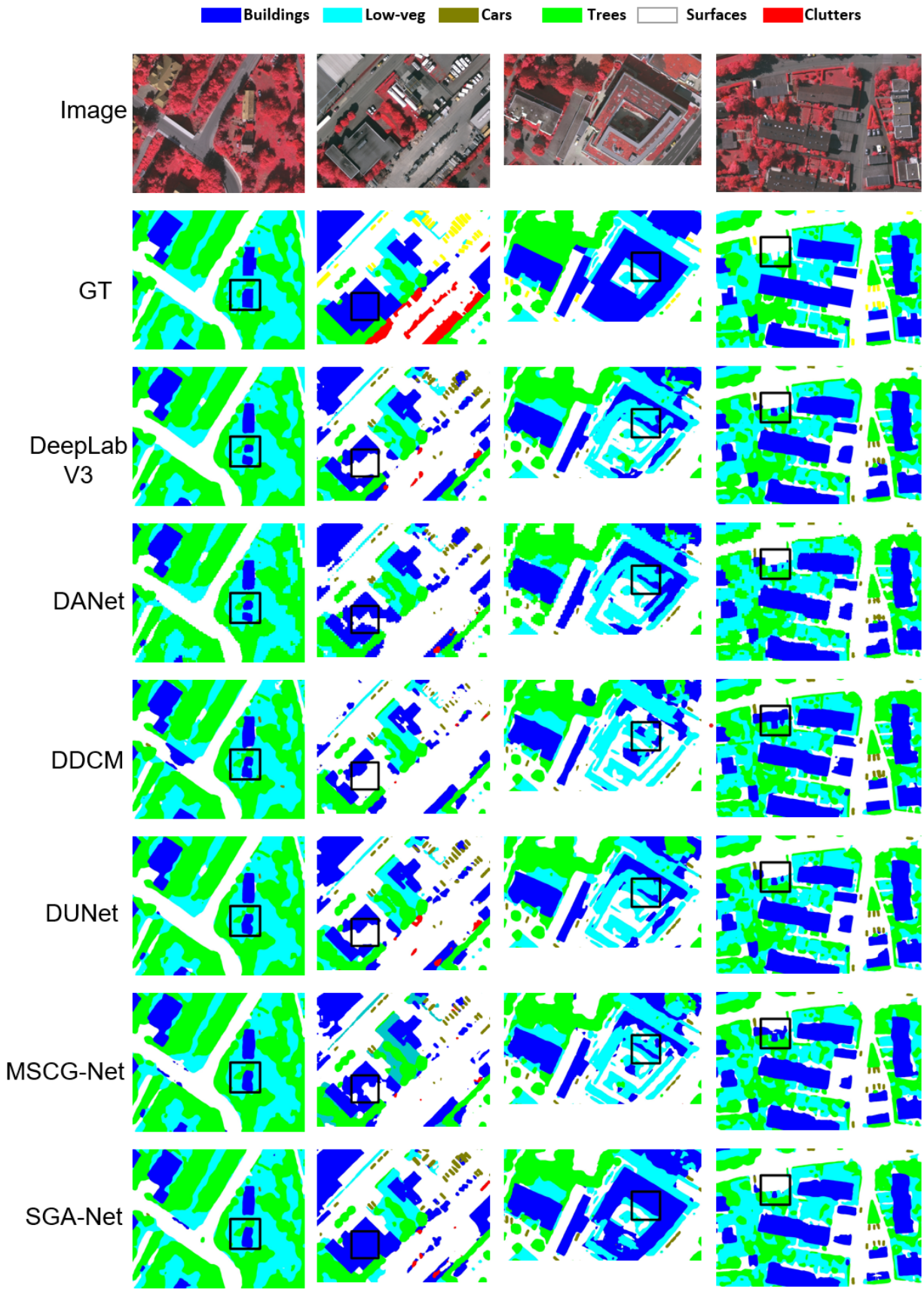

Figure 6. Visualization of prediction detail in the Vaihingen dataset.

Table 3. The ablation study about SGA-Net.

