Global-Local Deep Fusion: Semantic Integration with Enhanced Transformer in Dual-Branch Networks for Ultra-High Resolution Image Segmentation

Chenjing Liang , Kai Huang * , Jian Mao

Posted Date: 31 May 2024

doi: 10.20944/preprints202405.2059.v1

Keywords: ultra-high resolution image segmentation; enhanced transformer; feature fusion

Article

Global-Local Deep Fusion: Semantic Integration with Enhanced Transformer in Dual-Branch Networks for Ultra-High Resolution Image Segmentation

Chenjing Liang, Kai Huang * and Jian Mao

College of Computer Engineer, Jimei University, Xiamen, 361021, China; 202121331113@jmu.edu.cn (C.L.); maojian@jmu.edu.cn(J.M.)

* Correspondence: kaihuang@jmu.edu.cn(K.H.)

Abstract: The fusion of global contextual information with local cropped block details is crucial for segmenting ultra-high resolution images. In this study, we introduce a novel fusion mechanism termed Global-Local Deep Fusion (GL-Deep Fusion) based on an enhanced transformer architecture, which efficiently integrates global contextual information and local details. Specifically, we propose the Global-Local Synthesis Networks (GLSNet), a dual-branch network where one branch processes the entire original image, while the other branch handles cropped local patches as input. The feature fusion of different branches in GLSNet is achieved through GL-Deep Fusion, significantly enhancing the accuracy of ultra-high resolution image segmentation. Particularly effective is GLSNet in identifying tiny overlapping items. To optimize GPU memory utilization, we meticulously design a dual-branch architecture that proficiently leverages the features it extracts, seamlessly integrating them into the enhanced transformer framework of GL-Deep Fusion. Extensive experiments conducted on challenging benchmarks, including DeepGlobe and Vaihingen datasets, demonstrate that GLSNet achieves a new state-of-the-art performance in terms of GPU memory utilization and segmentation accuracy tradeoff.

Keywords: ultra-high resolution image segmentation; enhanced transformer; feature fusion

1. Introduction

The task of image segmentation, which is considered a crucial and difficult subject in the fields of artificial intelligence and computer vision, involves attributing semantic class labels to each pixel present within an image [1]. It divides the image into distinct regions with semantic information, providing crucial scene understanding and semantic context. These semantic insights are essential in many cutting-edge sectors, including autonomous driving, remote sensing, and medical imaging, where ultra-high resolution images deliver unparalleled detail and information [2,3].

Previously, the enhanced development of deep Convolutional Neural Networks (CNNs) has significantly improved the dependability of image segmentation models. Notable examples include DeepLab [4–7], UNet [8], BSNet [9], PSPNet [10], SegNet [11], ICNet [12], RefineNet [13], EncNet [14], etc. With the development of advancements in autonomous driving and remote sensing, the widespread use of ultra-high resolution images has posed new challenges for image segmentation. Currently, image datasets can be divided into different categories at the pixel level. 2K image resolution is at least 2048 $\times$ 1080 (approximately 2.2M) [15], 4K image resolution is at least 3840 $\times$ 1080 (approximately 4.1M) [16], and 4K ultra-high definition is at least 3840 $\times$ 2160 (approximately 8.3M) [17]. The enormous number of pixels is a considerable barrier to algorithm efficiency, especially given GPU memory constraints.

Downsampling is an effective way to reduce the number of pixels in an image, thus solving the problem of excessive GPU memory usage in ultra-high resolution image segmentation tasks. Nevertheless, an overabundance of downsampling might lead to the compromise of local details. GLNet has achieved good progress by using multi-level FPN (Feature Pyramid Network) [19] to fuse global contextual information from the downsampled input image and local cropped block details from cropped local patches. As shown in Figure 1(1) displays an image from the Vaihingen dataset [20], and its segmentation is presented in Figure 1(2). Ultra-high resolution orthophotos and digital surface

 doi:10.20944/preprints202405.2059.v1

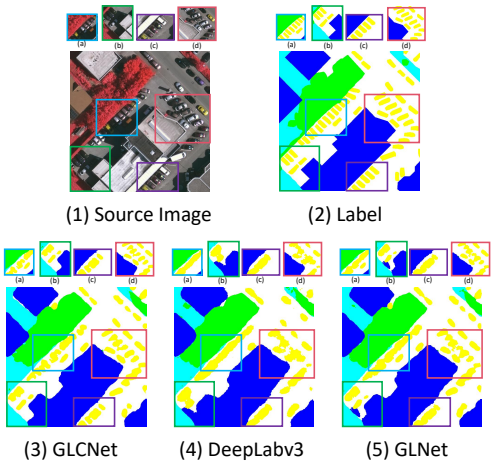

Figure 1. Segmentation Results Example of the Vaihingen dataset. (1) Source Image, (2) Segmentation Labels: white stands for "impervious surfaces", dark blue for "buildings", light blue for "low vegetation", green for "trees", and yellow for "cars", (3) GLSNet Segmentation Result, (4) DeepLabv3 [6] Segmentation Result, (5) GLNet [18] Segmentation Result. Zoomed-in panels (a), (b), (c), and (d) illustrate details. GLSNet demonstrates greater accuracy in handling boundary details compared to DeepLabv3 and GLNet.

