Scientific Research Publishing

Open Journal of Applied Sciences, 2024, 14, 288-319
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

New Fusion Approach of Spatial and Channel Attention for Semantic Segmentation of Very High Spatial Resolution Remote Sensing Images

Armand Kodjo Atiampo[1], Gokou Hervé Fabrice Diédié[2]

[1]Unité de Recherche et d'Expertise du Numerique (UREN), Université Virtuelle de Céte d'Ivoire, Abidjan, Ivory Coast
[2]Laboratoire de Mathématiques et Informatique, Université Peleforo Gon Coulibaly, Korhogo, Ivory Coast
Email: armand.atiampo@uvci.edu.ci

How to cite this paper: Atiampo, A.K. and Diédié, G.H.F. (2024) New Fusion Approach of Spatial and Channel Attention for Semantic Segmentation of Very High Spatial Resolution Remote Sensing Images. *Open Journal of Applied Sciences*, **14**, 288-319.
https://doi.org/10.4236/ojapps.2024.142020

Received: January 20, 2024
Accepted: February 6, 2024
Published: February 9, 2024

Abstract

The semantic segmentation of very high spatial resolution remote sensing images is difficult due to the complexity of interpreting the interactions between the objects in the scene. Indeed, effective segmentation requires considering spatial local context and long-term dependencies. To address this problem, the proposed approach is inspired by the MAC-UNet network which is an extension of U-Net, densely connected combined with channel attention. The advantages of this solution are as follows: 4) The new model introduces a new attention called propagate attention to build an attention-based encoder. 2) The fusion of multi-scale information is achieved by a weighted linear combination of the attentions whose coefficients are learned during the training phase. 3) Introducing in the decoder, the Spatial-Channel-Global-Local block which is an attention layer that uniquely combines channel attention and spatial attention locally and globally. The performances of the model are evaluated on 2 datasets WHDLD and DLRSD and show results of mean intersection over union (mIoU) index in progress between 1.54% and 10.47% for DLRSD and between 1.04% and 4.37% for WHDLD compared with the most efficient algorithms with attention mechanisms like MAU-Net and transformers like TMNet.

Keywords

Spatial-Channel Attention, Super-Token Segmentation, Self-Attention, Vision Transformer

1. Introduction

Semantic segmentation of remote sensing images is crucial for analyzing land

cover and land use, especially for assessing anthropization effects in rural and urban environments and managing natural disasters [1]-[6]. Traditional methods, relying on grayscale or color analysis [7] [8] and texture or similarity features [9], fall short in precise pixel-level classification, particularly for high spatial resolution images (≤4 m). The challenge in such images lies in representing land cover classes at the scale of objects, with varying spectral distributions between rural and urban areas [2]. Rural areas predominantly feature large natural objects, while urban areas exhibit high variability in man-made objects. Analyzing these images over large areas necessitates detailed spectral analysis alongside considering the spatial and semantic context for better discrimination [4] [5].

CNNs are widely used for semantic segmentation, excelling in local information extraction [10]. However, for remote sensing, considering overall context and long-range dependencies is crucial to avoid ambiguity [4] [11] [12] [13]. Recently, Attention mechanisms have gained importance in computer vision, particularly for tasks like classification, detection, object localization, and segmentation [14] [15]. The optimal performance in classification and object detection is achieved by integrating classical CNNs with attention mechanisms [16]. Consequently, various attention mechanisms are combined in these architectures, typically operating at distinct spatial resolution levels [4] [17] [18] [19].

This paper introduces SCGLU-Net, a semantic segmentation model for remote sensing images in complex urban and rural environments. Inspired by MACU-Net [20], our hybrid architecture combines a CNN encoder with specific transformers as decoders. We use asymmetric convolution [20] to analyze local context and reduce computational complexity. The SCGLU-Net differs by introducing Propagated attention to enhance relevant descriptors from the encoder during multi-scale fusion. The SCGL block influenced by [21], introduces a unique attention layer that combines channel and spatial attention simultaneously, addressing local and global semantic context. The SCGL block shown in **Figure 1(b)**, considers interactions between spatial and channel descriptors. In this SCGL block, spatial attention aggregates features into a regular grid of super-tokens to enable the use of self-attention [15] at high resolutions when estimating spatial attention [22]. The model uses a fine refinement head (FRH) to merge spatial and channel information at the original image resolution. Performance testing is conducted on WHDLD and DLRSD datasets with complex scenes in various environments [20] [23]. The main contributions of this study are:

- Introduction of propagate attention, an attention mechanism to prioritize relevant information and reduce artifacts from an encoder layer during integration into the multi-scale fusion proposal in the decoder.
- Introduction of the SCGL block, incorporating channel and spatial attention in a single block. Unlike conventional methods, this block allows simultaneous interaction capture between spatial and channel descriptors, overcoming the limitation on self-attention use at higher spatial resolutions due to quadratic complexity.

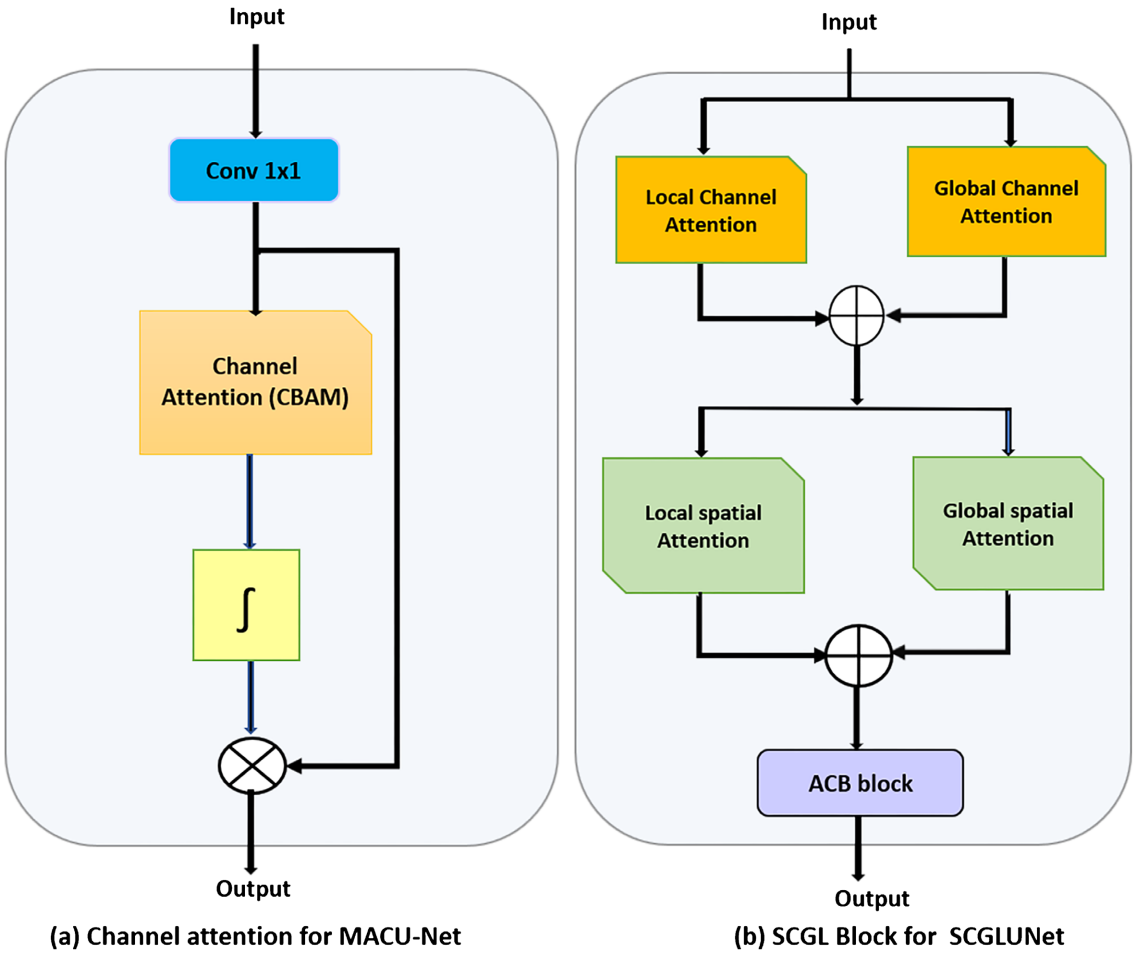

Figure 1. Illustration of attention block in (a) MACU-Net and (b) SCGLU-Net.

- To address the imbalance between target and nontarget areas, mitigating classifier bias towards the background class, a combination of Focal loss and Dice loss functions is employed to resolve sample imbalance.

The remainder of this paper is structured as follows. Section 2 is devoted to a synthetic review of previous works to show the interest of the new approach. Section 3 describes in detail the architecture of asymmetric convolution, propagate attention, SCGL, and the FRH blocks that constitute the core of the proposed model. In section 4, we present the results of our experiments and an analysis of the performances obtained compared with those of the most widely used methods in the literature. This paper ends with a conclusion followed by perspectives.

2. Related Work

Remote sensing image segmentation has progressed rapidly since the early 2000s, driven by the introduction of high and very high spatial resolution satellite imagers like IKONOS, QuickBird, and GeoEye. The fine spatial resolution of these images presented challenges for traditional pixel-based analysis [7] [9] [10] [24] [25] [26], leading to the development of new classification algorithms.

Those algorithms proved insufficient due to their inability to handle the internal variability of complex scenes [27] [28].

Inspired by works such as those in [29] [30], CNNs have become the standard for semantic segmentation in remote sensing due to their ability to extract spatial information. Two architectures have emerged, those based on pyramidal spatial pooling such as PSNet [31] and deepLab [17], and those based on the U-Net architecture [32]. U-Net employs is encoder-decoder that uses skip connections to concatenate information from the corresponding encoder layer and the layer below, allowing for multi-scale information capture and improving urban semantic segmentation [33] [34] [35]. Unlike U-Net family models, Models like PSPNet and Deeplab use spatial pyramid pooling to aggregate multi-scale information, from a fine-to-coarse level. Despite success on the PASCAL-VOC dataset [36], these models require pre-trained encoders and face limitations with very high-resolution images due to limited consideration of global spatial context. Another problem in the segmentation of fine remote sensing images is that they take care only of local spatial semantic context. To address the problem of global semantic context and improve performance in the semantic segmentation of remote sensing images, hybrid CNN has been proposed, in these architectures CNN models are combined with attention mechanisms, particularly in the decoder. Thus, several authors proposed to use various attention mechanisms like additive attention, self-attention, atrous convolution, spatial, and channel attention modules to enhance urban semantic segmentation [13] [37] [38] [39]. More recently, MACU-Net [20], featuring a densely connected CNN with CBAM-like channel attention [40], outperforms pure CNNs by increasing the mIoU score by over 1.5%, However, these attentions are built around the convolution product and therefore highly dependent on the local context.

Recently, transformers [15] have been adapted to computer vision, demonstrating excellence in classification tasks [16] and long-term dependency modeling [41] [42]. Two architectural trends have emerged for the semantic segmentation of very high spatial resolution images. Pure transformers, serving as both encoder and decoder in [43] [44], suffer from increasing computational complexity. The second trend involves a Transformer-based encoder and CNN-based decoder [45] [46]. Despite addressing local spatial and global semantic contexts, these models face increased complexity due to the quadratic computational complexity of transformers in the encoder. In [12] the authors show that optimal performance in object classification and detection tasks was achieved by combining classical CNNs with transformers. An alternative approach employs a CNN-based encoder and transformer-based decoder [21] [38] [47], featuring multi-scale feature fusion and a blend of attention mechanisms at different spatial resolution scales [4] [18] [21] [47]. Transformers and attention mechanisms are used separately in various processes and at the deepest spatial resolution levels [4] [18] [19]. However, in [21], the authors highlight the significant performance boost achieved by considering interactions between spatial

and channel features, which is overlooked in certain architectures. We propose a model that combines the benefits of pure transformers and hybrid architectures, featuring a CNN-based encoder and a transformer-based decoder within the MACU-Net framework. The model introduces a new attention mechanism to capture the energy of spatial features in constructing feature maps for each network layer. Additionally, the decoder integrates a mechanism to combine channel and spatial attention interaction at different spatial resolution levels, enhancing the model's capacity to consider both local and global semantic contexts in scenes.

3. Methods

In this section, we provide an in-depth analysis of the key components of the architecture. We begin by highlighting the architectural differences from the MACU-Net model. The focus then shifts to a detailed examination of attention mechanisms, particularly those employed in the decoder. The section is organized into sub-sections, covering a general presentation of the architecture (3.1), a review of Propagate Attention (3.2), an exploration of the Spatial-Channel-Global-Local block (SCGL) (3.3), a study of a version of the FRH block (3.4), and concludes with an estimation of the loss function in section 3.5.