Dataset	Method	Mean F1	Acc	mIoU
Vaihingen	ResNet50	0.826	0.944	0.753
	SGA-Net-ncl	0.849	0.946	0.761
	SGA-Net-one	0.876	0.948	0.798
	SGA-Net	**0.905**	**0.965**	**0.826**
Potsdam	ResNet50	0.873	0.934	0.783
	SGA-Net-ncl	0.906	0.960	0.821
	SGA-Net-one	0.912	0.957	0.825
	SGA-Net	**0.927**	**0.964**	**0.832**

From Figures 7 and 8, we know that the performance of the SGA-Net-ncl surpassed ResNet50 and that the SGA-Net outperformed the baselines of the ablation study in two real-world datasets. Owing to long-range global spatial dependency extraction by a self-constructing graph attention network, the SGA-Net-ncl had a better prediction result than ResNet50. Moreover, channel linear attention acquired a correlation among the channel outputs of the graph neural network, which is why the SGA-Net was superior to the SGA-Net-ncl in semantic segmentation.

From Figure 9, we know the target object had a strong similarity with the same object. On the right of Figure 9, the target object is a building, and the color of the building region is red, meaning that the target pixel had a strong similarity with these pixeles of the building region. On the left of Figure 9, the target objects are low-vegetation and road, and the color of all cars is blue, indicating a low similarity. This picture shows that our attention mechanism works.

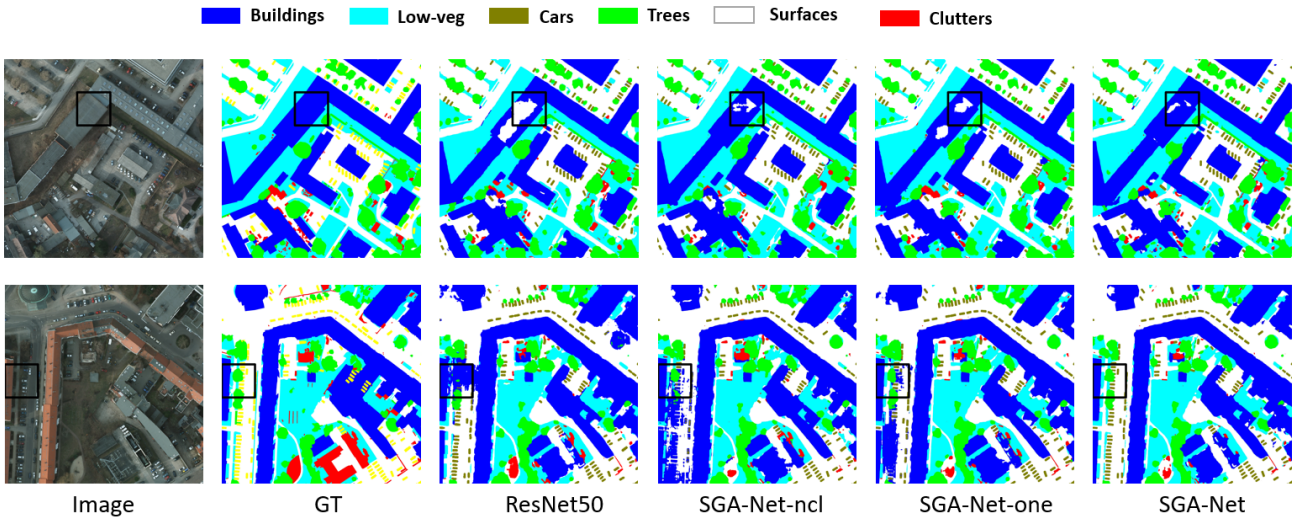

Figure 7. Visualization in the ablation study of Potsdam dataset.

