applied sciences

MDPI

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

Abstract: The fusion of global contextual information with local cropped block details is crucial for segmenting ultra-high resolution images. In this study, A novel fusion mechanism termed global–local deep fusion (GL-Deep Fusion) is introduced, based on an enhanced transformer architecture that 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. Identifying tiny overlapping items is a task where the model excels, demonstrating its particular effectiveness. To optimize GPU memory utilization, a dual-branch architecture was meticulously designed. This architecture proficiently leverages the features it extracts and seamlessly integrates them into the enhanced transformer framework of GL-Deep Fusion. Benchmarks on the DeepGlobe and Vaihingen datasets demonstrate the efficiency and accuracy of the proposed model. It significantly reduces GPU memory usage by 24.1% on the DeepGlobe dataset, enhancing segmentation accuracy by 0.8% over the baseline model. On the Vaihingen dataset, our model delivers a Mean F1 score of 90.2% and achieves a mIoU of 90.9%, highlighting its exceptional memory efficiency and segmentation precision.

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

check for updates

Citation: Liang, C.; Huang, K.; Mao, J. Global–Local Deep Fusion: Semantic Integration with Enhanced Transformer in Dual-Branch Networks for Ultra-High Resolution Image Segmentation. *Appl. Sci.* **2024**, *14*, 5443. https://doi.org/10.3390/app14135443

Academic Editor: Thomas Lindner

Received: 16 May 2024
Revised: 19 June 2024
Accepted: 20 June 2024
Published: 23 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

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. Furthermore, 2 K image resolution is at least 2048 × 1080 (approximately 2.2 M) [15], 4 K image resolution is at least 3840 × 1080 (approximately 4.1 M) [16], and 4 K ultra-high definition is at least 3840 × 2160 (approximately 8.3 M) [17]. The enormous number of pixels is a considerable barrier to algorithm efficiency, especially given GPU memory constraints.

Downsampling is recognized as an effective way to reduce the number of pixels in an image, addressing 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 a multi-level feature pyramid network (FPN) [18] to fuse global contextual information from the downsampled input image and local cropped block details from cropped local patches. Figure 1(1) displays an image from the Vaihingen dataset [19], and its segmentation is presented in Figure 1(2). Ultra-high resolution orthophotos and digital surface 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 [20] and DeepLabv3 [6] for prediction. Their results are displayed in Figure 1(4,5), 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.

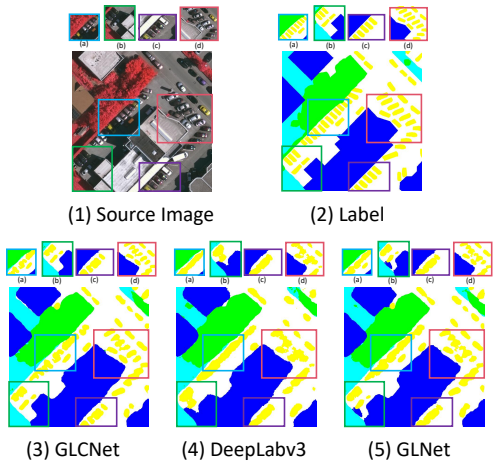


Figure 1. Segmentation Results Example from the Vaihingen dataset.

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 constructed a unique global–local deep fusion (referred to as GL-Deep Fusion) and used an improved transformer structure to better represent the connection between local details and global contextual information. Based on this, the global–local synthesis network (GLSNet) is proposed, featuring a dual-branch network structure with GL-Deep Fusion serving 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 feature pyramid networks (FPNs) [18] usually require stacking multiple layers to fuse branch information, which can lead to higher memory usage. In comparison, the transformer has greater potential. UNEPT [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. The study reveals that there are methods to significantly reduce the time taken for memory access without compromising overall system efficiency. Based on the analysis and findings, the proposed GL-Deep Fusion utilizes a dual-encoder and single-decoder attention mechanism. This mechanism, combined with the dual-branch structure, significantly reduces GPU memory usage. This structure demonstrates potential gains in accuracy and GPU memory usage on the Vaihingen and DeepGlobe [3] datasets.

To 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 [19] 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 6 introduces our conclusion and future work.