3.1. Structure of SCGLU-Net

The new model is inspired by the MACU-Net architecture [20] presented in Figure 2, a densely connected Convolutional Neural Network (CNN) with an encoder-decoder structure. In Figure 3 we decribe the architecutre of our new model. Like MACU-Net, the new model encoder employs asymmetric convolution

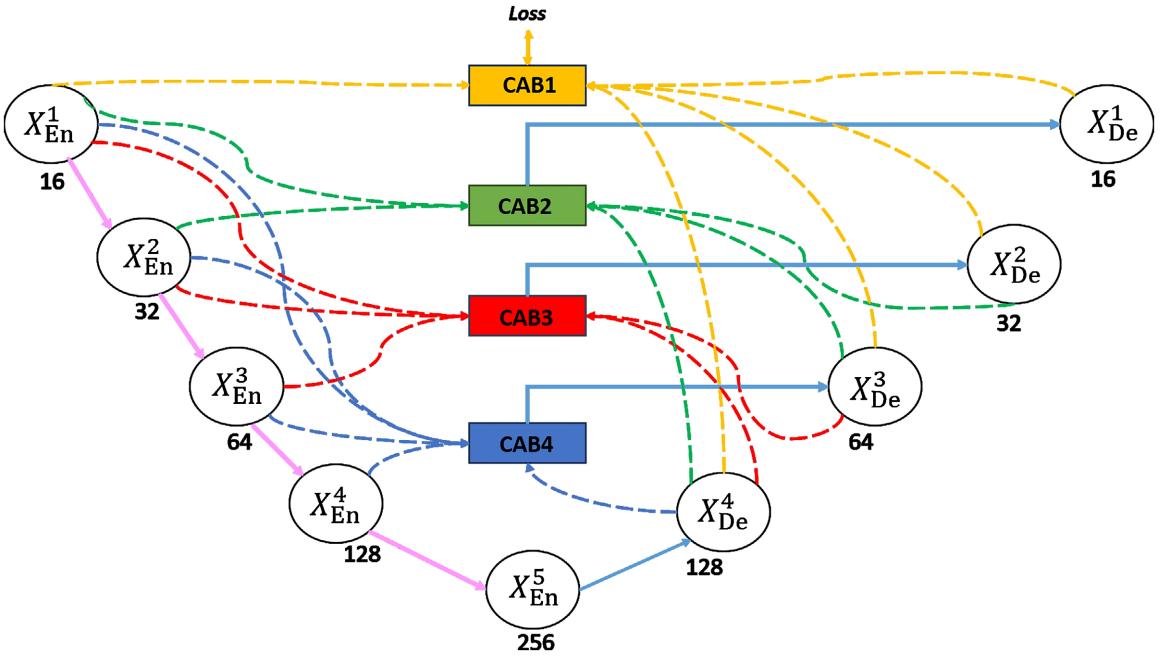

Figure 2. MACU-Net architecture.

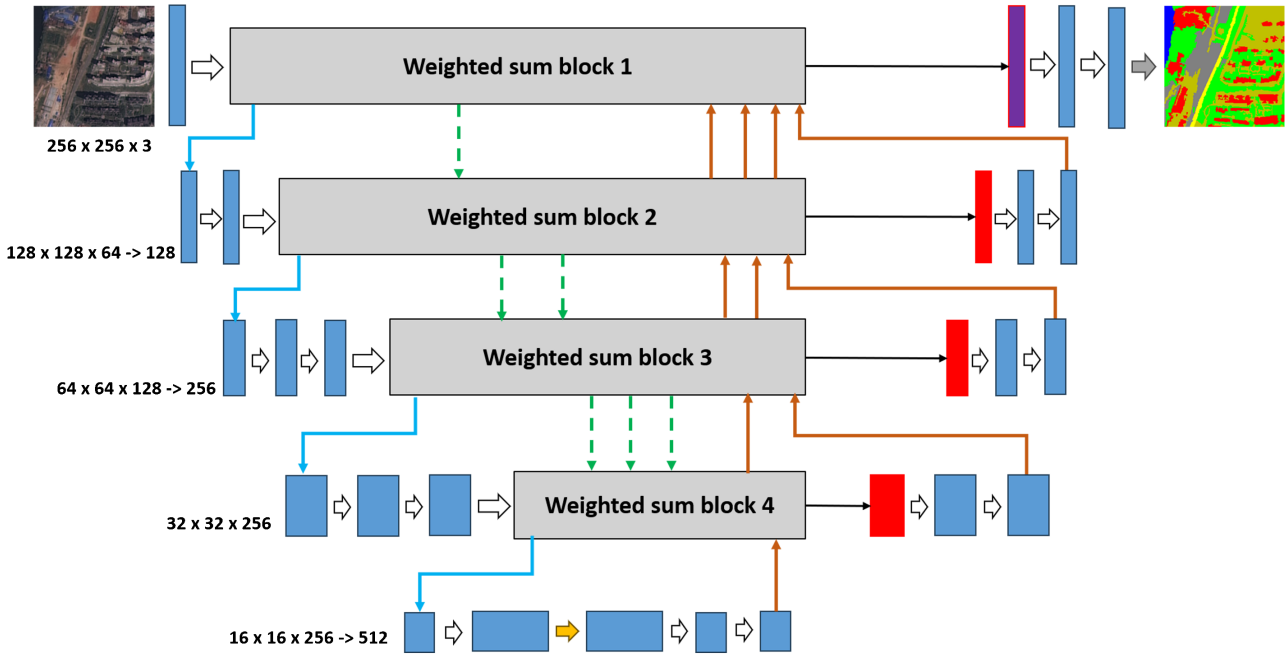

Figure 3. SCGLU-Net architecture.

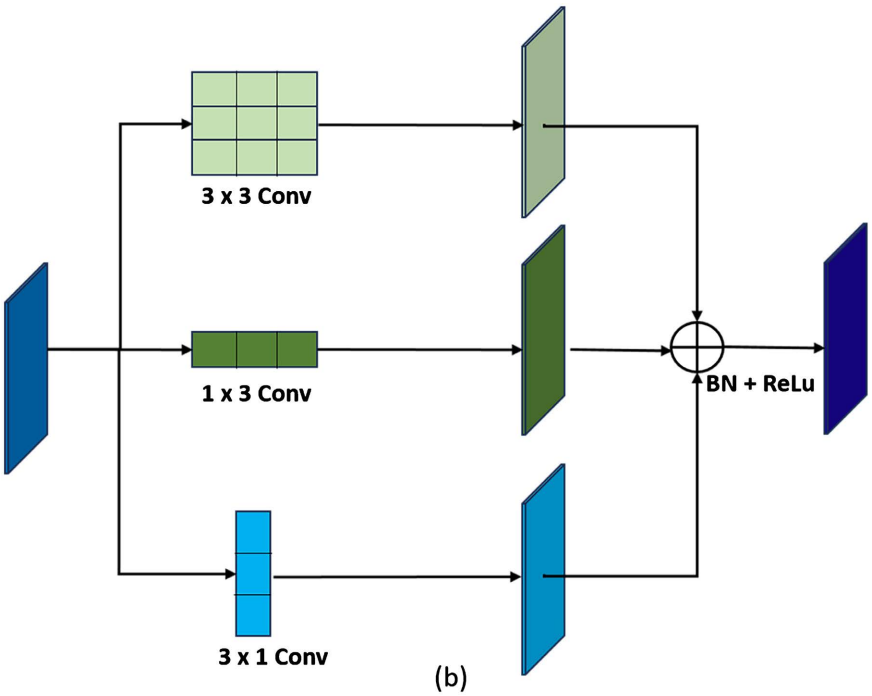

Figure 4. Asymmetric convolution block.

blocks (ACB) [20] to enhance representation power and capture local context with lower computational complexity [39]. It allows the encoder to extract descriptor maps at various spatial resolutions from coarse to fine, increasing channel dimensions. The principle of ACB block is illustrated in Figure 4. In the new architecture, the transition between encoder layers involves k ACB blocks, followed by size reduction using max-pooling with a factor of 2. The value of k is 2

for transitions from layer 1 to 2 and for layer 2 to 3, 3 for transitions from layer 3 to 4 and for layer 4 to 5.

The main difference between our model and MACU-Net is the decoder architecture. In MACU-Net, the decoder utilizes deconvolution and channel attention processes to reconstruct the original image's segmentation mask whereas in SCGLU-Net a combination of different attention mechanisms is used to reconstruct the segmentation mask. To capture global interactions, the transition from the deepest encoder layer to the decoder involves Multi-Head Self-Attention (MSA) [15] followed by 2 ACB blocks. Inspired by previous work [18] showing the performance benefits of combining multiple attention mechanisms, the decoder utilizes the new Spatial-Channel-Global-Local block (SCGL), which combines spatial and channel attention simultaneously at local and global scales. This block allows interactions between spatial and channel descriptors to be taken into account. Two ACB blocks follow each SCGL block before transposed convolution. Local attention uses 3 × 3 or 5 × 5 kernel convolutions, while global context is built around self-attention mechanisms, with pixels clustered into a regular grid of super-pixels at each spatial resolution level [22]. The multi-scale information fusion process shown in Figure 5 introduces a novel attention mechanism called propagate attention, which enables the encoder to extract feature

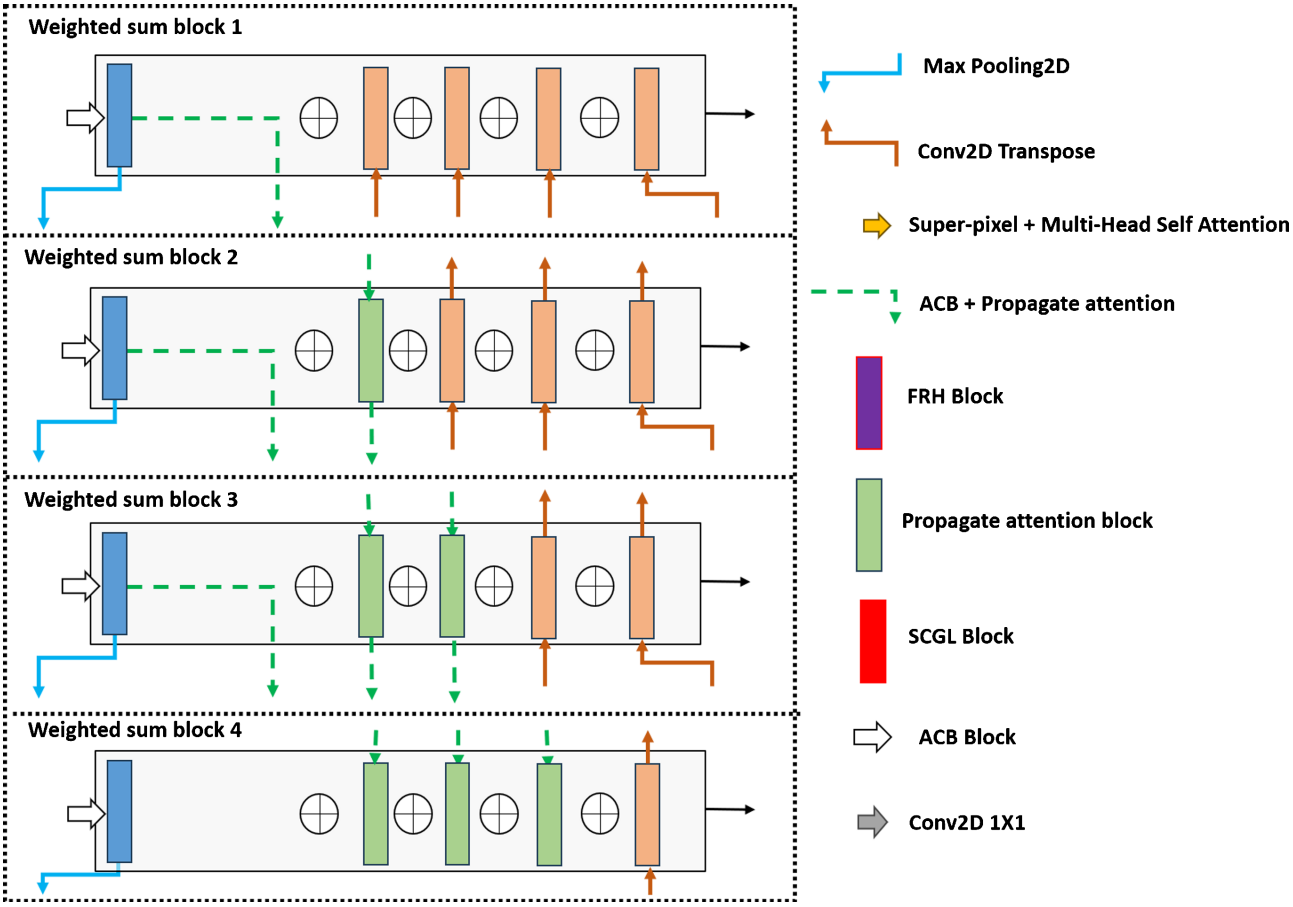

Figure 5. Fusion information block in SCGLU-Net architecture.

maps with both local and global context information. Information fusion from the encoder and lower decoder layers occurs at different spatial scales, using weighted summation according to Equation (1).

$$FF_l = \sum_{i=1}^{l} \alpha_i EF_{il} + \sum_{j=l+1}^{5} \beta_j DF_{jl} \tag{1}$$

In this equation FF_l represents the fused features at the input of layer l, EF_{il} represents the feature tensor coming from encoder layer i weighted by the propagate attention, DF_{jl} represents the features coming from decoder layer j, α_i and β_j are real numbers such that their sum gives 1

The approach ensures that information coming from each feature map from encoder layers is weighted based on its importance. The weights are learned during the training phase. The final layer includes a feature refinement head (FRH) block to combine spatial and channel information at the original image resolution, capturing semantic context from lower layers. The next sections outline the key blocks that form the core of the new model.