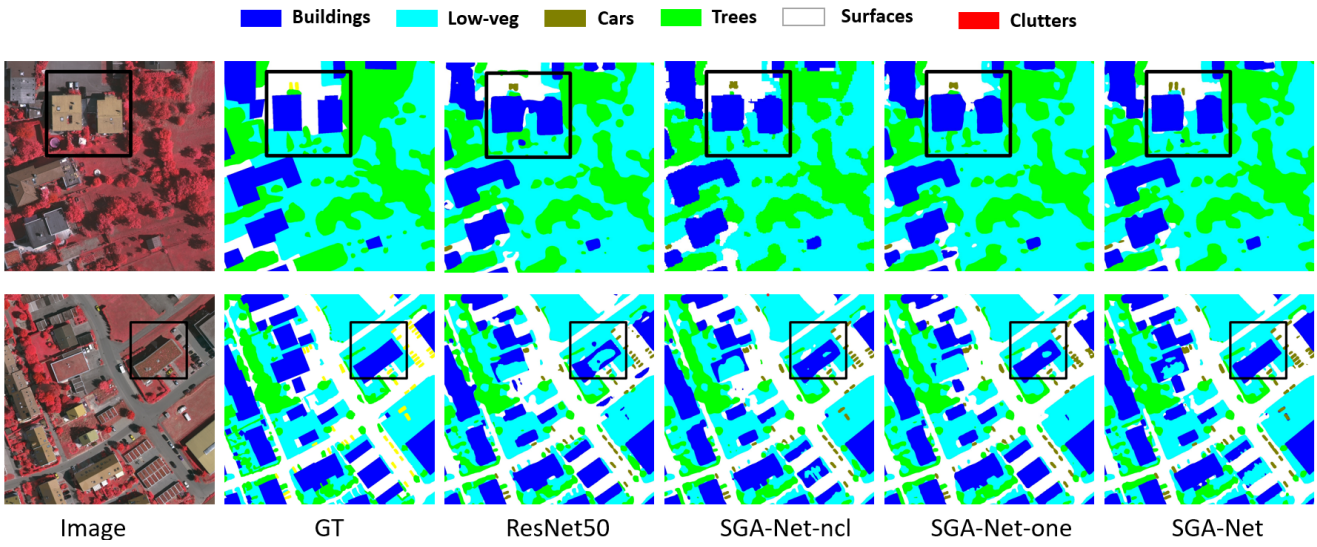

Figure 8. Visualization in the ablation study of Vaihingen dataset.

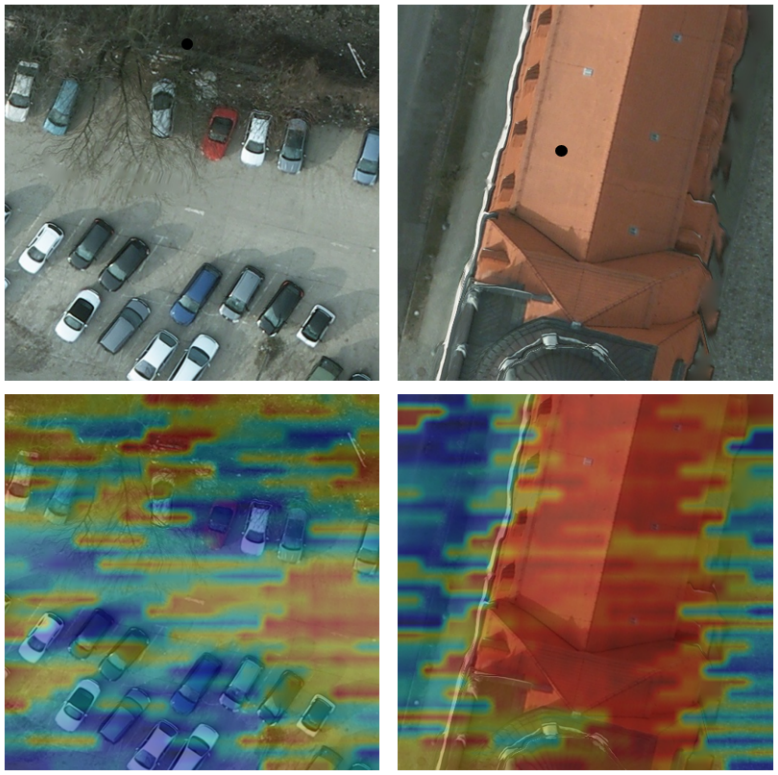

Figure 9. Visualization of the attention mechanism. The black dot is the target pixel or object. The red pixel color indicates that the target pixel is very similar to this pixel, and the blue color indicates that the target pixel is strongly different to this pixel.

