Class-guided Swin Transformer for Semantic Segmentation of Remote Sensing Imagery

Xiaoliang Meng, Yuechi Yang, Libo Wang*, Teng Wang, Rui Li and Ce Zhang

Abstract—Semantic segmentation of remote sensing images plays a crucial role in a wide variety of practical applications, including land cover mapping, environmental protection, and economic assessment. In the last decade, convolutional neural network (CNN) is the mainstream deep learning based method of semantic segmentation. Compared with conventional methods, CNN-based methods learn semantic features automatically, thereby achieving strong representation capability. However, the local receptive field of the convolution operation limits CNN-based methods from capturing global information. In contrast, Vision Transformer demonstrates its great potential in global information modelling and obtains superior results in semantic segmentation. Inspired by this, in this Letter, we propose a class-guided Swin Transformer (CG-Swin) for semantic segmentation of remote sensing images. Specifically, we adopt a Transformer-based encoder-decoder structure, which introduces the Swin Transformer backbone as the encoder and designs a class-guided Transformer block to construct the decoder. The experimental results on ISPRS Vaihingen and Potsdam datasets demonstrate the significant breakthrough of the proposed method over ten benchmarks, outperform both advanced CNN-based and recent Vision Transformers based approaches.

Index Terms—Fully Transformer network, class-guided mechanism, semantic segmentation, remote sensing.

I. INTRODUCTION

Semantic segmentation is one of the fundamental tasks in remote sensing imagery interpretation, which plays a vital role in widespread applications [4], such as land cover classification [7], urban planning [9], and ecological assessment [12]. Over the last decade, the convolutional neural network (CNN) has become the most popular framework for semantic segmentation due to its automatic and strong feature learning. The Fully Convolutional Network (FCN) [1] first proposes an end-to-end CNN structure for semantic segmentation. The outcome of FCN, although encouraging, appears to be coarse due to the over-simplified decoding process. To address such issue, the U-Net [16] presents an encoder-decoder framework with two symmetric paths, i.e. a contracting path for encoding semantic features and an expanding path for decoding and restoring the resolution. Benefiting from this encoder-decoder design, U-Net and its variants [18] obtain impressive segmentation performance. Subsequently, many CNN-based segmentation methods adopt the encoder-decoder structure and make great progress in the remote sensing community.

However, the convolutional kernel with a fixed size leads to a local receptive field, which makes CNNs incapable to model the long-range dependency or global context inherently [19]. As for semantic segmentation, the local pixel classification can be more accurate if with the support of global contextual information [20]. To circumvent this issue, some studies modified the convolutional operation [2, 3] or applied attention mechanisms [21]. The former aims to enlarge the receptive fields by using large kernel sizes, dilated convolutions, or feature pyramids, whereas the latter introduces spatial or point-wise attention modules into CNNs to better capture contextual information. Nevertheless, these methods fail to liberate the network from the dependence of a convolutional encoder, thus, biased toward local interactions.

More recently, several inspiring advances attempt to avoid convolution operations completely by a novel architecture, i.e. Vision Transformer (ViT) [22]. ViTs formulate the semantic segmentation task as a sequence-to-sequence problem and extract semantic features with long-range dependencies effectively, which obtain impressive performances on remotely sensed image classification [23-29]. Especially, the Swin Transformer [30] presents a hierarchical structure, which further enhances its potential for semantic segmentation and achieves state-of-the-art results.

Driven by this, in this Letter, we propose a class-guided Swin Transformer with a Transformer-based encoder-decoder architecture for remote sensing image segmentation. We introduce the Swin Transformer backbone as the encoder. Since category-based information is a vital factor for precise semantic
II. METHODOLOGY

The proposed CG-Swin follows the fully Transformer-based encoder-decoder architecture, which obtains pixel-level class annotations from non-overlapping image patches. The overview of the CG-Swin is shown in Fig. 1.

A. Encoder

The small version of the Swin Transformer (Swin-S) [30] is selected as the encoder due to its excellent trade-off between efficiency and accuracy. As shown in Fig. 1, the encoder takes an image $X \in \mathbb{R}^{3 \times H \times W}$ as the input and generates four semantic features at different channel dimensions and resolutions. Thus, we utilize standard $3 \times 3$ convolution layers and up-sampled operations to unify the shape of the four feature maps. Finally, we merge the four feature maps and generate an encoding feature $X_e \in \mathbb{R}^{d \times \frac{H \times W}{4}}$, where $d$ is the channel dimension and set as 96. This encoding feature is fed into the decoder for further processing.