2. Related Work

2.1. Image Segmentation

Advancements in image segmentation have been pivotal to the field of computer vision, with models like FCN [31], U-Net [8,32,33], and DeepLab [4–6] laying the groundwork for modern techniques. Recent innovations, such as BSNet [9], MaskFormer [22], and Mask2Former [23], have further pushed the boundaries by introducing novel attention mechanisms for segmentation tasks. These models have demonstrated remarkable efficacy in various applications, from biomedical imaging to autonomous vehicle perception. However, they also present challenges when applied to ultra-high resolution images, particularly in terms of GPU memory requirements and processing speed.

In this context, it is crucial to consider the balance between segmentation accuracy and computational efficiency. Figure 1, which provides a visual comparison of segmentation results on the Vaihingen dataset, is particularly instructive. It showcases the performance of GLSNet alongside DeepLabv3 [6] and GLNet [20], highlighting the ability of different models to capture semantic details and boundary precision. The source image and segmentation labels are depicted, with distinct colors representing different semantic classes: white for “impervious surfaces”, dark blue for “buildings”, light blue for “low vegetation”, green for “trees”, and yellow for “cars”. The segmentation results from GLSNet demonstrate a superior ability to handle boundary details, as evidenced by the zoomed-in panels (a), (b), (c), and (d), which reveal the nuances of each model’s performance. The visual evidence presented in Figure 1 underscores the significance of our approach, which aims to achieve high segmentation accuracy without compromising on computational efficiency. This balance is essential for the practical deployment of segmentation models in real-world applications.

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

As the dependency on image segmentation for real-time/low-latency tasks increases, the need to efficiently and qualitatively perform image segmentation on ultra-high resolution images becomes paramount. ENet [34] 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 [35] 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 [36], BiSeNet [37], and GUN [38]. 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.

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 comprises 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.

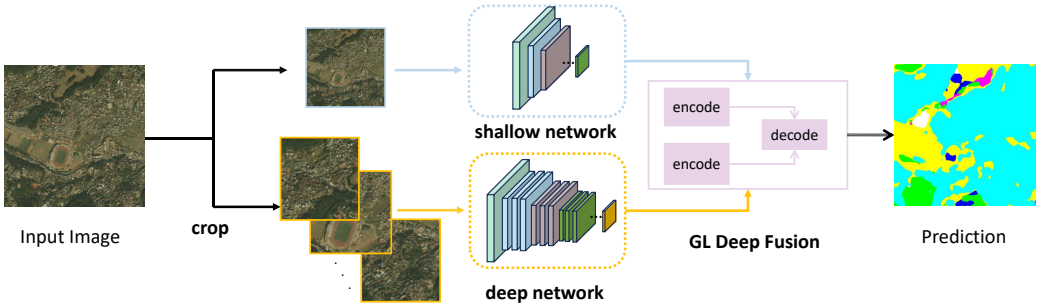


Figure 2. Overview of GLSNet. The global shallow branch and local deep branch use shallow neural networks and deep neural networks to capture global contextual information and local detail. Then, the GL-Deep Fusion completes the fusion of global–local information, ultimately completing the segmentation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

The two main modules proposed are GL-Deep Fusion and the global shallow branch, along with the local deep branch. In Section 3.2, the intricacies of GL-Deep Fusion are explored. Subsequently, Section 3.3 is dedicated to an exploration of the nuances of the global shallow branch and the local deep branch.

3.2. GL-Deep Fusion

As noted in Section 3.1, the integration of features from both branches is of paramount importance for the performance of GLSNet. To tackle this issue, the global–local deep fusion method (GL-Deep Fusion) was designed. The core idea is to leverage an enhanced transformer architecture that employs a dual-encoder and single-decoder mechanism to amalgamate global and local features effectively. Unlike traditional transformer frameworks that necessitate extensive computations for deriving the matrices of queries, keys, and values, the proposed encoders are adept at directly aligning with the features provided by the dual-branch network, thereby significantly reducing computational overhead and memory consumption. This innovative approach facilitates a memory-efficient fusion process that capitalizes on the distinct advantages of both the global and local branches, culminating in a more robust and accurate model for feature integration.

Transformer attention function [39]: The attention function is defined for matrices of queries Q, keys K, and values V where the dimensions are $d_k \times n$ and $d_v \times n$, respectively. Here, n denotes the number of elements within the set, while d_k and d_v are the dimensionalities associated with the keys and values. This function calculates the dot products between the queries and keys, applies a scaling factor of $\sqrt{d_k}$ to stabilize the softmax operation, and subsequently uses the softmax function to generate a weighted distribution across the values.