models produced by dense image-matching technologies are included in this dataset. The large number of freestanding and little multi-story structures in the collection is noteworthy. Under this dataset, we employed GLNet [18] and DeepLabv3 [6] for prediction. Their results are displayed in Figure 1(4) and (5), respectively. One can observe that the latter performs better in handling segmentation details, especially for overlapping cars (zoomed-in panels (a), (b), (c), (d)). GLNet shows some discrimination ability but still cannot accurately segment each car. This reflects the limited ability of traditional FPNs to maintain the relationship between global contextual information and local details. Since the fusion features are overly reliant on imprecise and one-sided global context information, significant boundary details are missing in the prediction results.

Since Visual Transformer (ViT) [21] introduced the transformer architecture into visual tasks, various state-of-the-art models such as the Masked-Attention Mask Transformer (Mask2Former) [22,23], BSNet [9], and EfficientUNetTransformer [24] have demonstrated the effectiveness of encoder-decoder structures and attention mechanisms. Thus, we construct a unique Global-Local Deep Fusion (referred to as GL-Deep Fusion), used improved transformer structure to better represent the connection between local details and global contextual information. Based on this, we propose the Global-Local Synthesis Networks (GLSNet), a dual-branch network structure with GL-Deep Fusion as the fusion module. By creating local deep branching structures and global shallow branching structures, more complex global contextual information and finer local details can be captured. As shown in Figure 1(3), GLSNet performs excellently in segmenting overlapping cars and object boundaries. Aside from segmentation accuracy, the GPU memory usage brought by the transformer is also a point of concern. Traditional FPN (Feature Pyramid Networks) [19] usually require stacking multiple layers to fuse branch information, which can lead to higher memory usage. In comparison, the transformer has greater potential. UN-EPT [25] employs an Efficient Pyramid Transformer structure for semantic segmentation tasks, resulting in a considerable reduction in GPU memory utilization, which greatly inspired us. In particular, the potential of GPUs and CPUs in terms of computational capacity is constrained by the delay in accessing memory [26–28], which significantly hampers the operational speed of transformers [29,30]. The memory inefficiency of the element-wise functions can be greatly reduced in the processes of multi-head self-attention (MHSA) and frequent tensor reshaping. We discover that there are ways to greatly minimize the time taken for memory access without compromising overall system efficiency. Based upon the analysis and findings, our GL-Deep Fusion employs a dual-encoder and single-decoder

 doi:10.20944/preprints202405.2059.v1

attention mechanism, which, in conjunction with the dual-branch structure, reduces the GPU memory usage significantly. This structure demonstrates potential gains in accuracy and GPU memory usage on the Vaihingen and DeepGlobe [3] datasets.

To summarize, we can summarize our contributions in the following manner:

- We introduce GL-Deep Fusion, which effectively holds the correlation between global semantics and ultra-high resolution image details through its integrated feature representation.
- The global contextual information and partially truncated block details captured by the dual-branch structure can be directly alternated between the dual encoding structures of the GL-Deep Fusion module, thereby avoiding redundant feature computations.
- Our proposed GLSNet has significantly improved GPU memory utilization and segmentation accuracy in the context of ultra-high resolution image segmentation. Compared to GLNet (baseline), it reduces GPU memory usage by 24.1% on the DeepGlobe dataset [3]. The Vaihingen dataset [20] also achieves groundbreaking results.

The organization of this paper is as follows: Section 2 presents an overview of the current state of the related research. Section 3 outlines the network architecture and fusion mechanism that have been designed. Furthermore, the results of our experiments are presented in Section 4. Finally, Section 5 introduces our conclusion and future work.

2. Related Work

2.1. Image Segmentation

FCN [31] is the first CNN structure to achieve high-quality segmentation results in the image segmentation domain. It displays encouraging relearned In contrast to image classification tasks, image segmentation models not only need to recognize pixel-level feature differences but also need to be able to translate distinguishing features learned at successive stages of the network back to the pixel space. The encoder-decoder structure has been frequently used in image segmentation models as a solution to this problem. In U-Net [8,32,33], skip connections are used to link low-level features with high-level features, enhancing the encoder-decoder structure. Numerous applications involving biomedical segmentation have demonstrated the remarkable efficacy of this concept. To increase the receptive field and boost model accuracy, DeepLab [4–6] uses multi-scale feature extraction in conjunction with dilated convolutions. Pyramid pooling modules are applied by PSPNet [10] to gather additional contextual data on a global scale. Dual attention modules are implemented by DANet [34] to take into account cross-channel information and long-range interdependence at various points. However, when applied to ultra-high resolution images, they will encounter significant challenges in terms of GPU memory requirements.