B. Decoder

The class-guided Transformer decoder is composed of three stages. In the first stage, it applies the class-level information extraction module to capture the category-based semantic feature. The class-level information extraction module employs three $3 \times 3$ convolution layers to capture the class-level context automatically. Each convolution layer is equipped with a Batch Normalization operation and a ReLU activation function. The extracted class-level feature can be denoted by $X_e \in \mathbb{R}^{\frac{H \times W}{k}}$, where $k$ is the number of categories.

In the second stage, we first flatten the concatenated feature to a 1D vector $X_{Cat} \in \mathbb{R}^{\frac{H \times W}{d+k}}$ and linearly project it to a sequence of tokens. Then, the learnable relative position embedding $X_{Pos} \in \mathbb{R}^{\frac{H \times W}{(d+k)}}$ is embedded with the $X_{Cat}$ for improving positional information. A consecutive Swin Transformer block with two layers is selected to optimize the $X_{Cat}$. The consecutive Swin Transformer block can be defined as:

$$X^1 = WMSA(LN(X^0)) + X^0$$  \hspace{1cm} (1)

$$X^1 = MLP(LN(X^1)) + X^1$$  \hspace{1cm} (2)

$$X^2 = SWMSA(LN(X^1)) + X^1$$  \hspace{1cm} (3)

$$X^2 = MLP(LN(X^2)) + X^2$$  \hspace{1cm} (4)

where $X^0$ denotes the input vector $X_{Cat}$, $X^1$ and $X^2$ represents the output vector of the window-based multi-head self-attention module (WMSA) and the shifted window-based multi-head self-attention(SWMHSA) [30]. $X^1$ and $X^2$ denotes the middle vector and the output vector, respectively. $LN$ is the Layer Normal operation, and $MLP$ denotes the multilayer perceptron.

Finally, we separate the optimized $X_{Cat}$ into the $X_e$ and $X_c$ and apply the class-guided operation to enhance the context interaction between them. The class-guided operation consists of two parallel paths, i.e. a compression path and an attention path. The compression path employs the Linear Projection and Layer Norm operation to reduce the channel dimension of the $X_e$ to the number of categories. The attention path applies an average pooling layer and a Sigmoid activation function to generate a class-guided attention matrix. The total function of the class-guided operation can be defined as:

$$S(X_e, X_c) = LN(LP_e(X_e)) \odot \text{Sigmoid}(\text{AvgPool}(X_c))$$  \hspace{1cm} (5)

Where $LN$ and $LP$ denote the Layer Norm and Linear Projection, respectively. $\odot$ represents the dot product operation. Processed by the class-guided operation, the compressed semantic feature $X_e$ gains rich class-level information, which means that each channel dimension of $X_e$ accurately characterized a category of ground objects. For precise per-pixel classification, it is crucial to obtain such a category-based semantic feature in the decoding period.