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}}) \cdot V, \quad (1)$$

where QK^T corresponds to the dot product of the query matrix with the transpose of the key matrix. The division by $\sqrt{d_k}$ serves as a form of normalization to maintain the balance of the softmax output.

Branch-Attention (B2A): B2A (Figure 3 right) facilitates the interaction between two distinct branches, allowing one to query information from the other. This is particularly useful for establishing relationships between different data representations within the model.

$$B2A_{branc1,branch2} = Attention(F^{(q)}_{branch1}, F^{(k)}_{branch2}, F^{(v)}_{branch2}), \quad (2)$$

where $F^{(q)}_{branch1}$ represents the query matrices from the branch1, and $F^{(k)}_{branch2}$, $F^{(v)}_{branch2}$ are the key and value matrices derived from branch2.

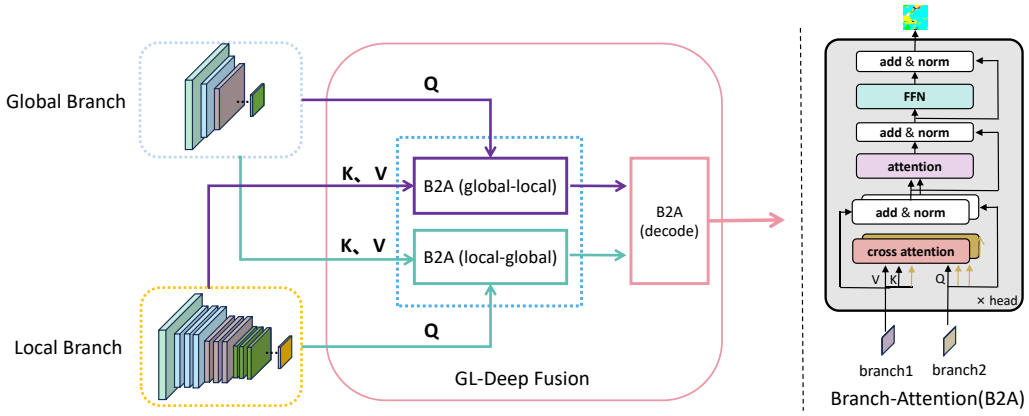


Figure 3. Overview of the GL-Deep Fusion module structure. As shown in the **left** figure, the core structure of GL-Deep Fusion includes a dual-cross encoder and a single decoder. As shown in the **right** figure, the two input sequences branch1 and branch2 correspond to the information sequences generated 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.

As shown in Figure 3, GL-Deep Fusion is based on the branch-attention and is designed with a dual-cross-encode and single-decode structure. Information sequences from both the global branch and the local branch are taken as input sequences directly, denoted as F_{global} and F_{local}. The design allows for a significant decrease in redundant computations, which enhances the effectiveness of the dual-branch structure's benefits. The first encoder generates a global–local sequence that encompasses the local relevance information. It can be represented as

$$B2A_{global-local} = Attention(F^{(q)}_{global}, F^{(k)}_{local}, F^{(v)}_{local}) \quad (3)$$

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

$$B2A_{local-global} = Attention(F^{(q)}_{local}, F^{(k)}_{global}, F^{(v)}_{global}) \quad (4)$$

Finally, the decoder merges the insights from both the global and local perspectives by employing a sophisticated B2A function, which treats the global–local and local–global sequences as distinct yet complementary branches. This integration is formulated as follows:

$$B2A_{decode} = B2A_{global-local,local-global} \quad (5)$$

This fusion process culminates in a rich set of features that capture the essence of both local details and global semantics. By deliberately excluding the FFN layer, our model streamlines the architecture, focusing on the efficient integration of features through the B2A mechanism, which is well-suited for the dual-branch structure of GLSNet. This decision reflects a strategic choice to prioritize memory efficiency without compromising the model's ability to understand complex scenes.

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 the presented study, the standard convolution-based ResNet [40] backbone network was utilized, which includes ResNet18 and ResNet50, with 18 and 50 layers, respectively (Table 1 for network structures). 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, the extraction of global contextual information is enabled, which encompasses a wide array of background environments and the semantic content 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.

Table 1. Architectures for ResNet18 and ResNet50 [40].