3.2. Propagate Attention

This attention is used in our model to fuse information coming from the encoder with those coming from the decoder. It aims to favor the most relevant spatial features at each spatial scale after downsampling since in our model, input data in each decoder layer is a combination of features coming from the below decoder layers and all above layers from the encoder. Indeed, as indicated by the authors in [18], the spatial features at a higher scale of spatial resolution have a greater impact during the process of merging information. Although uses only convolution products followed by pooling to propagate features from coarse-to-fine spatial levels suffers from the unique grasp of the spatial context and only guarantees the translational invariance of the network. This attention fills this gap by taking into account each spatial scale, the global context, and the local context with a similar computational complexity. This attention is inspired by that proposed by the authors in [18] to improve the residual blocks' capacities. Let X_{il} be the features map at layer l comes from encoder layer i by ACB block. It is a 4D tensor $\in \mathbb{R}^{B\times D\times H\times W}$, where B is the number of samples in the batch, H, W the spatial dimensions of layer l, D the number of channels or the depth of the features map. At the layer l, Att_l is calculated according Equation (2), Equation (3), and Equation (4):

$$\bar{X}_{il} = globAvgPool\left(X_{il}\right) \tag{2}$$

$$\hat{X}_{il} = softmax\left(\frac{Q\left(\bar{X}_{il}\right)K\left(\bar{X}_{il}\right)^{\mathrm{T}}}{\sqrt{D}}\right)^{\mathrm{T}} \cdot V\left(X_{il}\right) \tag{3}$$

$$Att_l = sigmoid\left(\frac{\tilde{X}_{il,\mu}}{4\left(v+\varepsilon\right)} + 0.5\right) \odot X_{il} \tag{4}$$

Tensor $\hat{X}_{il}$ was built from X_{il} by freezing spatial dimensions H and W with

global average pooling product combining with self-attention mechanism as described in Equation (2) and Equation (3).

In Equation (3), we use a 1D convolution with kernel size of 5 to infer Query (Q), and key (K) as tensors $\in \mathbb{R}^{B\times D}$. V is the tensor X_{il} reshaped as $\mathbb{R}^{B\times D\times (HW)}$ tensor. The tensor $\hat{X}_{il}$ is subsequently reshaped as a $\mathbb{R}^{B\times D\times H\times W}$ tensor.

In Equation (4), $\odot$ denotes element-wise matrix multiplication, v is the unbiasis variance of $\hat{X}_{il}$ tensor. This variance is calculated along spatial dimensions H and W. $\tilde{X}_{il,\mu}$ is the version of $\hat{X}_{il}$ tensor centered around the mean.

Figure 6 illustrates the flowchart of propagate attention.

3.3. Spatial-Channel-Gloabl-Local Block (SCGL)

The Spatial-Channel Global-Local (SCGL) block comprises a channel attention block followed by a spatial attention block, enabling consideration of spatial scale and channel dimension changes in input feature maps for multi-scale information fusion. Inspired by SACM in [48], the channel attention introduces a new branch to estimate interactions between channels, improving upon SACM [48] by avoiding neglect of interactions between spatial and channel features. The spatial attention is bifurcated into two branches: one capturing local spatial interactions and preserving details, and the other capturing long-term dependencies and global semantic context for scene interpretation. Figure 7 illustrates the new channel attention flowchart, while Figure 8 depicts the flowchart of spatial attention.

3.3.1. Channel Attention in SCGL

In the SCGL block, channel attention is the summation of local channel attention and global channel attention according to Equation (5) and Equation (6):

$$Y = conv2D_{1\times 1}(X) \tag{5}$$

$$Att_{ch}(X) = \breve{Y} + \hat{Y} \tag{6}$$

$\breve{Y}$ is the result of local transformation branch and is given by the following equation Equation (7):

$$\hat{Y} = Y \odot \left(F_{DH}(Y) \cdot F_{DW}(Y)\right) \tag{7}$$

$\hat{Y}$ is the result of global channel attention and is formulated with Equation

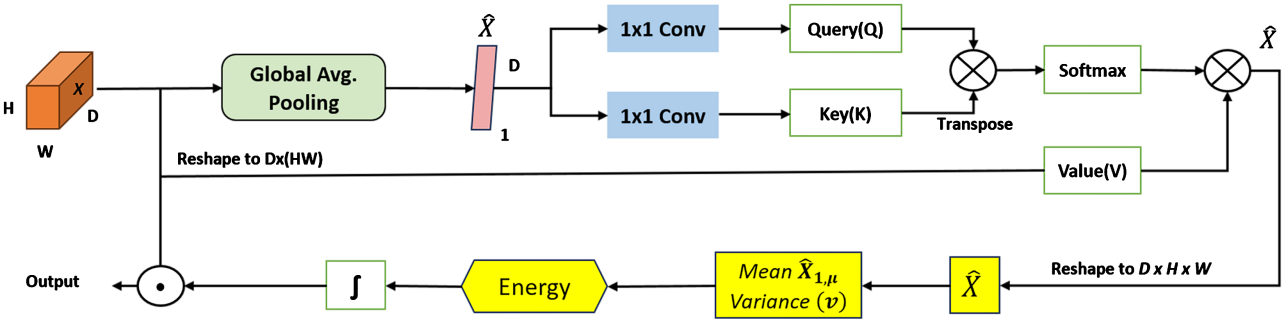

Figure 6. Propagate attention flowchart.

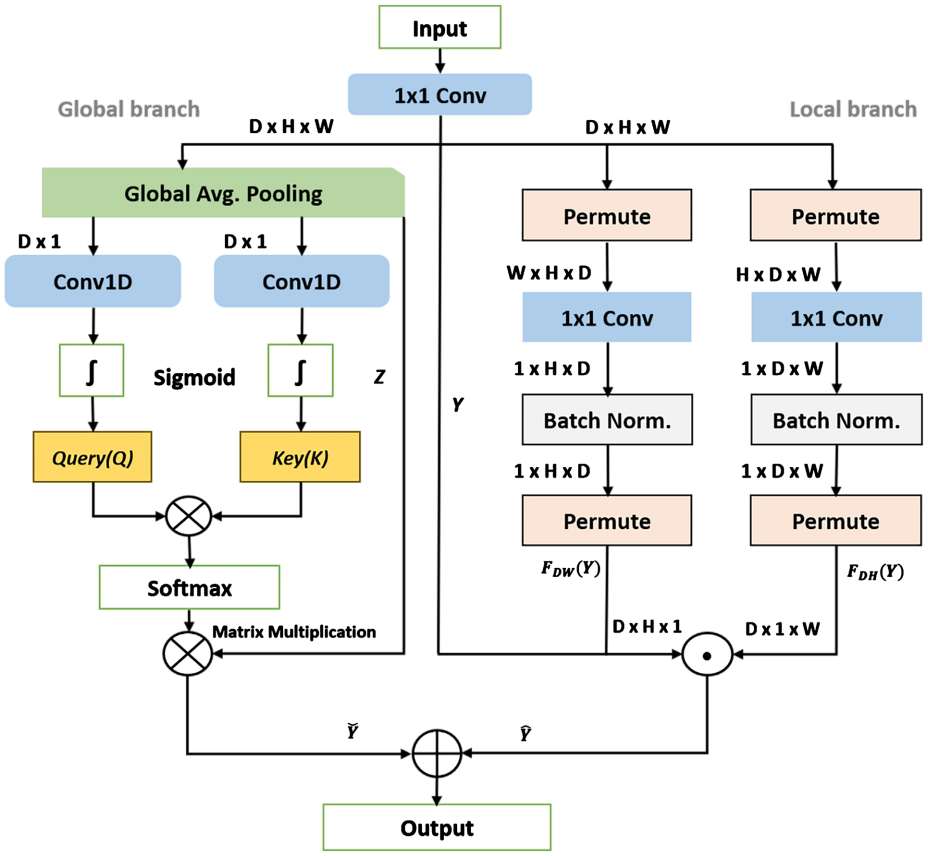

Figure 7. Channel attention in SCGL Block.

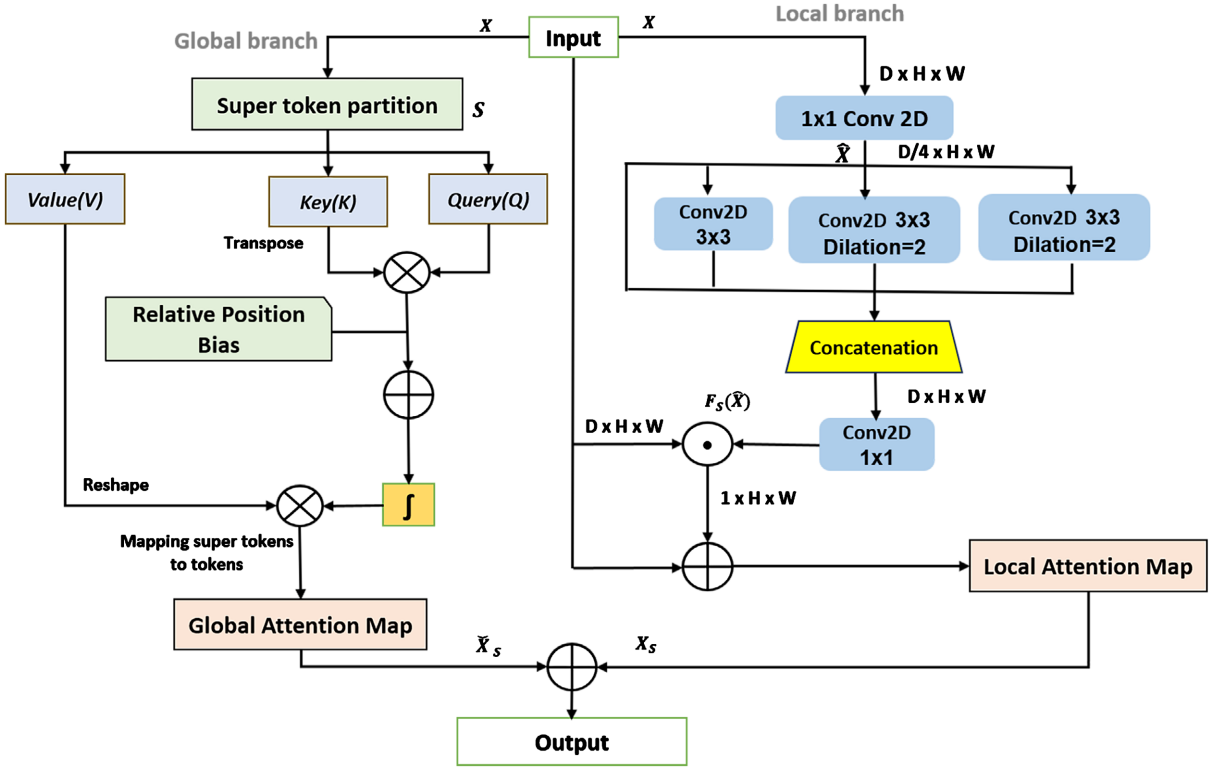

Figure 8. Spatial attention in SCGL.

(8) and Equation (9)

$$Z = glob_{Avg}\left(Y\right) \tag{8}$$

$$\breve{Y} = softmax\left(\frac{Q\left(Z\right)K\left(Z\right)^{\mathrm{T}}}{\sqrt{D_{out}}}\right)^{\mathrm{T}} \cdot V\left(Z\right) \tag{9}$$

In Equation (7), $\odot$ denotes element-wise matrix multiplication, $\cdot$ denotes matrix multiplication and $D = D_{out}$, $F_{DP}\left(X\right)$ with $P = W$ or H is formulated according Equation (10)

$$F_{DP}\left(Y\right) = Permute_{D,P}\left(conv_{1\times1}\left(Permute_{D,P}\left(Y\right)\right)\right) \tag{10}$$

In this equation $Permute_{D,P}\left(.\right)$ is transformations oparator for X tensor witihn P is permuted with channel dimension D. $conv_{1\times1}$ is a 1 × 1 2D-convolution product with one filter followed by a batch normalization to obtain a vector of dimension $D = 1$.