In the third stage, we utilize a multilayer perceptron to refine the category-based semantic feature and restore its resolution to the original, thereby producing the segmentation mask.
TABLE I
THE EXPERIMENTAL RESULTS ON THE VAIHINGEN DATASET.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Imp. surf.</th>
<th>Building</th>
<th>Low veg.</th>
<th>Tree</th>
<th>Car</th>
<th>Mean F1</th>
<th>OA</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN [1]</td>
<td>ResNet50</td>
<td>89.73</td>
<td>93.17</td>
<td>80.57</td>
<td>88.89</td>
<td>71.55</td>
<td>84.78</td>
<td>87.99</td>
<td>73.45</td>
</tr>
<tr>
<td>PSPNet [2]</td>
<td>ResNet101</td>
<td>92.79</td>
<td>95.46</td>
<td>84.51</td>
<td>89.94</td>
<td>88.61</td>
<td>90.26</td>
<td>90.85</td>
<td>82.58</td>
</tr>
<tr>
<td>DeepLabV3+ [3]</td>
<td>ResNet101</td>
<td>92.38</td>
<td>95.17</td>
<td>84.29</td>
<td>89.52</td>
<td>86.47</td>
<td>89.57</td>
<td>90.56</td>
<td>81.47</td>
</tr>
<tr>
<td>DANet [5]</td>
<td>ResNet101</td>
<td>91.63</td>
<td>95.02</td>
<td>83.25</td>
<td>88.87</td>
<td>87.16</td>
<td>89.19</td>
<td>90.44</td>
<td>81.32</td>
</tr>
<tr>
<td>DDCM-Net [6]</td>
<td>ResNet50</td>
<td>92.70</td>
<td>95.30</td>
<td>83.30</td>
<td>89.40</td>
<td>88.30</td>
<td>89.80</td>
<td>90.40</td>
<td>82.90</td>
</tr>
<tr>
<td>CASIA2 [11]</td>
<td>ResNet101</td>
<td>93.20</td>
<td>96.00</td>
<td>84.70</td>
<td>89.90</td>
<td>86.70</td>
<td>90.10</td>
<td>91.10</td>
<td>-</td>
</tr>
<tr>
<td>V-FuseNet [8]</td>
<td>FuseNet</td>
<td>91.00</td>
<td>94.40</td>
<td>84.50</td>
<td>89.90</td>
<td>86.30</td>
<td>89.20</td>
<td>90.00</td>
<td>-</td>
</tr>
<tr>
<td>UFMG_4 [15]</td>
<td>-</td>
<td>91.10</td>
<td>94.50</td>
<td>82.90</td>
<td>88.00</td>
<td>81.30</td>
<td>87.70</td>
<td>89.40</td>
<td>-</td>
</tr>
<tr>
<td>MANet [13]</td>
<td>ResNet50</td>
<td>93.02</td>
<td>95.47</td>
<td>84.64</td>
<td>89.98</td>
<td>88.95</td>
<td>90.41</td>
<td>90.96</td>
<td>82.71</td>
</tr>
<tr>
<td>BANet [14]</td>
<td>ResT-Lite</td>
<td>92.23</td>
<td>95.23</td>
<td>83.75</td>
<td>89.92</td>
<td>86.76</td>
<td>89.58</td>
<td>90.48</td>
<td>81.35</td>
</tr>
<tr>
<td>CMFNet [17]</td>
<td>-</td>
<td>92.36</td>
<td>97.17</td>
<td>86.37</td>
<td>90.32</td>
<td>85.47</td>
<td>90.24</td>
<td>90.43</td>
<td>80.82</td>
</tr>
<tr>
<td>CG-Swin (ours)</td>
<td>Swin-S</td>
<td>93.55</td>
<td>96.24</td>
<td>85.70</td>
<td>90.59</td>
<td>87.98</td>
<td>90.81</td>
<td>91.68</td>
<td>83.39</td>
</tr>
</tbody>
</table>

TABLE II
THE EXPERIMENTAL RESULTS ON THE POTSDAM DATASET.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Imp. surf.</th>
<th>Building</th>
<th>Low veg.</th>
<th>Tree</th>
<th>Car</th>
<th>Mean F1</th>
<th>OA</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN [1]</td>
<td>ResNet50</td>
<td>90.84</td>
<td>95.59</td>
<td>84.10</td>
<td>84.75</td>
<td>84.95</td>
<td>88.05</td>
<td>88.02</td>
<td>79.53</td>
</tr>
<tr>
<td>DeepLabV3+ [3]</td>
<td>ResNet101</td>
<td>92.95</td>
<td>96.88</td>
<td>86.72</td>
<td>88.15</td>
<td>96.02</td>
<td>92.12</td>
<td>90.88</td>
<td>84.32</td>
</tr>
<tr>
<td>PSPNet [2]</td>
<td>ResNet101</td>
<td>93.36</td>
<td>96.97</td>
<td>87.75</td>
<td>88.50</td>
<td>95.42</td>
<td>92.40</td>
<td>91.08</td>
<td>85.88</td>
</tr>
<tr>
<td>DDCM-Net [6]</td>
<td>ResNet50</td>
<td>92.90</td>
<td>96.90</td>
<td>87.70</td>
<td>89.40</td>
<td>94.90</td>
<td>92.30</td>
<td>90.80</td>
<td>-</td>
</tr>
<tr>
<td>AMA_1</td>
<td>-</td>
<td>93.40</td>
<td>96.80</td>
<td>87.70</td>
<td>88.80</td>
<td>96.00</td>
<td>92.54</td>
<td>91.20</td>
<td>-</td>
</tr>
<tr>
<td>SW1_2</td>
<td>ResNet101</td>
<td>94.40</td>
<td>97.40</td>
<td>87.80</td>
<td>86.70</td>
<td>94.70</td>
<td>92.38</td>
<td>91.70</td>
<td>-</td>
</tr>
<tr>
<td>V-FuseNet [8]</td>
<td>FuseNet</td>
<td>92.70</td>
<td>96.30</td>
<td>87.30</td>
<td>88.50</td>
<td>95.40</td>
<td>92.04</td>
<td>90.60</td>
<td>-</td>
</tr>
<tr>
<td>DST5_10</td>
<td>ResNet50</td>
<td>92.50</td>
<td>96.40</td>
<td>86.70</td>
<td>88.00</td>
<td>94.70</td>
<td>91.66</td>
<td>90.30</td>
<td>-</td>
</tr>
<tr>
<td>MANet [13]</td>
<td>ResNet50</td>
<td>93.40</td>
<td>96.36</td>
<td>88.32</td>
<td>89.36</td>
<td>96.48</td>
<td>92.19</td>
<td>91.44</td>
<td>86.95</td>
</tr>
<tr>
<td>BANet [14]</td>
<td>ResT-Lite</td>
<td>93.40</td>
<td>96.37</td>
<td>87.31</td>
<td>88.01</td>
<td>95.79</td>
<td>92.12</td>
<td>90.76</td>
<td>85.62</td>
</tr>
<tr>
<td>CMFNet [17]</td>
<td>-</td>
<td>92.84</td>
<td>97.63</td>
<td>88.00</td>
<td>87.40</td>
<td>95.68</td>
<td>92.10</td>
<td>91.16</td>
<td>85.63</td>
</tr>
<tr>
<td>CG-Swin (ours)</td>
<td>Swin-S</td>
<td>94.07</td>
<td>97.42</td>
<td>88.53</td>
<td>89.74</td>
<td>96.61</td>
<td>93.28</td>
<td>91.93</td>
<td>87.61</td>
</tr>
</tbody>
</table>