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 $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	28 × 28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
conv4_x	14 × 14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
conv5_x	7 × 7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	1 × 1	average pool, 1000-d fc, softmax	
FLOPs		1.8×10^9	3.8×10^9

4. Experiments and Results

4.1. Datasets

DeepGlobe [3]: The DeepGlobe dataset, known for its ultra-high resolution and challenging content, comprises 803 remote sensing images, each with a substantial pixel

dimension of 2448 × 2448. This dataset has been meticulously partitioned into training, validation, and test sets, containing 455, 207, and 142 images, respectively. Such a division facilitates efficient model training and evaluation and ensures the robustness of the model's performance when generalized across varied datasets. The diversity of land cover categories covered by DeepGlobe, including urban, agricultural, pastoral, forested, fire-stricken, waste, and unclassified areas, presents a comprehensive spectrum of seven distinct classes. These categories, in conjunction with the dataset's higher resolution compared to its predecessors, offer a more complex and realistic evaluation ground for the GLSNet model, particularly in handling fine details and diverse land cover types.

Vaihingen [19]: The Vaihingen dataset, with its 33 high-resolution images, each averaging 2494 × 2064 pixels and featuring a spatial resolution of 9 cm, provides a rich dataset for urban and natural environment analysis. The inclusion of red, green, and near-infrared (NIR) channels allows for the capture of a diverse range of visual and spatial attributes, which are crucial for accurate segmentation tasks. The six distinct categories represented in the Vaihingen dataset—impervious surfaces, buildings, low vegetation, trees, cars, and background—pose a demand for the model to exhibit high precision in distinguishing common urban and natural land cover types. The high spatial resolution and channel diversity of the Vaihingen dataset serves as an ideal testbed for evaluating the GLSNet's capability to accurately segment complex scenes with intricate details.

The characteristics of both the DeepGlobe and Vaihingen datasets, with their high resolution, diverse categories, and realistic scenarios, make them particularly suitable for evaluating the performance of GLSNet. These datasets not only challenge the model's ability to process and analyze large amounts of detailed spatial information but also verify its accuracy in segmenting various land cover types, which is essential for applications in autonomous driving, remote sensing, and urban planning.

4.2. Implementation Details

The model is optimized using the focal loss function [41], set with a weight of 1.0 and a γ value of 6, which serves as our primary objective to address class imbalance and focus on hard-to-classify examples. Similar to the methods employed in GLNet [20], we integrate two auxiliary losses and apply a regularization coefficient λ set to 0.15 to further refine the model's performance. As shown in Figure 4, the loss curves demonstrate a good fit, with both training and validation losses consistently decreasing and starting to plateau around the 20th epoch. This stabilization indicates that the model has effectively learned the underlying patterns in the data without overfitting, as evidenced by the convergence of the two curves to a similar minimum loss value.

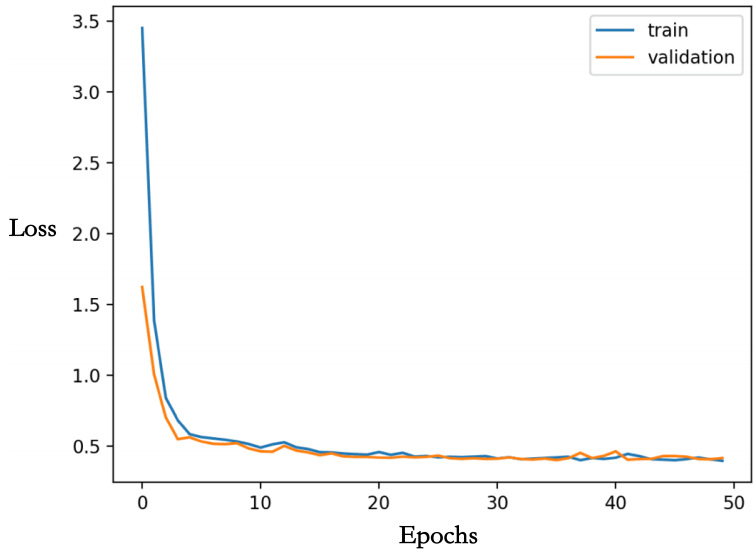


Figure 4. The training loss curve of GLSNet on the DeepGlobe dataset.