5. Conclusions

In this paper, we proposed a novel model, SGA-Net, which includes a self-constructing graph attention network and a channel linear attention. The Self-constructing graph was obtained from feature maps of images rather than prior knowledge or elaborately designed manual static graphs. In this way, the global dependency of pixels can be extracted efficiently from high-level feature maps and present pixel-wise relationships of the remote sensing images. Then, a self-constructing graph attention network was proposed that aligned with the actual situation by using current and neighboring nodes. After that,

a channel linear attention mechanism was designed to obtain the channel dependency of images and further improve the prediction performance of semantic segmentation. Comprehensive experiments were conducted on the ISPRS Potsdam and Vaihingen datasets to prove the effectiveness of our whole framework. Ablation studies demonstrated the validity of the self-constructing graph attention network to extract the spatial dependency of remote sensing images and the usefulness of channel linear attention mechanisms for mining correlation among channels. The SGA-Net achieved competitive performance for semantic segmentation in the ISPRS Potsdam and Vaihingen datasets.

In future research, we will re-evaluate the high-level feature map and the attention mechanism to improve the segmentation accuracy. Furthermore, we would like to employ our model to train other remote sensing images.

Author Contributions: Conceptualization, W.Z. and W.X.; Methodology, W.Z. and H.C.; Software, W.Z.; Validation, H.C., W.X. and N.J.; Data Curation, N.J.; Writing—Original Draft Preparation, W.Z.; Writing—Review and Editing, W.Z. and J.L.; Supervision, W.X.; Project Administration, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: The work in this paper is supported by the National Natural Science Foundation of China (41871248, 41971362, U19A2058) and the Natural Science Foundation of Hunan Province No. 2020JJ3042.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ignatiev, V.; Trekin, A.; Lobachev, V.; Potapov, G.; Burnaev, E. Targeted change detection in remote sensing images. In Proceedings of the Eleventh International Conference on Machine Vision (ICMV 2018), Munich, Germany, 1–3 November 2018; Volume 11041, p. 110412H.
2. Liu, Y.; Chen, H.; Shen, C.; He, T.; Jin, L.; Wang, L. ABCNet: Real-Time Scene Text Spotting With Adaptive Bezier-Curve Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020.
3. Panero Martinez, R.; Schiopu, I.; Cornelis, B.; Munteanu, A. Real-time instance segmentation of traffic videos for embedded devices. *Sensors* **2021**, *21*, 275.
4. Balado, J.; Martínez-Sánchez, J.; Arias, P.; Novo, A. Road environment semantic segmentation with deep learning from MLS point cloud data. *Sensors* **2019**, *19*, 3466.
5. Behrendt, K. Boxy vehicle detection in large images. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
6. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226.
7. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105.
8. Liu, Q.; Kampffmeyer, M.; Jenssen, R.; Salberg, A.B. Self-constructing graph neural networks to model long-range pixel dependencies for semantic segmentation of remote sensing images. *Int. J. Remote Sens.* **2021**, *42*, 6187–6211.
9. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.* **2019**, *38*, 1–12.
10. Qi, X.; Liao, R.; Jia, J.; Fidler, S.; Urtasun, R. 3d graph neural networks for rgbd semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5199–5208.
11. Liang, X.; Hu, Z.; Zhang, H.; Lin, L.; Xing, E.P. Symbolic graph reasoning meets convolutions. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; pp. 1858–1868.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, 5998–6008.
13. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
14. Ben-Cohen, A.; Diamant, I.; Klang, E.; Amitai, M.; Greenspan, H. Fully convolutional network for liver segmentation and lesions detection. In *Deep Learning and Data Labeling for Medical Applications*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 77–85.

15. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
16. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495.
17. Liang-Chieh, C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
18. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848.
19. Hua, Y.; Marcos, D.; Mou, L.; Zhu, X.X.; Tuia, D. Semantic segmentation of remote sensing images with sparse annotations. *IEEE Geosci. Remote Sens. Lett.* **2021**, doi:10.1109/LGRS.2021.3051053.
20. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
21. Zhang, L.; Xu, D.; Arnab, A.; Torr, P.H. Dynamic graph message passing networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3726–3735.
22. Hamaguchi, R.; Furukawa, Y.; Onishi, M.; Sakurada, K. Heterogeneous Grid Convolution for Adaptive, Efficient, and Controllable Computation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13946–13955.
23. Yao, L.; Mao, C.; Luo, Y. Graph convolutional networks for text classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 7370–7377.
24. Wang, H.; Xu, T.; Liu, Q.; Lian, D.; Chen, E.; Du, D.; Wu, H.; Su, W. MCNE: An end-to-end framework for learning multiple conditional network representations of social network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1064–1072.
25. Liu, Y.; Wang, W.; Hu, Y.; Hao, J.; Chen, X.; Gao, Y. Multi-agent game abstraction via graph attention neural network. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 7211–7218.
26. Liu, Q.; Kampffmeyer, M.C.; Jenssen, R.; Salberg, A.B. Multi-view Self-Constructing Graph Convolutional Networks with Adaptive Class Weighting Loss for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 44–45.
27. Su, Y.; Zhang, R.; Erfani, S.; Xu, Z. Detecting Beneficial Feature Interactions for Recommender Systems. In Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI), virtually, 2–9 February 2021.
28. Liu, B.; Li, C.C.; Yan, K. DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief. Bioinform.* **2020**, *21*, 1733–1741.
29. Lampropoulos, G.; Keramopoulos, E.; Diamantaras, K. Enhancing the functionality of augmented reality using deep learning, semantic web and knowledge graphs: A review. *Vis. Inf.* **2020**, *4*, 32–42.
30. Zi, W.; Xiong, W.; Chen, H.; Chen, L. TAGCN: Station-level demand prediction for bike-sharing system via a temporal attention graph convolution network. *Inf. Sci.* **2021**, *561*, 274–285.
31. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57–81.
32. Xie, Y.; Zhang, Y.; Gong, M.; Tang, Z.; Han, C. Mgat: Multi-view graph attention networks. *Neural Netw.* **2020**, *132*, 180–189.
33. Gao, J.; Zhang, T.; Xu, C. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8303–8311.
34. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
35. Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; Hengel, A.v.d. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1960–1968.
36. Guo, S.; Lin, Y.; Feng, N.; Song, C.; Wan, H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 922–929.
37. Hamilton, W.L.; Ying, R.; Leskovec, J. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1025–1035.
38. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
39. Huang, Y.; Jia, W.; He, X.; Liu, L.; Li, Y.; Tao, D. CAA: Channelized Axial Attention for Semantic Segmentation. *arXiv* **2021**, arXiv:2101.07434.
40. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical multi-scale attention for semantic segmentation. *arXiv* **2020**, arXiv:2005.10821.

41. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An Image is Worth 16× 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
42. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
43. Tran, P.T.; others. On the convergence proof of amsgrad and a new version. *IEEE Access* **2019**, *7*, 61706–61716.
44. Kampffmeyer, M.; Jenssen, R.; Salberg, A.B.; others. Dense dilated convolutions merging network for semantic mapping of remote sensing images. In Proceedings of the 2019 Joint Urban Remote Sensing Event (JURSE), Vannes, France, 22–24 May 2019; pp. 1–4.
45. Xue, H.; Liu, C.; Wan, F.; Jiao, J.; Ji, X.; Ye, Q. Danet: Divergent activation for weakly supervised object localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6589–6598.
46. Tian, Z.; He, T.; Shen, C.; Yan, Y. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3126–3135.
47. Florian, L.C.C.G.P.; Adam, S.H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.