2.2. Segmentation of Ultra-High Resolution Images: Efficiency & Quality

As the dependency on image segmentation for real-time/low-latency tasks, the need to efficiently and qualitatively perform image segmentation on ultra-high resolution images becomes paramount. ENet [35] successfully reduces floating-point computations by adopting an asymmetric encoder-decoder structure and early downsampling. ICNet [12] integrates multi-resolution feature maps for model compression to enhance efficiency. Recently, context aggregation has been a key tactic for overcoming the difficulties associated with ultra-high resolution image segmentation jobs. ParseNet [36] pools scene contexts globally at various levels to apply context aggregation techniques. To aggregate global contextual and high-resolution details, the deep/shallow branches were integrated into ContextNet [37], BiSeNet [38], and GUN [39]. However, these models are not specifically tailored for ultra-high resolution images. The challenge of balancing memory usage and segmentation accuracy remains unresolved. In contrast to the aforementioned studies, our objective is to develop a customized model that tackles the challenges in ultra-high resolution image segmentation tasks.

 doi:10.20944/preprints202405.2059.v1

3. Proposed Method

3.1. Overview

The overview of the entire network structure is shown in Figure 2. GLSNet revolves around three major modules: the global shallow branch, the local deep branch, and the global-local fusion module. In global shallow branching, shallow neural networks are used to collect global contextual information. The local deep branch uses a deep neural network to extract fine local features of the cropped array blocks in parallel. The global-local fusion module combines these branches with GL-Deep Fusion, which is comprised of a dual-cross encoder and a single decoder. Embracing the transformer's potential, GL-Deep Fusion skillfully combines high-quality features that hold both global semantics and local details. Using the cooperative relationship between these three fundamental components, GLSNet can effectively accomplish image segmentation assignments on extremely ultra-high resolution images.

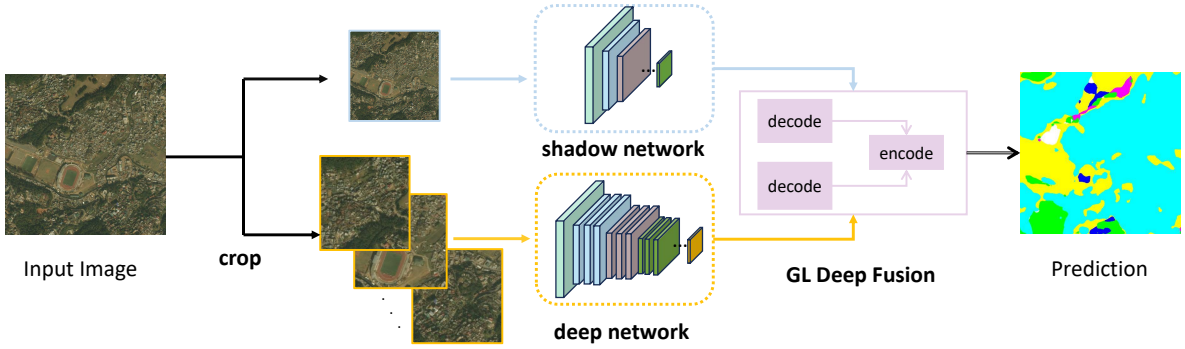

Figure 2. Overview of GLSNet. The global shallow branch and local deep branch use shallow neural networks and deep neural networks, to capture global contextual information and local detail. Then, the GL-Deep Fusion completes the fusion of global-local information, ultimately completing the segmentation task of ultra-high resolution images.

Our two main modules are GL-Deep Fusion and global shallow branch and local deep branch. In Section 3.2, we delve into the intricacies of GL-Deep Fusion. Following that, Section 3.3 is dedicated to exploring the nuances of the global shallow branch and local deep branch.

3.2. GL-Deep Fusion

Transformer attention function [40] can be characterized as a mechanism that transforms a given query and a collection of key-value pairs into a resulting output. The input comprises queries, keys with dimension d_k, and values with dimension d_v. We calculate the scalar products between queries and keys, divide them by $\sqrt{d_k}$, and use the softmax function to determine the weights of values. If we package queries, keys, and values into matrices Q, K, and V, then the attention function for a set of queries is defined as:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{V}) \tag{1}$$

In the context of cross-attention(CA), the queries are sourced from one input sequence, while the usual source for the keys and values is a different input sequence. This enables the model to efficiently establish and capture connections between two separate input sequences.

As shown in Figure 3, GL-Deep Fusion is based on the transformer and is designed with a dual-cross-encode and single-decode structure. We take information sequences from both the global branch and the local branch as input sequences, denoted as F_{global} and F_{local}. Specifically, unlike the conventional Transformer methodology, which requires a number of intricate computations to get the Q, K, and V matrices, we depart from this method to gain memory efficiency. Rather, in our architecture, the keys and values directly come from the local branch's features, and we generate the queries directly from the global branch's features. The design allows for a significant decrease in

 doi:10.20944/preprints202405.2059.v1

redundant computations, which enhances the effectiveness of the dual-branch structure's benefits. The first encoder generates a global-local sequence, which encompasses local relevance information. It can be represented as:

$$CA_{global-local} = Attention(F^{(q)}_{global}, F^{(k)}_{local}, F^{(v)}_{local}) \tag{2}$$

In parallel and symmetrically, the second encoder takes queries from F_{local} and key-value pairs from F_{global}, generating a local-global sequence with globally relevant information. Its representation is:

$$CA_{local-global} = Attention(F^{(q)}_{local}, F^{(k)}_{global}, F^{(v)}_{global}) \tag{3}$$

Finally, the decoder combines the global-local and local-global sequences, producing high-quality fused features with both local details and global semantic information. Specifically, the features extracted by our dual-branch structure exhibit distinct relationships and interdependencies. Therefore, we have opted to exclude the FFN layer, making our architecture lighter.