To effectively assess and optimize the graphics processing unit (GPU) utilization for our model, we have employed the "gpustat" command-line utility. This tool provides us with detailed insights into GPU usage, which is crucial for enhancing our model's computational efficiency. All training and testing are conducted on a single NVIDIA 1080Ti GPU card. This approach not only eliminates the need for gradient computation across multiple devices but also guarantees the replicability of our results. Moreover, to balance the GPU load and ensure stable training dynamics, we have established a batch size of six for all training iterations. Our experiments are executed within the PyTorch framework [42], chosen for its flexibility and powerful dynamic computational capabilities. For optimization, we have selected the Adam optimizer [43], renowned for its robust performance in handling sparse gradients.

Heeding research that demonstrates the benefits of assigning distinct learning rates to local and global branches for improved training outcomes, we have integrated this strategy into our model, GLNet [20]. Specifically, we have set the global branch learning rate to $\beta_1 = 1 \times 10^{-4}$ and the local branch learning rate to $\beta_2 = 2 \times 10^{-5}$. This differential rate allows each branch to learn at a pace that is well-matched to the nature of the data it processes.

4.3. Evaluation Metrics

The performance of the proposed GLSNet 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:

$$F_1 = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}, \beta = 1 \tag{6}$$

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:

$$IoU = \frac{TP}{FN + FP + TP} \tag{7}$$

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 comparing it with various models on the Vaihingen and DeepGlobe datasets.

4.4.1. Results and Analysis on the Vaihingen Dataset

In our meticulous evaluation on the Vaihingen dataset, the proposed GLSNet was quantitatively compared to several state-of-the-art approaches, including both transformer-integrated models such as BSNet [9], Mask2Former [23], and TransUNet [44], as well as traditional segmentation models like PSPNet [10], S-RA-FCN [45], DeepLabv3 [6], CCEM [46], RefineNet [13], EncNet [14], SegNet [11], UNet [8], and FCN-8s [45]. The assessment metrics included segmentation accuracy for large objects, small objects, and backgrounds, along with overall metrics such as mIoU, mean F1, and OA.

Table 2 presents a detailed comparison of these methods. GLSNet demonstrates superior performance across the overall metrics, with a particularly notable mean F1 score of 90.2%, an OA of 90.9%, and a mIoU of 90.9%. These results showcase the effectiveness of our model in segmenting various object categories with high precision. For instance, GLSNet outperformed BSNet by achieving a 2.3% improvement in accuracy for impervious surfaces and showed a significant enhancement of 3.7% for cars. It is also worth noting that

GLSNet's performance exceeds that of the popular Mask2Former method introduced in recent years. When compared with Mask2Former, which has gained significant attention for its innovative approach to segmentation, GLSNet demonstrates a higher mean F1 score and overall accuracy, indicating its robustness and potential for practical applications in scenarios requiring precise segmentation of ultra-high resolution images.

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 [45]	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 [44]	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**

However, we also observed that in scenarios with complex backgrounds or low-contrast objects, GLSNet's segmentation accuracy is slightly lower than in scenarios with clear object boundaries and distinct features. This observation indicates the necessity for enhancing the model's ability to adapt to challenging conditions. One potential avenue for improvement is expanding the diversity of the training data to include more examples of difficult-to-segment objects, which could help the model learn to better distinguish between similar textures and backgrounds.

In conclusion, the Vaihingen dataset evaluation highlights GLSNet's strengths in accurately segmenting a wide variety of objects. However, the analysis also reveals areas for improvement, particularly in the category of low vegetation, where GLSNet's performance was less pronounced compared to other object categories. This suggests that the model faces challenges in distinguishing objects with similar textural features from their surroundings. Addressing these challenges, such as enhancing the model's adaptability to complex segmentation tasks and optimizing memory usage without affecting accuracy, will be pivotal in guiding the future development of GLSNet. Our commitment to refining these aspects is aimed at achieving even higher standards of performance and efficiency.

4.4.2. Results and Analysis on the DeepGlobe Dataset

In our thorough evaluation on the DeepGlobe dataset, our proposed GLSNet was quantitatively compared against several state-of-the-art approaches, including TransUnet [44], Mask2Former [23], FCN-8s [45], DeepLabv3+ [7], SegNet [11], PSPNet [10], ICNet [12], UNet [8], and GLNet [20]. The comparison was not limited to segmentation accuracy (mIoU) but also encompassed the measurement of GPU memory usage (Memory).