In Equation (8) $glob_{Avg}\left(Y\right)$ represents global average pooling to freeze spatial dimensions before using self-attention to estimate long-range dependencies in Equation (9). In this Equation (9) Query (Q) and Key (K) tensors are estimated with the Equation (11) and Equation (12).

$$Q = \sigma\left(conv1_5\left(Z\right)\right) \tag{11}$$

$$K = \sigma\left(conv1_5\left(Z\right)\right) \tag{12}$$

within $conv1_5$ a 1D-convolution with a 5 kernel and σ a sigmoid function to maintain the values obtained in the interval [0; 1]. In Equation (11) and Equation (12), $conv1_5$ a 1D-convolution with a 5 kernel and σ a sigmoid function to maintain the values obtained in the interval [0; 1]. V, is the tensor Z with dimension $D_{out} \times H \times W$ rescaled to a tensor with dimension $D_{out} \times \left(HW\right)$ and $\cdot$ represents the classical matrix product.

3.3.2. Spatial Attention in SCGL

Spatial attention in SCGL was built around 2 branches that work in parallel. One branch is responsible for building spatial attention with local context and is based on the convolution product. The other branch is responsible for capturing global context and is based on a multi-head self-attention mechanism. Spatial attention is formulated according Equation (13):

$$Att_{spat}\left(X\right) = \dot{X}_S + \breve{X}_S \tag{13}$$

In this equation, X refers to the output tensor of the channel attention in SCGL. The principle of spatial attention is illustrated in **Figure 8**.

Local attention is inspired by the work of [49] who showed that in the case of semantic segmentation, atrous convolution is better suited for dense prediction without loss of resolution than the classical convolution product rather suitable for classification tasks. The local spatial attention was estimated according equations Equation (14) and Equation (15):

$$\hat{X} = conv2_{1\times1}(X) \tag{14}$$

$$\dot{X}_S = \hat{X} + F_s(\hat{X}) \odot \hat{X} \tag{15}$$

with

$$F_s(\hat{X}) = conv2_{1\times1}\left(concat\left(MultiConv(\hat{X}), \hat{X}\right)\right) \tag{16}$$

$$MultiConv(\hat{X}) = \left(F_{3\times3}^{dil=1}(\hat{X}), F_{3\times3}^{dil=3}(\hat{X}), F_{3\times3}^{dil=5}(\hat{X})\right) \tag{17}$$

In Equation (16) a 1 × 1 2D-convolution product is used to obtain the weight tensor due to each spatial feature.

In Equation (17) $F_{3\times3}^{dil=\cdot}(.)$ relies on 3 2D-convolution products with 3 × 3 kernel size and $D/4$ filters but with dilation rate of 1, 3 and 5. Moreover, to reduce the computational complexity and as proposed in [50] the 3 × 3 convolution of dilation 1 is seen as a combination of the asymmetric 2D-convolutions with 3 × 1 kernel size and 1 × 3 kernel size.

Global spatial attention was estimated according Equation (18):

$$\breve{X}_S = Atttn(S) \times Q \tag{18}$$

In this equation, S means super-tokens features tensor, and $Q \in \mathbb{R}^{m\times n}$ is the mapping matrix of the features tensor $X \in \mathbb{R}^{D\times H\times W}$ into super-tokens features tensor $S \in \mathbb{R}^{D\times m}$. W, H are the spatial dimensions of X, D is the channel dimensions, $n = HW$, and m is the number of super-tokens. $Atttn(.)$ represents multi-head self-attention mechanism. This attention is introduced in [15] because as the authors indicated it is the best approach to capture long-term dependencies and take into account the global context [4] [18]. To limit the effects of computational complexity, we will base our multi-head self-attention mechanism by estimation on an Atttn (adaptation of super-pixel clustering as proposed by [51] [22]. $Atttn(S)$ is estimated according to Equation (19):

$$Attn(S) = softmax\left(\frac{q(S)k(S)^{\mathrm{T}} + q(S)k(R)^{\mathrm{T}}}{\sqrt{D}}\right)^{\mathrm{T}} \cdot v(S) = A(S)v(S) \tag{19}$$

with A (S), the attention map with the relative position embeddings [52]. Vectors $q(S) = SW_q$, $K(S) = SW_k$, $v(S) = SW_v$, $k(R) = RW_k'$ are the results of the linear transformations by the weight matrix W_q, W_k, W_v, W_k' obtained by a 1 × 1 2D-convolution. The vector R represents the embedded relative position vector introduced to improve self-attention performance and guarantee its permutation equivariant. Its formulation is identical to that indicated in [53]. In the Equation (19), to reduce the complexity, for each token, only its 3 × 3 surrounding super-tokens are used to compute $q(S)k(R)^{\mathrm{T}}$. In practice, we use the Unfold and Fold Python functions to extract and combine the corresponding 3 × 3 super-tokens, respectively. The relative position embeddings tensor R is constructed for each element $(i, j) \in \{1, 2, 3, \cdots, H\} \times \{1, 2, \cdots, W\}$ by determining the relative distance of (i, j) to each position $(a, b) \in \mathbb{V}_3(i, j)$, where $\mathbb{V}_3(i, j)$

is 3 × 3 neighbor around position (i, j). Each element $(a,b) \in \mathbb{V}_3(i, j)$ receives two distances: a row offset $a-i$ and column offset $b-j$ as it is shown in Figure 9.

The row and column offsets are associated with an embedding r_{a-i} and r_{b-j} respectively each with dimension $\frac{1}{2}D$. The row and column offset embeddings are concatenated to form $r_{a-i,b-j}$.

The tensor S and the matrix Q are constructed iteratively:

1) linear normalization of input tensor X with a 1 × 1 2D-convolution with D filters according Equation (20)

$$\hat{X} = conv2_{1\times1}(X) \tag{20}$$

2) Creation of initial super-tokens tensor $S^0 \in \mathbb{R}^{D\times m}$ by calculating the local average of the tokens on a regular sliding window with size $h \times w$ such that $m = \frac{H}{h} \times \frac{W}{w}$. Tokens tensor is the features tensor $\hat{X} \in \mathbb{R}^{D\times H\times W}$ resized to a tensor $\in \mathbb{R}^{D\times HW}$.

3) At each iteration t, matrix Q was estimated according to Equation (21) and Equation (22)

$$Q^t = softmax\left(\frac{\left(S^{t-1}\right)^{\mathrm{T}} \hat{X}}{\sqrt{D}}\right) \tag{21}$$

where D is the number of Channels. The set S^t of super-tokens is updated by the weighted sum of tokens

$$S^t = \hat{X}\left(Q^t\right)^{\mathrm{T}} \tag{22}$$

Global spatial attention is obtained by resizing $\breve{X}_S$ into a tensor with dimension $D_{out} \times H \times W$. For clarity, we have presented the results for a single attention head. In practice, multiple heads of attention are used by partitioning the features map depthwise into N groups to learn multiple distinct representations of the input tensor. The final result comes from the concatenation of the results obtained for each attention head. For our model, we have retained after our experiments the following distribution presented in Table 1 below which shows for each decoder layer, the number of attention heads, the depth of the input

(-1,-1)	(-1,0)	(-1,1)
(0, -1)	(0, 0)	(0,1)
(1, -1)	(1, 0)	(1,1)

Figure 9. The principle of relative distances. The row offset is in blue color, the column offset is in red color.

Table 1. Parameters for building spatial global attention.

Layer	Channels (H, W)	Heads	Features size (H, W)	Super-tokens size (h, w)	iterations
05	0512	08	00 (16, 16)	00 (1, 1)	3
04	0256	04	00 (32, 32)	00 (2, 2)	3
03	0256	04	00 (64, 64)	00 (4, 4)	3
02	0128	02	00 (128, 128)	0 (16, 16)	3

tensor, the spatial dimension of the tensor, as well as the spatial dimension of the super-tokens and number of iterations to estimate them.

The table above summarizes the parameters for constructing the global spatial attention of the SCGL block from the deepest layer (layer 5) to the uppermost layer (layer 2). The output layer, based on the fine refinement head in the next section, is not listed. As one progresses from lower to higher layers, the number of heads decreases due to a halving of channel count, while super-token size doubles and is limited to 16 × 16 [16] for global relation extraction. Three iterations were used to estimate the super-token count, as increasing it did not significantly improve results in our experiments.

3.4. Fine Refinement Head Block (FRH)

This block, inspired by [4], merges rich semantic data from lower network layers with spatial descriptors from the original image. Comprising two branches, one focuses on channel interactions, using the Convolutional Block Attention Module (CBAM) [40] for channel attention. The channel attention map is generated through a weight-shared network. The other branch addresses spatial interactions through depth-wise convolution. The attentions are combined by summation, processed by two asymmetric convolution blocks (ACB), and a 1 × 1 2D-convolution produces the segmentation mask. Unlike the original module, this approach avoids over-sampling and linear interpolation, reducing errors. **Figure 10** illustrates a visual representation of the fine refinement head block (FRH).

3.5. Loss Function

To address the challenge of gradient vanishing in deep networks, particularly in semantic segmentation of remote sensing images with unbalanced classes, a robust loss function is crucial for optimal convergence during training. To mitigate the impact of class imbalance, the focal loss introduced by Lin in 2017 [54] is employed, defined by Equation (23)

$$\mathcal{L}_{Focal} = -\alpha\left(1-p_t\right)^{\gamma}\log\left(p_t\right) \tag{23}$$

with $\gamma = 2$ and $\alpha = 0.25$ and p_t is the probability of the pixel belonging to the object class estimated from a softmax function at the network output. This modified cross-entropy penalizes over-represented classes, reducing their impact on

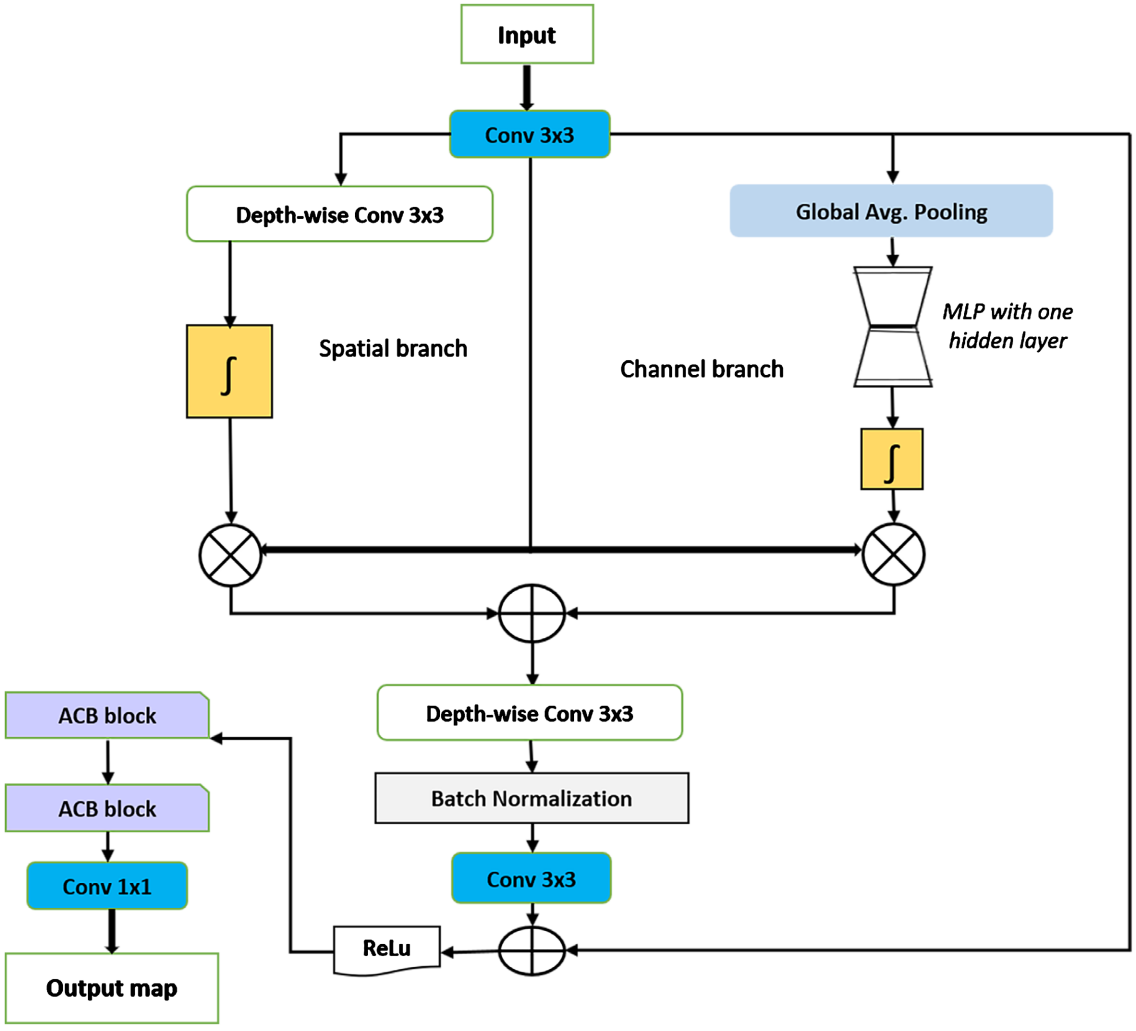

Figure 10. Fine refinement head (FRH) block.

loss estimation bias. Notably, for $\gamma = 0$, the focal loss is equivalent to cross-entropy. Additionally, to ensure accurate localization of various object categories and consider interactions between classes, Dice's loss [55] is used to minimize information loss between the reconstructed and original masks. This loss function is thus formalized by Equation (24)

$$\mathcal{L}_{Dice} = 1 - \frac{2\hat{y}y}{\hat{y} + y + 1} \tag{24}$$

where $\hat{y}$ is the mask tensor predicted by the network and y is the ground truth mask tensor. The final loss function is the sum of the 2 functions of focal loss and Dice loss defined in Equation (25)

$$\mathcal{L}_{total} = \mathcal{L}_{Focal} + \mathcal{L}_{Dice} \tag{25}$$

4. Experiments and Results

To assess our model's effectiveness in semantic segmentation of high-resolution remote sensing images, we tested it on two datasets with diverse urban and rural

complex scenes. Our model's performance was compared against state-of-the-art algorithms from the scientific literature. Two sets of experiments were conducted: the first series focused on metrics like mIoU, Precision, Recall, and mean Pixel Accuracy (mPA) for result comparisons. The second series evaluated the model's computational efficiency, considering factors such as complexity (G) in Flops, required memory (MB), number of parameters (M), and inference speed (Fps). Subsequent sections will detail the datasets, experiments, and analysis of the obtained results.