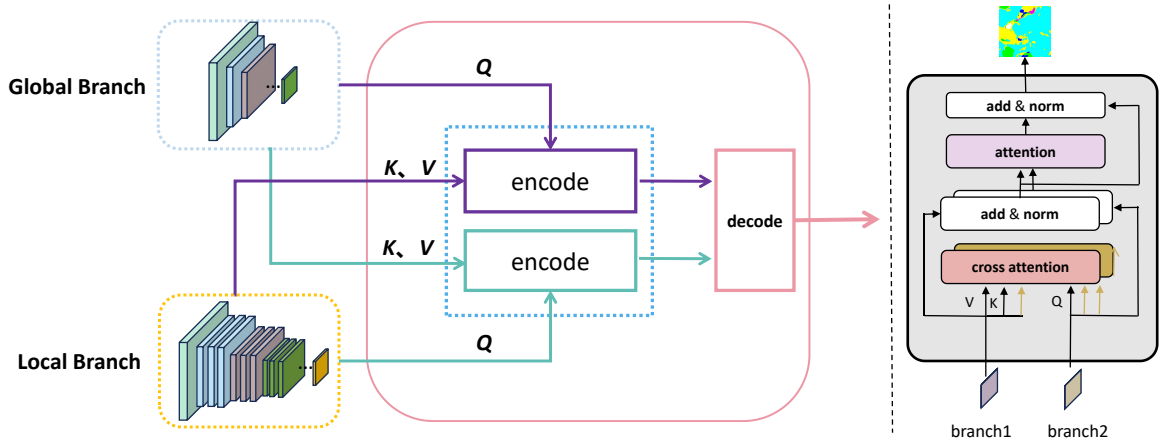

Figure 3. Overview of the GL-Deep Fusion module structure. As shown in the left figure, the core structure of GL-Deep Fusion includes a dual-cross encoder and a single decoder. As shown in the right figure, the two input sequences branch1 and branch2 correspond to the information sequences generated after processing by the global branch and local branch in the left figure, respectively. They are cross-used as the input sequences of the two encodes, obtaining the global-local and local-global sequences. Finally, the two types of sequences are fused into the ultimate feature by the decoder.

3.3. Global Shallow Branch and Local Deep Branch

The global shallow branch and local deep branch of the GLSNet are compatible with various backbone network structures. In our work, we used the standard convolution-based ResNet [41] backbone network (ResNet18 and ResNet50, with 18 and 50 layers, respectively). For large-scale images, employing shallower neural networks can effectively extract global features without incurring significant computational overhead. Accordingly, the shallow branch architecture is ResNet18. Notably, the original ultra-high resolution images are used as input directly, without any preliminary downsampling, in the global shallow branch of GLSNet. With this design, we can extract global contextual information that covers a broad range of the background environments and the semantics of the entire image. Implementing a shallow neural network in global branching improves segmentation accuracy and memory utilization compared to deep neural network designs. In addition, ResNet50 has been used as a deep branch processing architecture.

 doi:10.20944/preprints202405.2059.v1

Table 1. Architectures for ResNet18 and ResNet50 [41]

layer name	output size	18-layer	50-layer
conv1	112×112	7×7, 64, stride 2	
conv2_x	56×56	3×3 max pool, stride 2 $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
	1×1	average pool, 1000-d fc, softmax	
FLOPs		1.8×10^9	3.8×10^9

4. Experiments

4.1. Datasets

DeepGlobe [3]: The DeepGlobe dataset is an ultra-high resolution and challenging remote sensing image dataset consisting of 803 ultra-high resolution images, all with a pixel size of 2448×2448. A random partition of the dataset, consisting of 455, 207, and 142 images, has been made into training, validation, and test sets to efficiently perform model training and evaluation. This division helps ensure the model's performance across different datasets. The data covers multiple different land cover categories, including urban, agriculture, pasture, forest, fire, wasteland, and unknown, totaling seven categories. Compared with previous land cover classification datasets, DeepGlobe offers significantly higher resolution and presents more challenging aspects.

Vaihingen [20]: The dataset from Vaihingen consists of 33 images, each with a spatial resolution of 9 centimeters. These images consist of three channels and have an average size of 2494×2064 pixels. including red, green, and the channels that detect electromagnetic radiation in the near-infrared (NIR) range. The availability of a diverse range of information enables researchers to explore different visual and spatial attributes effectively. The Vaihingen dataset includes six different categories, including impervious surfaces, buildings, low vegetation, trees, cars, and background, covering common land cover types in urban and natural environments, requiring the model to accurately distinguish them.

4.2. Implementation Details

The model utilizes the focal loss function [42] with a weight of 1.0 and a γ value of 6 as its main optimization objective. Additionally, it incorporates two auxiliary losses while setting the regularization coefficient λ to 0.15.