As depicted in Table 3, all methods demonstrated improved mIoU results with the incorporation of a global branch, as opposed to relying solely on local patches. However, this enhancement came at the cost of a significant increase in GPU memory consumption. The majority of the methods struggled to effectively balance the segmentation accuracy with GPU memory usage. Among the listed approaches, only GLNet, which employs global–local information sharing, achieved a higher mIoU with reduced memory consumption, thus being selected 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	10,339
DeepLabv3+ [7]	63.1	1279	63.5	3199
FCN-8s [45]	64.3	1963	70.1	5227
Mask2Former [23]	66.7	3458	70.3	23,577
TransUnet [44]	68.2	2436	70.2	6283
	Local and Global			
	mIoU [%]		Memory [MB]	
GLNet [20] (baseline)	71.6		1865	
GLSNet	**72.4**		**1414**	

In contrast to the baseline model GLNet, the proposed GLSNet achieved significant breakthroughs in both mIoU and GPU memory usage. The mIoU score reached 72.4%, marking a 0.8% improvement over the baseline model. Most notably, the GPU memory usage was substantially decreased by 451 MB, a reduction of 24.1%. These advancements position GLSNet as a more advantageous model in terms of operational speed and resource utilization, offering enhanced potential for practical applications. We conducted an in-depth analysis to elucidate the performance differences between GLSNet and the baseline GLNet. The innovative dual-branch structure and the enhanced transformer attention mechanism of GLSNet allow for more efficient feature extraction and fusion, leading to superior segmentation accuracy while maintaining low memory usage. Specifically, the optimized architecture of GLSNet minimizes redundant computations and capitalizes on the strengths of both global and local information, resulting in a notable increase in segmentation performance.

Furthermore, we have considered the practical implications of our findings, recognizing that the balanced approach of GLSNet to accuracy and efficiency could be beneficial for real-world applications, especially in environments with constrained computational resources. The performance of GLSNet suggests that it may offer a viable solution for 3D object detection tasks that require a balance between precision and resource utilization. We believe that the results achieved by GLSNet contribute valuable insights for the ongoing research and may support the development of more effective 3D object detection systems in the future.

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.

Network Architectures	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 conducted ablation studies to evaluate the impact of different network architectures on mIoU, GPU memory usage, and the frames per second (FPS) metric, which provides additional insight into the models' performance. As shown in Table 5, the deep–deep model, enhanced by the GL-Deep Fusion strategy, achieved a 1%

higher mIoU than the GLNet baseline. This confirms that our fusion approach contributes positively to segmentation accuracy.

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

Network Architectures	mIoU [%]	Memory [MB]	FPS [f/s]
GLNet (baseline)	71.6	1865	0.05
Shallow–Deep	72.4	1414	1.14
Shallow–Shallow	71.9	**1044**	**1.34**
Deep–Deep	**72.6**	2903	0.50

Notably, the Shallow-Shallow model, while using less complex ResNet18 for both branches, not only reduced GPU memory usage significantly to 1044 MB but also improved the FPS to 1.34, demonstrating its efficiency. The Shallow-Deep model, our primary choice for GLSNet, offers a balanced performance with a mIoU of 72.4%, memory usage of 1414 MB, and an FPS of 1.14, highlighting a good trade-off between accuracy and computational speed.

4.5.2. The Effect of Transformer Attention

To assess the efficacy of diverse fusion module strategies on model performance, we undertook ablation studies utilizing the DeepGlobe dataset, with a continued focus on mIoU, GPU memory usage, and the frames per second (FPS) metric. Table 6 encapsulates the findings, which indicate that the GL-Deep Fusion module not only boosts mIoU by 1% and trims GPU memory usage by 96 MB when contrasted with Attention (DANet) but also realizes a noteworthy FPS of 1.14. This enhancement is markedly superior to the baseline GLNet's 0.05 FPS and Attention (DANet)'s 0.02 FPS, highlighting the transformer attention mechanism's role in achieving both efficiency and precision in segmentation. The FPS metric accentuates the model's practical viability, suggesting that the GL-Deep Fusion strategy is particularly fit for settings that necessitate swift processing capabilities.

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

Network Architectures	mIoU [%]	Memory [MB]	FPS [f/s]
GLNet (baseline) [20]	71.6	1865	0.05
Attention (DANet) [47]	71.4	1510	0.02
Attention (transformer) [39]	**72.4**	**1414**	**1.14**

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 5, 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.