III. EXPERIMENTAL RESULTS

A. Dataset

We select the ISPRS Vaihingen and Potsdam semantic labelling datasets to test the effectiveness of the proposed method. There are 33 tiles with an average size of 2494×2064 pixels in the Vaihingen dataset and 38 tiles with a size of 6000×6000 pixels in the Potsdam dataset. Following previous pieces of literature [6], in the Vaihingen dataset, we use the officially provided 16 images for training and 17 for testing, while the setting in the Potsdam dataset is 24 tiles for training and 14 tiles for testing. The image tiles are cropped into 1024×1024 px patches as the input. As for data augmentation, we use the random scale with a range from 0.5 to 1.5, random flip and random crop. Besides, the multi-scale strategy is applied in the testing period.

B. Experimental Setting

All of the experiments are implemented with PyTorch on a single RTX 3090, and the optimizer is set as AdamW with a learning rate of 6e-4. For each method, the overall accuracy (OA), mean Intersection over Union (mIoU), and F1-score (F1) are selected as evaluation metrics:

\[
OA = \frac{\sum_{k=1}^{N} TP_k + FP_k + FN_k}{\sum_{k=1}^{N} TP_k + FP_k + FN_k}
\]

\[
mlIoU = \frac{1}{N} \sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k + FN_k}
\]

\[
precision = \frac{1}{N} \sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k}
\]

\[
recall = \frac{1}{N} \sum_{k=1}^{N} \frac{TP_k}{TP_k + FN_k}
\]

\[
F1 = 2 \times \frac{precision \times recall}{precision + recall}
\]

where \( TP_k, FP_k, TN_k, \) and \( FN_k \) indicate the true positive, false positive, true negative, and false negatives, respectively, for the specific object indexed as class \( k \). OA is computed for all categories including the background.

C. Semantic Segmentation Results and Analysis

1) Quantitative Comparison: The experimental results on the Vaihingen and Potsdam datasets are listed in Table I and Table II. The quantitative metrics illustrate the effectiveness of the proposed method. Specifically, the CG-ViT achieves 90.81% in mean F1 score, 91.68% in OA, and 83.39% in mIoU for the Vaihingen dataset, with 93.28%, 91.93%, and 87.61% for the Potsdam dataset, outperforming the advanced CNN-based methods like DDCM-Net [6] and UFMG_4 [15]. Benefitting from the category-level semantic information modelled by the class-guided Transformer block, the performance of our method also outperforms recent Vision Transformers designed initially for remote sensing images, such as BANet [14] and CMFNet [17]. The visual comparisons with the FCN further illustrate the advantages of the proposed method (Fig. 2).

2) Ablation Study: To investigate the contribution of the class-guided mechanism to accuracy, we conduct the ablation study on the Vaihingen and Potsdam datasets. As shown in Table III, the utilization of the class guidance provides significant improvements, i.e. 2.21% and 1.65% mIoU for the Vaihingen and Potsdam test sets, respectively.
IV. CONCLUSION

In this Letter, we construct a fully Transformer network, namely the class-guided Swin Transformer (CG-Swin), for semantic segmentation of remote sensing images. We design a class-guided mechanism and combined it with the Swin Transformer block to construct a class-guided Transformer decoder. Numerical experiments conducted on the ISPRS Vaihingen and Potsdam datasets with benchmark comparison demonstrate the benefits of the class-guided mechanism and the effectiveness of the proposed method in segmentation accuracy.

REFERENCE