To assess the utilization of the graphics processing unit (GPU) for the given model. The "gpustat" command-line utility gathers information about GPU usage. Training and testing are done on a single GPU card to prevent the computation of any gradients. Additionally, all training is done in batches of six, and each experiment is run on a workstation that has an NVIDIA 1080Ti GPU card installed. In the context of programming, experiments within the PyTorch framework [43] are executed, with the Adam optimizer [44] selected. Take note that research shown that employing distinct learning rates

 doi:10.20944/preprints202405.2059.v1

for local and global branches can improve training outcomes. GLNet has likewise incorporated this finding. [18] (global branch learning rate $\beta_1 = 1 \times 10^{-4}$, local branch learning rate $\beta_2 = 2 \times 10^{-5}$).

4.3. Evaluation Metrics

The proposed GLSNet's performance is assessed using three widely-used metrics: overall accuracy (OA), F_1 score, and mean intersection over union (mIoU) for each class. The OA is a metric that evaluates the accuracy of pixel classification by determining the ratio of accurately classified pixels to the overall number of pixels. The F1 score can be calculated for every category:

$$F_1 = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}, \beta = 1 \tag{4}$$

Additionally, the mean F1 score is determined by averaging all of the F1 scores. If TP is defined as the number of true positives. The numbers for false positives and false negatives are represented as FP and FN, respectively. The IoU formula can be written as follows:

$$IoU = \frac{TP}{FN + FP + TP} \tag{5}$$

Next, we calculate the mean intersection over union (mIoU) by averaging the IoU values across all semantic categories to facilitate comparison.

4.4. Experimental Results

The advantages of GLSNet were verified through experiments that compared it with various models on the Vaihingen and DeepGlobe datasets.

4.4.1. Results and Analysis on the Vaihingen Dataset

On the Vaihingen dataset, the proposed GLSNet was quantitatively compared with state-of-the-art approaches, including BSNet [9], Mask2Former [23] and TransUNet [48] that combine transformer architecture, and classic segmentation models, such as PSPNet [10], S-RA-FCN [47], DeepLabv3 [6], CCEM [46], RefineNet [13], EncNet [14], SegNet [11], UNet [8], and FCN-8s [45]. The comparison metrics included segmentation accuracy for large objects (Low Vegetation, Building), small objects (Tree, Car), background (Impervious Surface), and overall evaluation metrics such as mIoU, mean F1, and overall accuracy OA.

As shown in Table 2, GLSNet outperforms other approaches in overall metrics (mIoU, mean F1, OA). Additionally, it also shows improvement in the segmentation of categories like Impervious Surface, Low Vegetation, Building, and Car. From the results, FCN performed well in segmenting large-scale buildings but lacked attention to detailed information, resulting in mediocre predictions for small objects. S-RA-FCN improved the results by aggregating long-range spatial relationships between pixels based on FCN. UNet and SegNet applied different jump connections to combine low-level and high-level features. RefineNet used a chain residual pool module to capture background context to improve the fuzzy segmentation at the boundary of high spatial resolution images. It achieves better segmentation results at the boundary than UNet and SegNet. CCEM used a height map as an additional channel to extract boundary information. It performs better than RefineNet. DeepLabv3+ used dilated convolution to capture contextual information, further strengthening the segmentation advantage for large objects.

 doi:10.20944/preprints202405.2059.v1

Table 2. Comparison results of various approaches on the Vaihingen dataset.

Method	Impervious Surface	Building	Low Vegetation	Tree	Car	OA	Mean F1	mIoU
FCN-8s [45]	90.0	93.0	77.7	86.5	80.4	88.3	85.5	75.5
UNet [8]	90.5	93.3	79.6	87.5	76.4	89.2	85.5	75.5
SegNet [11]	90.2	93.7	78.5	85.8	83.9	88.5	86.4	76.8
EncNet [14]	91.2	94.1	79.2	86.9	83.7	89.4	87.0	77.8
RefineNet [13]	91.1	94.1	79.8	87.2	82.3	88.9	86.9	77.1
CCEM [46]	91.5	93.8	79.4	87.3	83.5	89.6	87.1	78.0
DeepLavb3 [6]	91.4	94.7	79.6	87.6	85.8	88.9	87.8	79.0
S-RA-FCN [47]	90.5	93.8	79.6	87.5	82.6	89.2	86.8	77.3
PSPNet [10]	90.6	94.3	79.0	87.0	70.7	89.1	84.3	74.1
TransUNet [48]	92.2	93.9	83.7	**88.3**	87.4	89.3	89.1	80.4
Mask2Former [23]	91.4	94.2	82.0	86.4	86.0	88.3	88	78.1
BSNet [9]	92.1	94.4	83.1	**88.3**	86.7	90.3	88.9	80.2
GLSNet	**94.4**	**95.1**	**83.4**	87.6	**90.4**	**90.9**	**90.2**	**81.4**

Our model, GLSNet, fully integrates global contextual information and local detail information, surpassing previous state-of-the-art approaches in multiple metrics. Compared to BSNet, our model achieved gains of 1.28%, 0.6%, and 1.2% in mean F1, OA, and mIoU, respectively. In terms of segmentation accuracy for specific categories of Impervious Surfaces, Buildings, Low Vegetation, and Cars, there were improvements of 2.3%, 0.7%, 0.3%, and 3.7%, respectively. Most importantly, the mean F1 of GLSNet is the only one among all the approaches mentioned above that reached 90%.