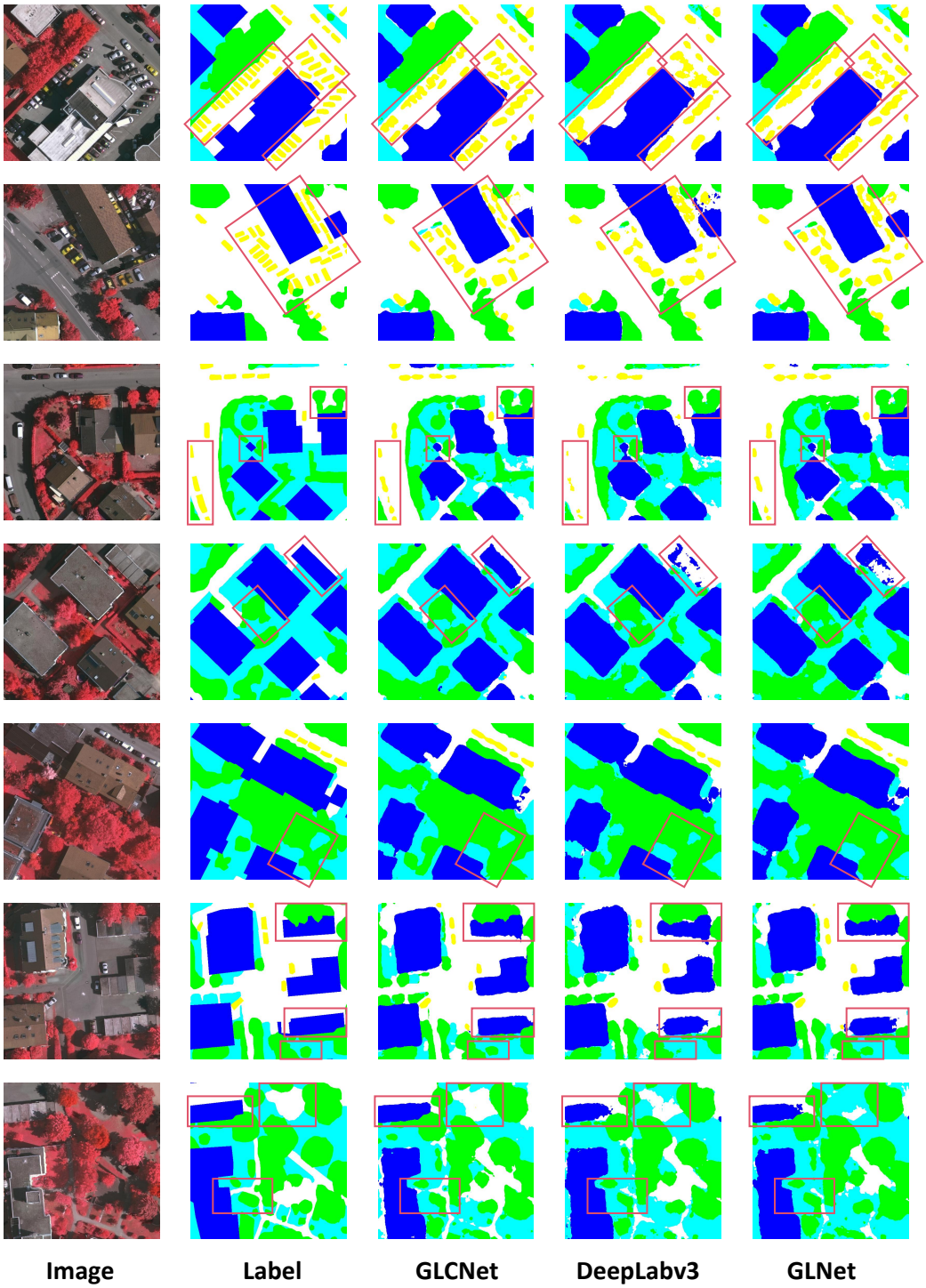


Figure 5. Comparison of segmentation results between GLSNet, DeepLabv3, and GLNet on the Vaihingen dataset. Note that the red boxes in the figure are mainly used to indicate parts where there are significant differences in segmentation results.

5. Discussion

The exceptional performance of our GLSNet method, as reflected in the comparative Table 7, is fundamentally attributed to the unique architectural design tailored specifically for ultra-high resolution image segmentation. This innovative framework forms the cornerstone of our success. The superior performance metrics observed when juxtaposed with other state-of-the-art methods substantiate the effectiveness of our approach in addressing the complexities of high-resolution imagery.