4.1. Datasets

The first dataset is WHDLD which is a public dataset provided by Wuhan University [20] [23] [56]. It is composed of 4940 RGB color images of dimensions 256 × 256 pixels provided by the Gaofen 1 and ZY-3 satellite sensors over the urban area of Wuhan with a spatial resolution of 2 m. The segmentation masks represent 6 classes of objects namely bare soil, buildings, sidewalks, roads, vehicles, and water. For our experiments, the data was randomly partitioned into 3 subsets, training, validation, and testing according to the ratio 0.7:0.1:0.2. Figure 11 shows the images and labels in WHDLD datasets.

As for the DLRSD dataset, is a dataset containing 2100 RGB color images with a dimension of 256 × 256 pixels [23] [56]. It is composed of images of segmentation masks representing 17 classes of objects encountered both in rural and urban areas. These are airplanes, bare ground, buildings, cars, chaparral, land, docks, mobile-home, pavement, sand, sea, ships, water tanks or fuel, trees and water. The images used to build this dataset come from UC Merced Land Use

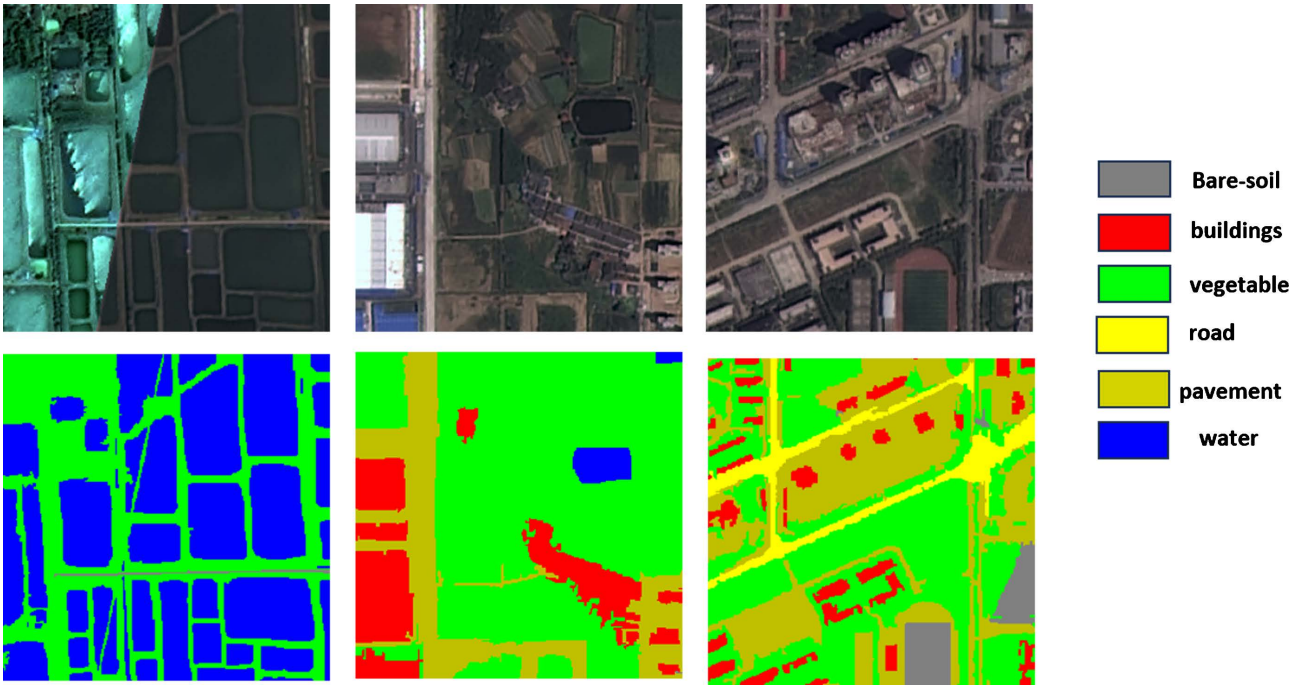

Figure 11. WHDLD images and labels.

data proposed by [57] which includes 2100 images divided into 17 land cover classes of 100 images each. The images have a spatial resolution of 0.3 m. For our experiments, the data was randomly separated into 3 subsets according to the ratio 0.7:0.1:0.2 for training, validation, and testing. Figure 12 shows the images and labels in DLRSD datasets.

These 2 datasets contain a large number of objects to be identified present at different scales within the same image. There we have cars and trees, having resolutions lower than 20 × 20, and buildings, lakes, roads, etc. having resolutions greater than 200 × 200 with chaotic distribution and fuzzy borders. This makes it difficult to classify pixels between neighboring objects.

4.2. Experimental Hypotheses

To study the performance of our algorithm, the test environment included, the operating system Pop!_os in its version 22.04, CUDA12, PyTorch 1.13, and python 3.10. During the training phase, the size of the input images of the different models was fixed at 256 × 256 pixels, the optimizer is of the Adam type

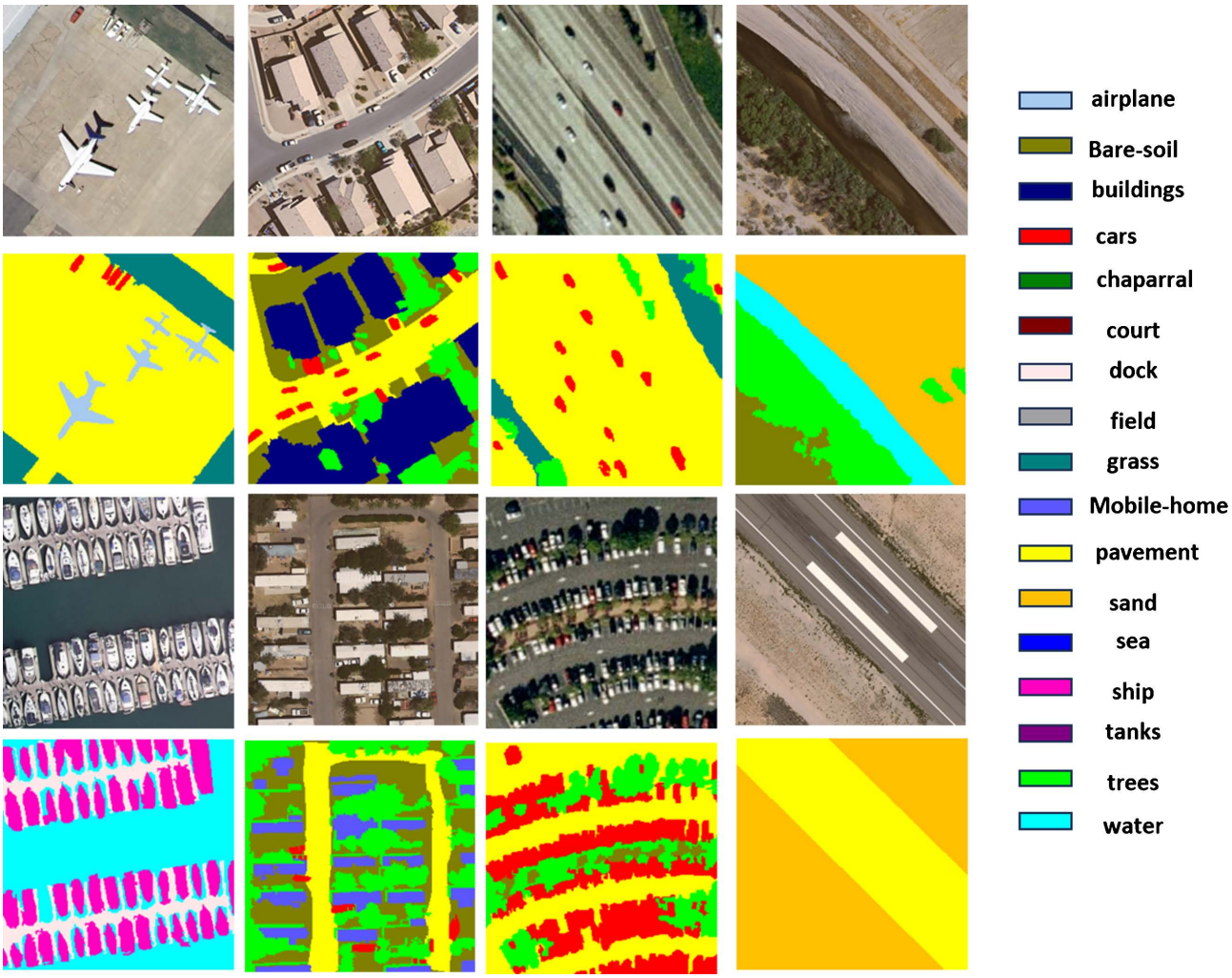

Figure 12. DLRSD images and labels.

introduced by [58], the learning rate was respectively 0.0003 and 0.0001 for WHDLD and DLRSD with a cosine annealing decay strategy [59]. All experiments were implemented on an NVIDIA GeForce RTX 3070 Max-Q GPU with 8GB of VRAM. The datasets were randomly separated into 3 subsets of data including 70% data for training, 10% for validation, and 20% data used for testing. The loss function to be minimized consists of the summation of the Dice loss function proposed by [60], and the focal loss [54] as shown in the previous section in order to be able to mitigate the impact of unbalanced data. The efficiency of our model has been compared with those of the algorithms which, to our knowledge, are among the most efficient in semantic segmentation of satellite images through metrics such as the average intersection over Union (mIoU), the mean Pixel Accuracy (mPA), precision (P), and recall (R) by class [18] [61]. The quantitative evaluation of the performance of our model was done by comparing it to those of the models used in the semantic segmentation of satellite images using the WHDLD or DLRSD datasets. Among these models, we have:

1) CNN models for semantic segmentation: U-Net [32] and U-Net3+ [35], MulitlabelRSIR [56]

2) Those who use pyramidal spatial pooling: DeepLabv3+ [17], PSPNet [31], DPPNet [62], Segment Anything Model (SAM) in Ref. [63].

3) CNN-based attentional networks: MACU-Net introduced in [20], MAU-Net in [18], and Multi-scale network with HL module provided by [19], AttU-Net U-Net with addtitive attention [37], CAU-Net [64].

4) Fully transformer-based networks with a transformer-based decoder: SegFormer intorduced by [44], HrVit Multi-scale vision transformer [65], TMNet multi-branch transformer [66], and Fursformer [67].

4.3. Results and Analysis

The results of the mIoU, Precision, and Recall metrics by object class as well as their mean value of the WHDLD database are summarized in **Table 2** for each class. Those of DLRSD are summarized in **Table 3** for each class. **Table 4** presents global results. These results show the high capacity of our model to correctly locate the objects present in the scene. The results on mean pixels

Table 2. mIoU, Precision (P), Recall (R) and mPA in (%) results by object class for WHDLD dataset.

Object Category	mIoU (%)	P (%)	R (%)	mPA (%)
Building	59.17	73.68	75.02	75.25
Road	60.96	75.76	75.74	77.33
Pavement	43.90	61.53	60.51	44.47
Vegetation	81.01	88.84	90.19	90.97
Bare Soil	40.45	61.17	54.41	61.27
Water	94.34	97.56	96.62	97.56

Table 3. mIoU, Precision (P), Recall (R) and mPA results by object class for DLRSDD dataset.