4.4.2. Results and Analysis on the DeepGlobe Dataset

On the DeepGlobe dataset, the proposed GLSNet was also quantitatively compared with state-of-the-art approaches, including TransUnet [48], Mask2Former [23], FCN-8s [45], DeepLabv3+ [7], SegNet [11], PSPNet [10], ICNet [12], UNet [8], and GLNet [18] (as the baseline for this dataset). This comparison not only focused on segmentation accuracy (mIoU) but also measured GPU memory usage (Memory).

As shown in Table 3, all approaches achieved better mIoU results after adding the global branch compared to using only local patches. But this also resulted in a sharp increase in GPU memory usage. Most approaches failed to balance segmentation accuracy and GPU memory usage effectively. Among all the methods listed in the table, only GLNet employs global-local information sharing and achieves higher mIoU with less memory consumption. Therefore, GLNet is used as the benchmark model for this dataset.

Table 3. mIoU and inference GPU memory usage for predictions on the DeepGlobe test set.

Model	Local Inference		Global Inference	
	mIoU [%]	Memory [MB]	mIoU [%]	Memory [MB]
U-Net[8]	37.3	949	38.4	5507
ICNet[12]	35.5	1195	40.2	2557
PSPNet[10]	53.3	1513	56.6	6289
SegNet[11]	60.8	1139	61.2	10339
DeepLabv3+[7]	63.1	1279	63.5	3199
FCN-8s[45]	64.3	1963	70.1	5227
Mask2Former[23]	66.7	3458	70.3	23577
TransUnet[48]	68.2	2436	70.2	6283
	Local & Global			
	mIoU [%]		Memory [MB]	
GLNet[18](baseline)	71.6		1865	
GLSNet	**72.4**		**1414**	

 doi:10.20944/preprints202405.2059.v1

Compared to the baseline model GLNet, the proposed GLSNet made breakthroughs in both mIoU and GPU memory usage. The mIoU reached 72.4%, an improvement of 0.8% compared to the baseline model. More importantly, GPU memory usage was significantly reduced by 451MB, a decrease of 24.1%. This makes GLSNet more advantageous in terms of running speed and resource usage, providing better possibilities for practical applications.

4.5. Ablation Experiments

4.5.1. The effects of Shallow-Deep Branch and GL-Deep Fusion

As shown in Table 4, we designed three models: Shallow-Deep, Shallow-Shallow, and Deep-Deep. These three models differ in their design for the Global backbone, Local backbone, and Fusion module. They are used to evaluate the impact of the Shallow-Deep branch collaborative strategy and GL-Deep Fusion structure on ultra-high resolution image segmentation. Note, the benchmark model, GLNet, is also included for comparison.

Table 4. Illustrations of network architectures for various model designs.

	Global backbone	Local backbone	Fusion
GLNet(baseline)	ResNet50	ResNet50	FPN
Shallow-Deep	ResNet18	ResNet50	GL-Deep Fusion
Shallow-Shallow	ResNet18	ResNet18	GL-Deep Fusion
Deep-Deep	ResNet50	ResNet50	FPN + GL-Deep Fusion

On the DeepGlobe dataset, we tested the mIoU and GPU Memory Usage metrics for various models. As shown in Table 5, the Deep-Deep model with the GL-Deep Fusion strategy achieved a 1% higher mIoU than the benchmark model GLNet. This indicates that the introduction of the GL-Deep Fusion enhances segmentation performance. Compared to the Shallow-Deep model using ResNet50, the Shallow-Shallow model with ResNet18 reduced GPU memory usage by 370 MB. This suggests that the choice of neural network layers significantly impacts model GPU memory usage. The GPU memory usage for Shallow-Shallow, Shallow-Deep, and GLNet models shows an increasing trend. Therefore, the cooperative strategy of shallow and deep branching plays a key role in reducing GPU memory consumption. Moreover, the GL-Deep Fusion fusion strategy not only improves segmentation accuracy but also slightly reduces GPU memory usage.

Table 5. Changes in mIoU and GPU memory usage for different network architecture models.

	mIoU [%]	Memory [MB]
GLNet(baseline)	71.6	1865
Shallow-Deep	72.4	1414
Shallow-Shallow	71.9	**1044**
Deep-Deep	**72.6**	2903

Considering segmentation accuracy and GPU memory usage, we chose the Shallow-Deep model as the primary structure for GLSNet. Compared to the benchmark model, GLNet, GLSNet improved the mIoU metric by 0.8% and reduced GPU memory usage by 451 MB. In summary, the introduction of the Shallow-Deep branch collaborative strategy and the GL-Deep Fusion structure has significantly enhanced both the segmentation precision and memory efficiency of GLSNet.