Table 7. The main contributions of those already reported in the state of the art.

Method	Core Contribution
FCN (Series) [31,45]	Introduced fully convolutional networks for semantic segmentation.
U-Net (Series) [8,25,32,33]	Utilizes an encoder–decoder architecture with skip connections to integrate features from various levels, particularly effective for medical image segmentation.
DeepLab (Series) [4–7]	Utilizes atrous convolution and pyramid pooling modules to effectively expand the receptive field and capture multi-scale contextual information.
GLNet [20]	a dual-branch network that leverages multi-level feature pyramid networks (FPNs) to exchange features between branches, improving feature utilization.
MaskFormer (Series) [22,23]	Introduces Transformer decoders and proposes a mask classification model that unifies semantic, instance, and panoptic segmentation tasks by predicting a set of binary masks.
Our Method	Utilizes a global shallow branch and a local deep branch in conjunction with GL-Deep Fusion based on branch-attention, achieving complete collaboration between the dual-branch structure, which is highly suitable for the field of ultra-high resolution image segmentation.

6. Conclusions

GLSNet, a segmentation model optimized for ultra-high resolution images prioritizing memory efficiency, has been presented. 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 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: C.L., K.H. and J.M.; data collection: C.L.; analysis and interpretation of results: C.L.; draft manuscript preparation: C.L., K.H. and J.M. All authors have read and agreed to the published version of the manuscript.

Funding: 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: The data are 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 Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542. [CrossRef] [PubMed]
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. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 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. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 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* **2014**, arXiv:1412.7062.
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 Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
6. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
7. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
8. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18 ; Springer: Berlin/Heidelberg, Germany, 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 Trans. Geosci. Remote Sens.* **2022**, *60*, 5624022. [CrossRef]
10. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
11. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
12. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 405–420.
13. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
14. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
15. Ascher, S.; Pincus, E. *The Filmmaker's Handbook: A Comprehensive Guide for the Digital Age*; Penguin: London, UK, 2007.
16. Lilly, P. Samsung launches insanely wide 32: 9 aspect ratio monitor with hdr and freesync 2. *PC Gamer*, 10 June 2017.
17. Initiatives, D.C. Digital Cinema System Specification, Version 1.3. 2018. Available online: http://dcimovies.com/specification/DCI%20DCSS%20Ver1-3%202018-0627.pdf (accessed on 11 July 2023).
18. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

19. ISPRS Vaihingen Dataset. Available online: https://paperswithcode.com/dataset/isprs-vaihingen (accessed on 15 September 2023).
20. Chen, W.; Jiang, Z.; Wang, Z.; Cui, K.; Qian, X. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8924–8933.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
22. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17864–17875.
23. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
24. AlMarzouqi, H.; Saoud, L.S. Semantic Labeling of High Resolution Images Using EfficientUNets and Transformers. *IEEE Trans. Geosci. Remote. Sens.* **2023**, *61*, 4402913. [CrossRef]
25. Zhu, F.; Zhu, Y.; Zhang, L.; Wu, C.; Fu, Y.; Li, M. A unified efficient pyramid transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2667–2677.
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. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 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 Trans. Cloud Comput.* **2020**, *10*, 2381–2397. [CrossRef]
28. Venkat, A.; Rusira, T.; Barik, R.; Hall, M.; Truong, L. SWIRL: High-performance many-core CPU code generation for deep neural networks. *Int. J. High Perform. Comput. Appl.* **2019**, *33*, 1275–1289. [CrossRef]
29. Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Adv. Neural Inf. Process. Syst.* **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. *Proc. Mach. Learn. Syst.* **2021**, *3*, 711–732.
31. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 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* **2017**, arXiv:1706.04284.
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 Trans. Image Process.* **2020**, *29*, 3695–3706. [CrossRef]
34. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
35. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
36. Poudel, R.P.; Bonde, U.; Liwicki, S.; Zach, C. Contextnet: Exploring context and detail for semantic segmentation in real-time. *arXiv* **2018**, arXiv:1805.04554.
37. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
38. Mazzini, D. Guided upsampling network for real-time semantic segmentation. *arXiv* **2018**, arXiv:1807.07466.
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
42. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
43. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
44. 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* **2021**, arXiv:2102.04306.
45. Mou, L.; Hua, Y.; Zhu, X.X. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12416–12425.

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 J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
47. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.