Object category	mIoU (%)	P (%)	R (%)	mPA (%)
Airplane	62.53	75.14	78.84	83.24
Bare-soil	43.42	66.07	55.87	74.17
Buildings	59.25	79.08	70.26	87.18
Cars	71.77	83.10	84.04	91.20
Chapparal	62.72	68.51	88.12	76.61
Court	68.45	76.17	87.10	84.27
Dock	53.30	74.61	65.11	82.71
Field	98.33	99.86	98.47	98.33
Grass	60.04	76.13	73.96	84.23
Mobile-home	55.47	71.16	71.56	79.26
Pavement	74.57	83.21	87.77	91.31
Sand	61.93	73.28	79.99	81.38
Sea	88.05	89.92	97.69	98.02
Ship	76.81	83.62	90.41	91.72
Tanks	50.72	65.04	69.72	73.14
Trees	66.86	79.98	80.31	88.08
Water	78.84	85.81	90.65	93.91

Table 4. Global statistics for WHDLD and DLRSD datasets.

Dataset	mIoU (%)	P (%)	R (%)	mPA (%)
WHDLD	63.28	76.40	75.41	76.75
DLRSD	66.64	78.28	80.58	79.65

(mPA) by class around 76.43% for WHDLD and 79.56% for DLRSD show that although there are misclassifications, in general, the pixels are mainly represented within the object when it is correctly located.

4.3.1. Comparison Results for WHDLD Dataset

Figure 13 shows the visual results of the segmentation of our model compared to that of MACU-Net. The use of several attentions and the choice of a loss function that takes into account the less represented pixel classes increase the results of the segmentation. As a result, the large homogeneous areas are relatively well identified by the 2 models even if, as for line 1, we can see that our model better discerns the edges and shapes of the buildings, which is not the case with MACU-Net. Moreover, the original model misclassifies two objects of quite similar classes by confusing the pavement with the road. Our model is

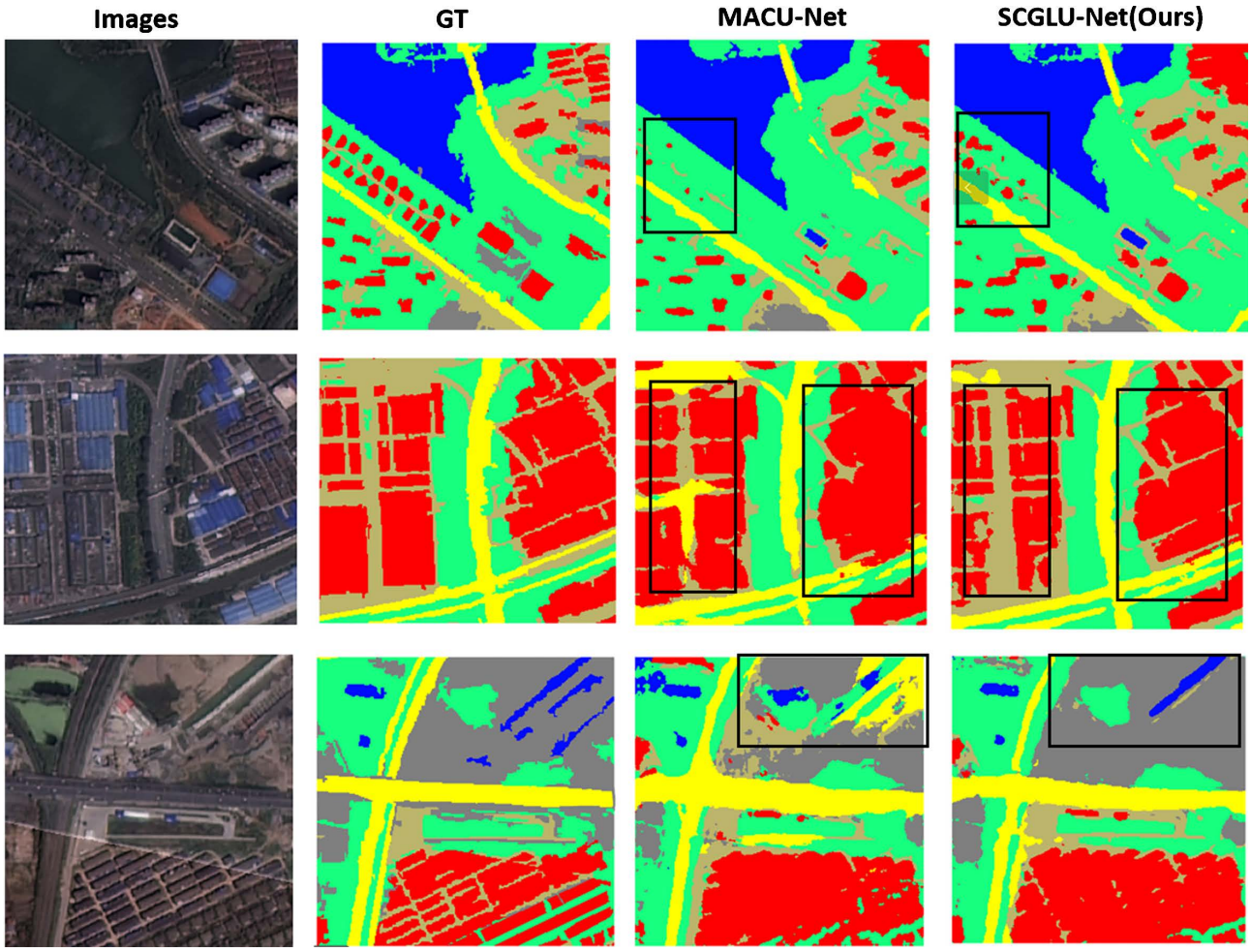

Figure 13. WHDLD test visualization results of MACU-Net and SCGLU-Net.

more sensitive to objects that are very poorly represented in an image because as we can see in line 2, the original model does not identify buildings that are less represented in the image than other objects, while our model manages to detect its presence. In line 3, the original model fails to discriminate fine objects containing large objects, such as the presence of water in bare soil, which is not the case with our model which detects its presence. In the case of the DLRSD dataset, Figure 14 illustrates the segmentation results of our model compared to MACU-Net. The observations made previously are confirmed, since in line 1, the original model very poorly classifies the objects present in the scene, while in line 2, the vehicles are not identified because they are relatively small in size at the mobile home and grass areas. In line 3, the original model is not able to sufficiently discriminate between two close classes such as bare soil and pavements. Unlike the original model, our model exhibits relatively better performance in each of these situations. In order to measure the efficiency of our algorithm during the experiments conducted on the WHDLD and DLRSD datasets, we compared the results obtained with those given by state-of-the-art approaches. The following tables present the results of the mIoU, Precision, Recall, and mPA

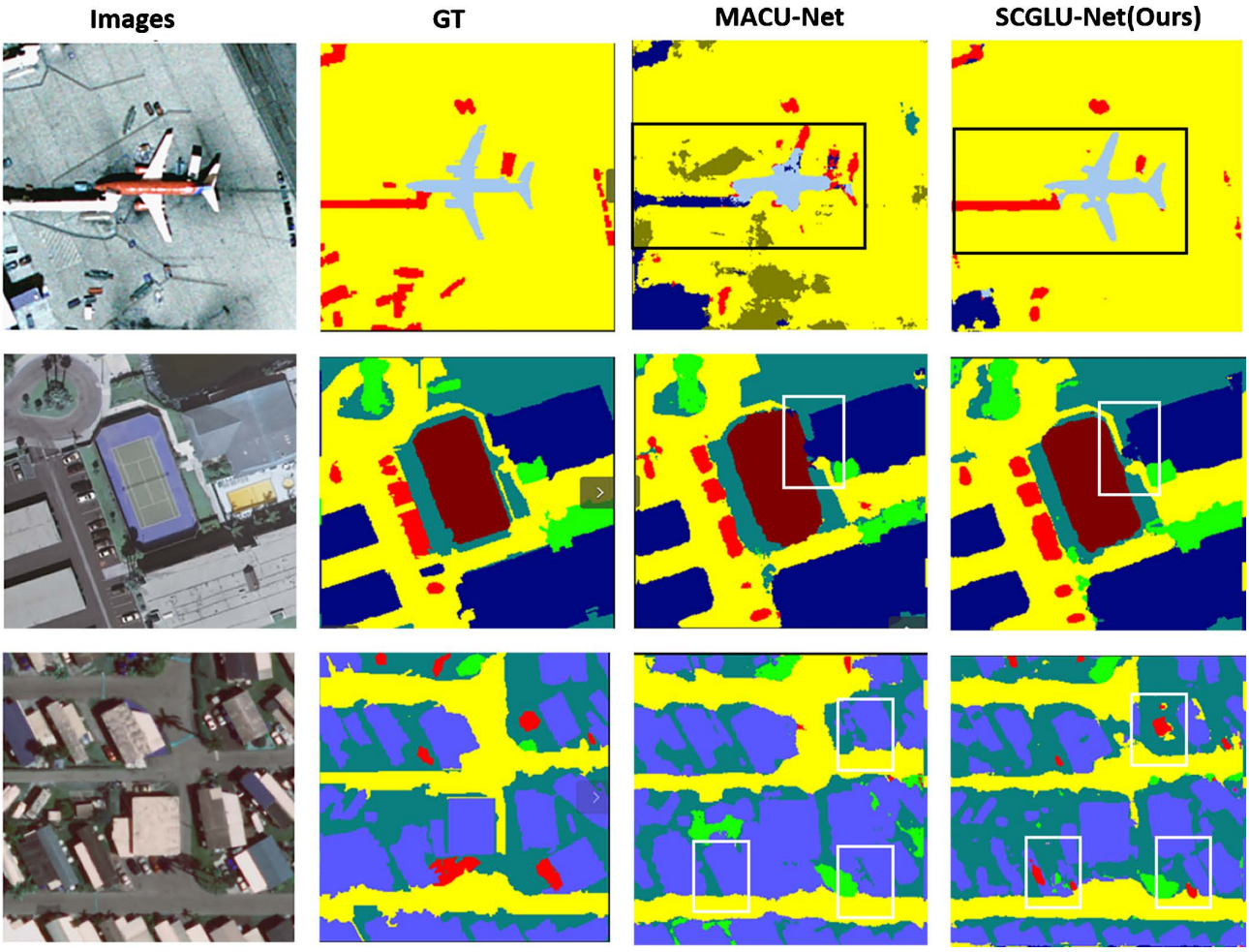

Figure 14. DLRSD test visualization results of MACU-Net and SCGLU-Net.

metrics for WHDLD in Table 5 and DLRSD in Table 6.

As we can see in Table 5, our approach performs better than all the models in terms of mIoU with a gain of +1.54% compared to the best model AttU-Net and +4.37% compared to U-Net which presents the weakest results. The methods combining several attention mechanisms and transformers give better results in mIoU followed by those with spatial pyramidal pooling and pure CNN. Regarding Precision (P) these gains are 7.28% respectively 1.80% relative to multi-lablRSI which has the worst performance with 69.12% and UNet3+ architectures with 74.60%. Only CAU-Net gives better precision than our approach, but we can see the performance is relatively close to 76.40% for our approach and 76.57% for CAU-Net. We can argue that simultaneously local semantic context and range dependencies increase the capacity of the model to detect the class of the objects in the image, unlike an approach that considers only local context. According to the Recall (R) and the mean Pixel Accuracy (mPA), our approach performs all the models with a gain of 3.72% for PSPNet with the worst performance and AttU-Net with a gain of 1.22%. This result shows that our model is among the best models to correctly identify and affect objects in the correct class

Table 5. Performances on WHDLD dataset. The best values are in bold.

Method	mIoU (%)	P (%)	R (%)	mPA (%)
U-Net	58.91	74.00	71.92	71.87
U-net3	61.26	74.60	72.00	72.63
MultilabelRSIR	-	69.12	73.84	74.63
DeepLabV3+	60.27	73.87	72.04	73.49
PSPNet	57.43	73.19	70.57	72.63
MACU-Net	59.57	73.80	72.28	72.51
MAU-Net	60.53	73.11	73.75	69.56
Rmg + HL	59.70	74.10	72.01	75.00
AttU-Net	61.54	75.18	73.39	74.36
CAU-Net	61.27	**76.57**	74.95	74.86
SegFormer	61.5	74.25	71.36	73.02
HrViT	60.71	-	-	73.57
TMNet	61.23	-	-	72.57
SCGLU-Net (ours)	**63.28**	76.40	**75.39**	**76.75**

Table 6. Performances on DLRSD dataset.