4.5.2. The Effect of Transformer Attention

To delve deeper into the impact of the transformer attention mechanism, we adopted three distinct fusion module strategies: (1) The deep feature map sharing strategy based on FPN (from GLNet); (2) The self-attention mechanism of the Dual Attention Network (DANet), which we refer to

 doi:10.20944/preprints202405.2059.v1

as Attention (DANet) [34]; and (3) The GL-Deep Fusion based on the transformer attention mechanism. We conducted thorough testing and analysis of these three strategies on the DeepGlobe dataset.

As shown in Table 6, the information fusion effect of GL-Deep Fusion is significantly superior to the other two approaches. Compared with using Attention(DANet), the model employing GL-Deep Fusion as the fusion module improves 1% in the mIoU metric and also reduces GPU memory usage by 96MB. This highlights the distinct advantage of the transformer attention mechanism in integrating global contextual information and local details. The GL-Deep Fusion module plays a pivotal role in enhancing the accuracy of ultra-high resolution image segmentation and reducing GPU memory usage.

Table 6. Changes in mIoU and GPU memory usage for different fusion module designs

	mIoU [%]	Memory [MB]
GLNet(baseline) [18]	71.6	1865
Attention(DANet) [34]	71.4	1510
Attention(transformer) [40]	**72.4**	**1414**

4.6. Visualization Results and Analysis

The segmentation outputs of various common techniques exhibit distinct patterns when visualized in the Vaihingen dataset. As shown in Figure 4, in the DeepLabv3 and GLNet, there are problems where large independent areas cannot be accurately segmented due to interference from boundary details. In contrast, GLSNet excels in choosing segmentation boundaries. For the stacking of small target objects, GLSNet has proven to be more precise in segmenting these objects compared to GLNet and DeepLabv3.

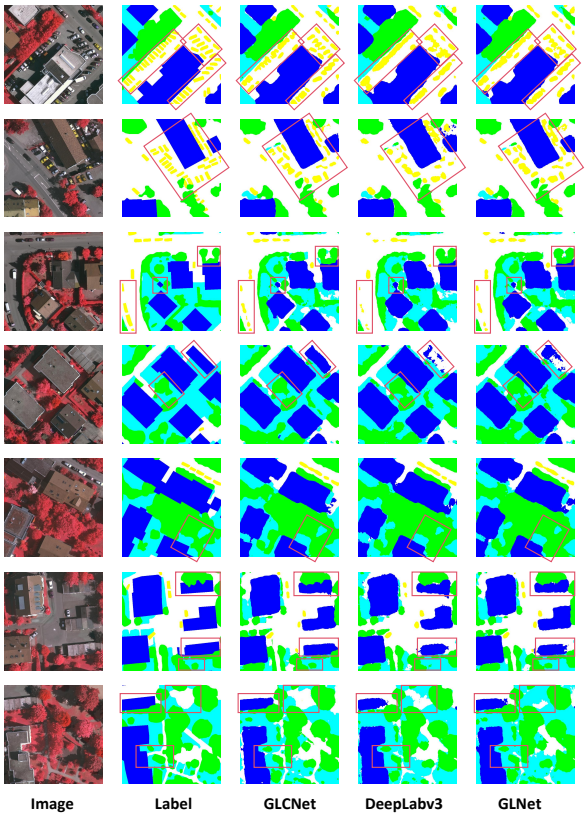

Figure 4. Comparison of segmentation results between GLSNet, DeepLabv3, and GLNet on the Vaihingen dataset. Note that the red boxes in the figure are mainly used to indicate parts where there are significant differences in segmentation results.

 doi:10.20944/preprints202405.2059.v1

5. Conclusion

We have presented GLSNet, a segmentation model optimized for ultra-high resolution images, that prioritizes memory efficiency. It creates a network structure composed of a shallow branch that covers the global context and a deep branch that focuses on local details, ensuring the effective collection of both global and local information. The innovative GL-Deep Fusion seamlessly combines global contextual information and local intricacies, bringing about a transformative effect. GLSNet showcases its competitive performance on both the DeepGlobe and Vaihingen datasets using this method. Specifically, it excels at producing exceptional outcomes in the process of separating overlapping small objects within an image. We consider it essential to strike an improved balance between the utilization of GPU memory and the accuracy of segmentation when exploring ultra-high resolution image research. Therefore, the GLSNet network we designed proved to be a key solution. It is a powerful tool for solving the problem of ultra-high resolution image segmentation.

Although GLSNet has already shown efficient memory usage, there remains untapped potential for additional optimization. In our upcoming research, we aim to further investigate various prospects related to ultra-high-resolution image segmentation. We intend to expand the range of uses for GLSNet to incorporate a greater variety of real-life situations. Simultaneously, We will explore the fusion of multi-modal data, including the integration of ultra-high resolution images with LiDAR, radar, or hyperspectral imagery, aiming to improve both segmentation accuracy and contextual comprehension. These endeavors will contribute to further enhancing the performance and applicability of ultra-high resolution image segmentation technology.