Method	mIoU (%)	P (%)	R (%)	mPA (%)
U-Net	56.17	68.46	72.52	71.11
U-Net3+	60.56	73.75	73.41	76.28
MultilabelRSIR	-	76.16	78.80	71.02
DeepLabV3+	59.36	72.35	72.14	74.96
PSPNet	55.81	69.54	68.49	71.28
SAM	58.22	-	-	-
MACU-Net	61.70	76.6	70.78	77.11
MAU-Net	61.90	75.47	74.96	76.72
Rmg + HL	64.10	76.40	74.57	78.95
AttU-Net	59.52	74.87	72.29	76.13
CAU-Net	63.27	76.57	74.95	74.86
SegFormer	60.97	70.62	69.36	73.63
Fursformer	63.32	74.04	74.40	74.04
SCGLU-Net (ours)	**66.64**	77.40	**80.58**	**79.65**

despite the complexity of the scene. In terms of mPA, our approach performs all the models with a gain of 6.59% for MAU-Net with the worst performance and Rmg + HL with a gain of 1.75%.These results show the ability of our model to

correctly classify each object in the WHDLD dataset. This implies that our model assigns object classes to pixels better than any other algorithm. In addition, the power of locating objects remains better than the original model. When we compare our model with those that use an attention mechanism, we can see that using transformers increases performance in the mIou between 1.04% for HrVitand 2% for SegFormer compared to MACU-Net that use channel attention whereas our model increases the mIou of 3.73%. Let's compare the performance in mIou between our model and those with attention like channel attention or spatial attention. We observe that those models increase mIou between 0.96% for MAU-Net and 2% for AttU-net compared to MACU-Net. Consequently, the use of combination attentions and transformers greatly improves the ability of networks to identify objects in the datasets such as WHDLD with à lot of large area objects.

4.3.2. Comparison Results for DLRSD Dataset

Regarding the DLRSD dataset, the results also showed that our algorithm outperforms all other algorithms in all metrics. Compared to the best models, our SCGLU-nets increase mIoU by 3.32% compared with Fursformer, increase precision by 0.88% relative to MACU-Net, increase recall by 5.62% compared with CAU-Net in Recall and 0.70% in mPA compared with Rmg + HL. Unlike some models such as MAU-Net, DeepLabV3+, U-Net, U-Net3+, PSPNet, AttU-net, and Segformer which have seen their performance deteriorate due to the large number of object categories present and their large-scale variability, our model, CAU-Net, Rmg+ Hl, on the contrary, experienced an improvement in their performance in all metrics. This is because models like MAU-Net or MACU-net do not take into account abrupt changes between object scales. By taking into account the interactions between spatial and channel features at different spatial and channel resolutions, our model manages to be sensitive to them compared to Rmg + HL where these interactions are defined in the lowest layers. PSPNet and U-Net obtain the worst performance in all the metrics while U-Net3+ and DeepLabV3+ which combine multi-scale fusion with the convolution product and atrous convolution respectively experience a notable improvement in performance with a respective gain of 5.09% and 3.19% in mIoU, 8.29% and 3.99% in accuracy, 0.9% and recall for U-Net3+ and quite similar results for DeepLabV3+ of 5.12% and 3.85% in mPA. Multi-scale fusion alone is not sufficient for datasets like DLRSD to discriminate the objects. Compared with multiple attention models and more particularly to our model, the performance of pure CNN and spatial pyramidal polling models is much lower if we compare our results to that of the most efficient in these families of models. Our algorithm presents a gain of 6.12% in mIoU, of 4.52% in Precision compared to U-Net3+, of 7.17% in recall, and of 3.37% in mPA. In all metrics, the model's based attentions and transformers give the best performances compared to all model families. The results show that attention helps models improve their capabilities to detect object classes and their locations in the images. Compared to the other

models with attention and transformers, the gain of our approach is comprised of between 2.5% for Rmg + HL and 7.12% for AttU-Net. It is due to the combination of different kinds of attention and transformers.

A study of the results on these 2 datasets shows us that the introduction of multi-scale information fusion as well as the introduction of attention mechanisms greatly increase the capacities of CNN networks in the segmentation of spatial images at very high spatial resolution. However, the combined use of several types of attention, although allowing a performance improvement, is not enough in the case of images in which objects of variable large sizes interact as in DLRSD. In this case, taking into account the interactions between spatial and channel features greatly increases the results obtained and the sensitivity of the network.

4.3.3. Comparison of Network Efficiency

We compared our SCGLU-Net with efficient segmentation networks based on the mIoU, and GPU memory footprint in the number of parameters, and complexity, on the WHDLD test set. The comparison results are listed in Table 7. When we compare the number of parameters and complexity (FLOPs) of each method, our approach performed moderately well in both aspects, indicating that SCGLU-Net does not simply pile up computational effort to obtain high accuracy. Compared with attention models in terms of complexity, our approach

Table 7. Quantitative comparison results on the WHDLD test set with state-of-the-art models. The complexity and number of parameters are measured for a 256 × 256 input on a single NVIDIA GTX 3070 GPU. The best values are in bold.

Method	Parameters (*M*)	Complexity (*G*)	mIoU (%)
U-Net	99.074	54.67	58.91
U-Net3+	26.986	198.67	61.26
MultilabelRSIR	51.95	140.680	59.87
DeepLabV3+	59.34	**22.243**	60.27
PSPNet	72.31	70.058	57.43
DPPNet	61.26	74.60	59.63
MACU-Net	**5.2**	29.65	62.67
MAU-Net	14.55	41.263	60.53
Rmg + HL	40.75	68.23	59.70
AttU-Net	34.87	66.49	61.54
CAU-Net	45.24	70.65	63.27
SegFormer	47.34	79.105	60.45
HrViT	28.62	27.415	60.71
TMNet	20.55	47.49	61.23
SCGLU-Net (ours)	28.46	66.62	**63.28**

with 66.62 Flops is the median value between pure transformers like Segformer, HrVit, and TMNet and hybrid CNN with channel attention or spatial attention like MAU-Net, CAU-Net, Rmg + HL. In terms of the number of parameters, despite its complexity, our model needs fewer parameters than modern transformers like SegFormer and HrVit with better mIoU of 3.45% compared to Segformer and 3.71% compared to HrVit. Compared to hybrid CNN with attention mechanism, except MACU-Net and MAU-Net, SCGLU-Net needs fewer parameters than AttU-Net, CAU-Net, and Rmg + HL. it is because SCGLU-net combines transformers and attention mechanisms and benefits from their advantages. Compared to pure CNN and the spatial pyramidal polling models, in terms of parameters, our model is less than 32.8 M compared to DPPNet, 43.85 M compared to PSPNet, 30.88 M compared to DeepLabV3+ and 23.65 M compared to multilabelRSIR with increases in mIoU by 3.65% for DPPNet, by 5.87% for PSPNet, 3.01% for DeepLabV3+ and 3.41% for multilabelRSIR. Compared to pure CNN, our model needs a bit more parameters +1.47 M than U-Net3+ and less more parameters −70.61 M compared to U-Net models with an increase of mIoU by 2% for U-Net3+ and by 4.37% for U-Net. The results show that using a combination of channel and spatial attention is more computationally efficient and reduces the number of parameters than using only local context for image segmentation. The complexity in Flops also confirms this tendency. In this case, except for U-net and DeepLabV3+, the complexity of SCGLU-Net is less than all the models for pure CNN and spatial pyramidal polling with the best mIoU.

4.4. Abalation Study

To assess the impact of each proposed attention mechanism on our model's performance, ablation experiments were conducted on the WHDLD and DLRSD databases. The evaluation focused on mIoU metrics, as well as complexity (flops), memory (MB), and model speed (fps). Results are summarized in Table 8 for WHDLD and Table 9 for DLRSD. In these experiments, U-Net served as the baseline, lacking any attention mechanism and considering only local context with convolution, in contrast to MACU-Net with ACB convolution, densely connected architecture, and channel attention mechanism CAB.

The baseline is U-Net architecture which only models the local contextual information in the decoder. The loss function of the baseline is the classical categorical cross-entropy.

Propagate attention: We add propagate at the inputs of skip connections in U-Net architecture to add attention to features coming from the encoder layer. The propagate attention achieves less increase of mIoU by 0.17% for WHDLD and 0.23% for DLRSD with a relatively low impact on memory and in terms of Complexity and memory requirement in terms of parameters.

Baseline + Propagate attention + Channel attention: Adding only channel attention increases mIoU by 1.05% for WHDLD and 1.28% for DLRSD. Channel attention has also an impact on memory requirements because the number of

Table 8. Abaltion studies on WHDLD dataset.

Method	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6
Baseline	√	√	√	√	√	
Propagate attention		√	√	√	√	
Channel attention			√			
Spatial attention				√		
SCGL					√	√
FRH						√
mIoU (%)	58.90	59.07	59.95	59.34	61.3	**61.45**
Complexity (G)	54.67	59.74	74.57	78.95	55.57	**43.605**
Memory (MB)	**460.76**	540.28	858.14	1003.78	736.46	1105.79
Speed (fps)	32.80	33.32	33.32	29.65	52.53	**53.26**

Table 9. Abaltion studies on DLRSD dataset.

Method	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6
Baseline	√	√	√	√	√	√
Propagate attention		√	√	√	√	√
Channel attention			√			
Spatial attention				√		√
SCGL					√	
FRH						√
mIoU (%)	56.17	56.40	57.45	56.95	60.17	**60.45**
Complexity (G)	64.10	76.40	74.57	78.95	55.41	**59.02**
Memory (MB)	**466.80**	547.00	864.14	1009.78	742.23	1111.56
Speed (fps)	33.32	34.20	34.84	30.24	51.03	**51.83**

parameters increases by 61% for those of baseline + propagate attention and increases by 86% for baseline parameters for DLRSD. For WHDLD, this augmentation represents 0.34% for baselin + propagate attention and 86% for baseline. The addition of channel attention does not improve notably the speed (Fps) of the model because the speed is still identic for baseline + propagate attention and base + propagate attention + channel attention and in less augmentation of 0.6% for DLRSD.

Baseline + Propagate attention + Spatial attention: Adding only spatial after propagate attention does not increase the performance of the model as shown by the increase of mIou for WHDLD and DLRSD. Also, spatial attention is responsible for the augmentation of memory requirements and complexity in flops. For our two datasets, adding only spatial attention decreases the inference speed of the model by 2.85% for WHDLD and by 4.6% for DLRSD. The conse-

quence is that spatial attention when he was used alone is not optimal in our model for large resolution size because of its quadratic complexity.

Baseline + Propagate attention + SCGL: The impact of this block is so significant. In terms of memory requirements, the SCGL block increases baseline memory by 275.43 MB for the two datasets. It is less than the sum of memory requirements of channel and spatial attention taken individually from baseline, 940.3 MB for WHDLD and DLRSD. Also, with the use of scale block, we notice an increase of speed by 60% in the case of WHDLD and by 53% in the case of DLRSD from baseline. Concerning me, we notice a significant augmentation by 2.85% for WHDLD and by 4% for DLRSD. These results show that combining spatial attention and channel attention globally and locally manner is more suitable than using this attention alone.

Baseline + Propagate attention + SCGL+ FRH: As we see in table Table 8 and table Table 9, adding FRH block also increases less significantly the mIoU by 0.15% for WHDLD and by 0.28% for DLRSD, and the complexity of the model despite the notable augmentation in term of memory requirement by 639 MB from baseline + SCGL and for inference speed by ≈0.80% for the tow datasets. The usage of a standard 2D convolution product can explain this augmentation because this convolution product requires a lot of calculations to estimate the results.

5. Conclusions

Semantic segmentation of high-resolution remote sensing images poses challenges due to the complexity and variability of scenes. This complexity requires considering both local semantic context and long-term dependencies. Our proposed hybrid architecture, SCGLU-Net, integrates CNNs as encoder, combination of transformers and channel attention mechanisms as decoder to address this issue. The SCGL block within this architecture processes spatial and channel attention locally and globally, capturing interactions between descriptors. This is an advancement over conventional methods. Additionally, the architecture introduces Propagate attention in multi-scale fusion to selectively retain pertinent information from encoder, mitigating artifacts observed in concatenation-based approaches. Results on mIoU scores on WHDLD and DLRSD datasets demonstrate enhanced segmentation capabilities for CNN networks in very fine spatial resolution images, with controlled computational complexity.

Future work aims to enhance object boundary segmentation through the integration of a self-attention module.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Chen, B., Xia, M. and Huang, J. (2021) MFANet: A Multi-Level Feature Aggregation

Network for Semantic Segmentation of Land Cover. *Remote Sensing*, **13**, Article 731. https://doi.org/10.3390/rs13040731

[2] Jensen, J.R., Qiu, F. and Patterson, K. (2001) A Neural Network Image Interpretation System to Extract Rural and Urban Land Use and Land Cover Information from Remote Sensor Data. *Geocarto International*, **16**, 21-30. https://doi.org/10.1080/10106040108542179

[3] Wang, J., Zheng, Z., Ma, A., Lu, X. and Zhong, Y. (2021) Loveda: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. ArXiv: 2110.08733.

[4] Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X. and Atkinson, P.M. (2022) UNetFormer: A UNet-Like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, **190**, 196-214. https://doi.org/10.1016/j.isprsjprs.2022.06.008

[5] Zhang, T., Su, J., Liu, C. and Chen, W.-H. (2021) State and Parameter Estimation of the AquaCrop Model for Winter Wheat Using Sensitivity Informed Particle Filter. *Computers and Electronics in Agriculture*, **180**, Article 105909. https://doi.org/10.1016/j.compag.2020.105909

[6] Witharana, C., Bhuiyan, M.A.E., Liljedahl, A.K., Kanevskiy, M., Epstein, H.E., Jones, B.M., Daanen, R., Griffin, C.G., Kent, K. and Jones, M.K.W. (2020) Understanding the Synergies of Deep Learning and Data Fusion of Multispectral and Panchromatic High Resolution Commercial Satellite Imagery for Automated Ice-Wedge Polygon Detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, **170**, 174-191. https://doi.org/10.1016/j.isprsjprs.2020.10.010

[7] Blake, A., Criminisi, A., Cross, G. and Kolmogorov, V. (2010). Image Segmentation of Foreground from Background Layers. US Patent 7,676,081.