Author Contributions: The authors confirm their contribution to the paper as follows: study conception and design: Chenjing Liang, Kai Huang, Jian Mao; data collection: Chenjing Liang; analysis and interpretation of results: Chenjing Liang; draft manuscript preparation: Chenjing Liang, Kai Huang, Jian Mao. All authors reviewed the results and approved the final version of the manuscript.

Funding Statement: Kai Huang reports financial support was provided by the Natural Science Foundation (3502Z202372018) of Xiamen, China, and the Department of Education (JAT232012) of the Fujian Province of China. Jian Mao reports financial support was provided by the Natural Science Foundation (2021J01858) of Fujian Province of China and the Xiamen Science and Technology Subsidy Project (2023CXY0318).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available within the article. The authors confirm that the data supporting the findings of this study are available within the article.

Acknowledgments: We express our gratitude to the faculty members of the College of Computer Engineering at Jimei University for providing the necessary technical resources and instructional guidance for this research.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence* **2021**, *44*, 3523–3542.

 doi:10.20944/preprints202405.2059.v1

2. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.
3. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 172–181.
4. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* **2014**.
5. 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 transactions on pattern analysis and machine intelligence* **2017**, *40*, 834–848.
6. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* **2017**.
7. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.
8. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 2015, pp. 234–241.
9. Hou, J.; Guo, Z.; Wu, Y.; Diao, W.; Xu, T. BSNet: Dynamic hybrid gradient convolution based boundary-sensitive network for remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–22.
10. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.
11. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* **2017**, *39*, 2481–2495.
12. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. Icnet for real-time semantic segmentation on high-resolution images. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 405–420.
13. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1925–1934.
14. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.
15. Ascher, S.; Pincus, E. *The filmmaker's handbook: A comprehensive guide for the digital age*; Penguin, 2007.
16. Lilly, P. Samsung launches insanely wide 32: 9 aspect ratio monitor with hdr and freesync 2, 2017.
17. Initiatives, D.C. Digital Cinema System Specification, Version 1.3 **2018**. Available online at http://dcimovies.com/specification/DCI%20DCSS%20Ver1-3%202018-0627.pdf.
18. Chen, W.; Jiang, Z.; Wang, Z.; Cui, K.; Qian, X. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8924–8933.
19. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
20. ISPRS Vaihingen Dataset. https://paperswithcode.com/dataset/isprs-vaihingen.
21. 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 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
22. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems* **2021**, *34*, 17864–17875.
23. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 1290–1299.
24. AlMarzouqi, H.; Saoud, L.S. Semantic Labeling of High Resolution Images Using EfficientUNets and Transformers. *IEEE Transactions on Geoscience and Remote Sensing* **2023**.
25. Zhu, F.; Zhu, Y.; Zhang, L.; Wu, C.; Fu, Y.; Li, M. A unified efficient pyramid transformer for semantic segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2667–2677.

 doi:10.20944/preprints202405.2059.v1

26. Gu, J.; Zhu, H.; Feng, C.; Liu, M.; Jiang, Z.; Chen, R.T.; Pan, D.Z. Towards memory-efficient neural networks via multi-level in situ generation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5229–5238.
27. Jiang, C.; Qiu, Y.; Shi, W.; Ge, Z.; Wang, J.; Chen, S.; Cérin, C.; Ren, Z.; Xu, G.; Lin, J. Characterizing co-located workloads in alibaba cloud datacenters. *IEEE Transactions on Cloud Computing* **2020**, *10*, 2381–2397.
28. Venkat, A.; Rusira, T.; Barik, R.; Hall, M.; Truong, L. SWIRL: High-performance many-core CPU code generation for deep neural networks. *The International Journal of High Performance Computing Applications* **2019**, *33*, 1275–1289.
29. Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems* **2022**, *35*, 16344–16359.
30. Ivanov, A.; Dryden, N.; Ben-Nun, T.; Li, S.; Hoefler, T. Data movement is all you need: A case study on optimizing transformers. *Proceedings of Machine Learning and Systems* **2021**, *3*, 711–732.
31. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
32. Liu, D.; Wen, B.; Liu, X.; Wang, Z.; Huang, T.S. When image denoising meets high-level vision tasks: A deep learning approach. *arXiv preprint arXiv:1706.04284* **2017**.
33. Liu, D.; Wen, B.; Jiao, J.; Liu, X.; Wang, Z.; Huang, T.S. Connecting image denoising and high-level vision tasks via deep learning. *IEEE Transactions on Image Processing* **2020**, *29*, 3695–3706.
34. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3146–3154.
35. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147* **2016**.
36. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579* **2015**.
37. Poudel, R.P.; Bonde, U.; Liwicki, S.; Zach, C. Contextnet: Exploring context and detail for semantic segmentation in real-time. *arXiv preprint arXiv:1805.04554* **2018**.
38. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 325–341.
39. Mazzini, D. Guided upsampling network for real-time semantic segmentation. *arXiv preprint arXiv:1807.07466* **2018**.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
42. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
43. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch **2017**.
44. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
45. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
46. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing* **2018**, *135*, 158–172.
47. Mou, L.; Hua, Y.; Zhu, X.X. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12416–12425.
48. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* **2021**.