[8] Qi, S., Ma, J., Lin, J., Li, Y. and Tian, J. (2015) Unsupervised Ship Detection Based on Saliency and S-Hog Descriptor from Optical Satellite Images. *IEEE Geoscience and Remote Sensing Letters*, **12**, 1451-1455. https://doi.org/10.1109/LGRS.2015.2408355

[9] Goncalves, H., Corte-Real, L. and Goncalves, J.A. (2011) Automatic Image Registration through Image Segmentation and Sift. *IEEE Transactions on Geoscience and Remote Sensing*, **49**, 2589-2600. https://doi.org/10.1109/TGRS.2011.2109389

[10] Simonyan, K., Vedaldi, A. and Zisserman, A. (2013) Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. ArXiv: 1312.6034.

[11] Ding, L., Lin, D., Lin, S., Zhang, J., Cui, X., Wang, Y., Tang, H. and Bruzzone, L. (2022) Looking outside the Window: Wide-Context Transformer for the Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, **60**, Article No. 4410313. https://doi.org/10.1109/TGRS.2022.3168697

[12] Yang, M.Y., Kumaar, S., Lyu, Y. and Nex, F. (2021) Real-Time Semantic Segmentation with Context Aggregation Network. *ISPRS Journal of Photogrammetry and Remote Sensing*, **178**, 124-134. https://doi.org/10.1016/j.isprsjprs.2021.06.006

[13] Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L. and Atkinson, P. M. (2021) ABCNet: Attentive Bilateral Contextual Network for Efficient Semantic Segmentation of Fine-Resolution Remotely Sensed Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, **181**, 84-98. https://doi.org/10.1016/j.isprsjprs.2021.09.005

[14] Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv: 1409.0473.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention Is All You Need. In: Guyon, I., *et al.*, Eds.,

Advances in Neural Information Processing Systems 30, Neural Information Processing Systems Foundation, Inc. (NeurIPS), Long Beach, 6000-6010.

[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.* (2020) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. ArXiv: 2010.11929.

[17] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018) Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 801-818. https://doi.org/10.1007/978-3-030-01234-2_49

[18] Sun, Y., Bi, F., Gao, Y., Chen, L. and Feng, S. (2022) A Multi-Attention UNet for Semantic Segmentation in Remote Sensing Images. *Symmetry*, **14**, Article 906. https://doi.org/10.3390/sym14050906

[19] Wang, G., Zhai, Q. and Lin, J. (2022) Multi-Scale Network for Remote Sensing Segmentation. *IET Image Processing*, **16**, 1742-1751. https://doi.org/10.1049/ipr2.12444

[20] Li, R., Duan, C., Zheng, S., Zhang, C. and Atkinson, P.M. (2022) MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images. *IEEE Geoscience and Remote Sensing Letters*, **19**, Article No. 8007205. https://doi.org/10.1109/LGRS.2021.3052886

[21] Song, C.H., Han, H.J. and Avrithis, Y. (2022) All the Attention You Need: Global-Local, Spatial-Channel Attention for Image Retrieval. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, HI, 3-8 January 2022, 2754-2763. https://doi.org/10.1109/WACV51458.2022.00051

[22] Yang, F., Sun, Q., Jin, H. and Zhou, Z. (2020) Superpixel Segmentation with Fully Convolutional Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 13-19 June 2020, 13964-13973. https://doi.org/10.1109/CVPR42600.2020.01398

[23] Shao, Z., Yang, K. and Zhou, W. (2018) Performance Evaluation of Single-Label and Multi-Label Remote Sensing Image Retrieval Using a Dense Labeling Dataset. *Remote Sensing*, **10**, Article 964. https://doi.org/10.3390/rs10060964

[24] Thanh Noi, P. and Kappas, M. (2017) Comparison of Random Forest, K-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors*, **18**, Article 18. https://doi.org/10.3390/s18010018

[25] Han, B. (2015) Watershed Segmentation Algorithm Based on Morphological Gradient Reconstruction. 2015 2*nd International Conference on Information Science and Control Engineering*, Shanghai, 24-26 April 2015, 533-536. https://doi.org/10.1109/ICISCE.2015.124

[26] Radman, A., Zainal, N. and Suandi, S.A. (2017) Automated Segmentation of Iris Images Acquired in an Unconstrained Environment Using HOG-SVM and GrowCut. *Digital Signal Processing*, **64**, 60-70. https://doi.org/10.1016/j.dsp.2017.02.003

[27] Blaschke, T. (2010) Object Based Image Analysis for Remote Sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, **65**, 2-16. https://doi.org/10.1016/j.isprsjprs.2009.06.004

[28] Carleer, A., Debeir, O. and Wolff, E. (2005) Assessment of Very High Spatial Resolution Satellite Image Segmentations. *Photogrammetric Engineering & Remote Sensing*, **71**, 1285-1294. https://doi.org/10.14358/PERS.71.11.1285

[29] Kirillov, A., Girshick, R., He, K. and Dollár, P. (2019) Panoptic Feature Pyramid Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-*

tern Recognition, Long Beach, 15-20 June 2019, 6399-6408. https://doi.org/10.1109/CVPR.2019.00656

[30] Kotaridis, I. and Lazaridou, M. (2021) Remote Sensing Image Segmentation Advances: A Meta-Analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, **173**, 309-322. https://doi.org/10.1016/j.isprsjprs.2021.01.020

[31] Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2017) Pyramid Scene Parsing Network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2881-2890. https://doi.org/10.1109/CVPR.2017.660

[32] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention-MICCAI* 2015: 18*th International Conference*, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[33] Diakogiannis, F.I., Waldner, F., Caccetta, P. and Wu, C. (2020) ResUNet-a: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, **162**, 94-114. https://doi.org/10.1016/j.isprsjprs.2020.01.013

[34] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N. and Liang, J. (2018) UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *DLMIA* 2018, *ML-CDS* 2018: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Granada, 20 September 2018, 3-11. https://doi.org/10.1007/978-3-030-00889-5_1

[35] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W. and Wu, J. (2020) UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. *ICASSP* 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 1055-1059. https://doi.org/10.1109/ICASSP40776.2020.9053405

[36] Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W. and Lau, R.W. (2021) Location-Aware Single Image Reflection Removal. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 5017-5026. https://doi.org/10.1109/ICCV48922.2021.00497

[37] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., *et al.* (2018) Attention U-Net: Learning Where to Look for the Pancreas. ArXiv: 1804.03999.

[38] Shi, H., Fan, J., Wang, Y. and Chen, L. (2021) Dual Attention Feature Fusion and Adaptive Context for Accurate Segmentation of Very High-Resolution Remote Sensing Images. *Remote Sensing*, **13**, Article 3715. https://doi.org/10.3390/rs13183715

[39] Ding, X., Guo, Y., Ding, G. and Han, J. (2019) ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 1911-1920. https://doi.org/10.1109/ICCV.2019.00200

[40] Woo, S., Park, J., Lee, J.-Y. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[41] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y. and Liu, W. (2019) CCNet: Criss-Cross Attention for Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October 2-2 November 2019, 603-612. https://doi.org/10.1109/ICCV.2019.00069

[42] Yuan, Y., Huang, L., Guo, J., Zhang, C., Chen, X. and Wang, J. (2021) OCNet: Object Context for Semantic Segmentation. *International Journal of Computer Vision*, **129**, 2375-2398. https://doi.org/10.1007/s11263-021-01465-9

[43] Strudel, R., Garcia, R., Laptev, I. and Schmid, C. (2021) Segmenter: Transformer for Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 7262-7272. https://doi.org/10.1109/ICCV48922.2021.00717

[44] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P. (2021) Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in Neural Information Processing Systems*, **34**, 12077-12090.

[45] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

[46] Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P. and Vateekul, P. (2021) Transformer-Based Decoder Designs for Semantic Segmentation on Remotely Sensed Images. *Remote Sensing*, **13**, Article 5100. https://doi.org/10.3390/rs13245100

[47] Sun, L., Zou, H., Wei, J., Cao, X., He, S., Li, M. and Liu, S. (2023) Semantic Segmentation of High-Resolution Remote Sensing Images Based on Sparse Self-Attention and Feature Alignment. *Remote Sensing*, **15**, Article 1598. https://doi.org/10.3390/rs15061598

[48] Liu, T., Luo, R., Xu, L., Feng, D., Cao, L., Liu, S. and Guo, J. (2022) Spatial Channel Attention for Deep Convolutional Neural Networks. *Mathematics*, **10**, Article 1750. https://doi.org/10.3390/math10101750

[49] Yu, F. and Koltun, V. (2015) Multi-Scale Context Aggregation by Dilated Convolutions. ArXiv: 1511.07122.

[50] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[51] Huang, H., Zhou, X., Cao, J., He, R. and Tan, T. (2023) Vision Transformer with Super Token Sampling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 17-24 June 2023, 22690-22699.

[52] Shaw, P., Uszkoreit, J. and Vaswani, A. (2018) Self-Attention with Relative Position Representations. *Proceedings of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **2**, 464-468. https://doi.org/10.18653/v1/N18-2074

[53] Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A. and Shlens, J. (2019) Stand-Alone Self-Attention in Vision Models. In: Wallach, H., *et al.*, Eds., *Advances in Neural Information Processing Systems* 32, NeurIPS 2019, Vancouver, 68-80.

[54] Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017) Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2980-2988. https://doi.org/10.1109/ICCV.2017.324

[55] Jadon, S. (2020) A Survey of Loss Functions for Semantic Segmentation. 2020 *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology* (*CIBCB*), Via del Mar, 27-29 October 2020, 1-7.

https://doi.org/10.1109/CIBCB48159.2020.9277638

[56] Shao, Z., Zhou, W., Deng, X., Zhang, M. and Cheng, Q. (2020) Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **13**, 318-328. https://doi.org/10.1109/JSTARS.2019.2961634

[57] Yang, Y. and Newsam, S. (2010) Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. *Proceedings of the* 18*th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, San Jose, 2-5 November 2010, 270-279. https://doi.org/10.1145/1869790.1869829

[58] Kingma, D. P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization. ArXiv: 1412.6980.

[59] Loshchilov, I. and Hutter, F. (2016) Sgdr: Stochastic Gradient Descent with Warm Restarts. ArXiv: 1608.03983.

[60] Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S. and Jorge Cardoso, M. (2017) Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. *DLMIA* 2017, *ML-CDS* 2017: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Québec City, 14 September 2017, 240-248. https://doi.org/10.1007/978-3-319-67558-9_28

[61] Sokolova, M., Japkowicz, N. and Szpakowicz, S. (2006) Beyond Accuracy, F-Score and Roc: A Family of Discriminant Measures for Performance Evaluation. *Australasian Joint Conference on Artificial Intelligence*, Hobart, 4-8 December 2006, 1015-1021. https://doi.org/10.1007/11941439_114

[62] Sravya, N., Lal, S., Nalini, J., Reddy, C.S., Dell'Acqua, F., *et al*. (2022) Dppnet: An Efficient and Robust Deep Learning Network for Land Cover Segmentation from High-Resolution Satellite Images. *IEEE Transactions on Emerging Topics in Computational Intelligence*, **7**, 128-139. https://doi.org/10.1109/TETCI.2022.3182414

[63] Qi, X., Wu, Y., Mao, Y., Zhang, W. and Zhang, Y. (2023) Self-Guided Few-Shot Semantic Segmentation for Remote Sensing Imagery Based on Large Vision Models. ArXiv: 2311.13200.

[64] Jia, J., Song, J., Kong, Q., Yang, H., Teng, Y. and Song, X. (2023) Multi-Attention-Based Semantic Segmentation Network for Land Cover Remote Sensing Images. *Electronics*, **12**, Article 1347. https://doi.org/10.3390/electronics12061347

[65] Gu, J., Kwon, H., Wang, D., Ye, W., Li, M., Chen, Y.-H., Lai, L., Chandra, V. and Pan, D.Z. (2022) Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, 18-24 June 2022, 12094-12103. https://doi.org/10.1109/CVPR52688.2022.01178

[66] Gao, Y., Zhang, S., Zuo, D., Yan, W. and Pan, X. (2023) TMNet: A Two-Branch Multi-Scale Semantic Segmentation Network for Remote Sensing Images. *Sensors*, **23**, Article 5909. https://doi.org/10.3390/s23135909

[67] Zhang, Z., Liu, B. and Li, Y. (2023) FURSformer: Semantic Segmentation Network for Remote Sensing Images with Fused Heterogeneous Features. *Electronics*, **12**, Article 3113. https://doi.org/10.3390/electronics12143113