mathematics

MDPI

Article

Mask2Former with Improved Query for Semantic Segmentation in Remote-Sensing Images

Shichen Guo [1,2], Qi Yang [2,3], Shiming Xiang [2,3], Shuwen Wang [4] and Xuezhi Wang [1,*]

[1] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China; guoshichen@cnic.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China; yangqi2021@ia.ac.cn (Q.Y.); smxiang@nlpr.ia.ac.cn (S.X.)
[3] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[4] Department of Computer Science, Portland State University, Portland, OR 97201, USA; shuwen@pdx.edu
* Correspondence: wxz@cnic.cn

Abstract: Semantic segmentation of remote sensing (RS) images is vital in various practical applications, including urban construction planning, natural disaster monitoring, and land resources investigation. However, RS images are captured by airplanes or satellites at high altitudes and long distances, resulting in ground objects of the same category being scattered in various corners of the image. Moreover, objects of different sizes appear simultaneously in RS images. For example, some objects occupy a large area in urban scenes, while others only have small regions. Technically, the above two universal situations pose significant challenges to the segmentation with a high quality for RS images. Based on these observations, this paper proposes a Mask2Former with an improved query (IQ2Former) for this task. The fundamental motivation behind the IQ2Former is to enhance the capability of the query of Mask2Former by exploiting the characteristics of RS images well. First, we propose the Query Scenario Module (QSM), which aims to learn and group the queries from feature maps, allowing the selection of distinct scenarios such as the urban and rural areas, building clusters, and parking lots. Second, we design the query position module (QPM), which is developed to assign the image position information to each query without increasing the number of parameters, thereby enhancing the model's sensitivity to small targets in complex scenarios. Finally, we propose the query attention module (QAM), which is constructed to leverage the characteristics of query attention to extract valuable features from the preceding queries. Being positioned between the duplicated transformer decoder layers, QAM ensures the comprehensive utilization of the supervisory information and the exploitation of those fine-grained details. Architecturally, the QSM, QPM, and QAM as well as an end-to-end model are assembled to achieve high-quality semantic segmentation. In comparison to the classical or state-of-the-art models (FCN, PSPNet, DeepLabV3+, OCRNet, UPerNet, MaskFormer, Mask2Former), IQ2Former has demonstrated exceptional performance across three publicly challenging remote-sensing image datasets, 83.59 mIoU on the Vaihingen dataset, 87.89 mIoU on Potsdam dataset, and 56.31 mIoU on LoveDA dataset. Additionally, overall accuracy, ablation experiment, and visualization segmentation results all indicate IQ2Former validity.

Keywords: semantic segmentation; remote-sensing image; transformer; Mask2Former; query

MSC: 68T45

Citation: Guo, S.; Yang, Q.; Xiang, S.; Wang, S.; Wang, X. Mask2Former with Improved Query for Semantic Segmentation in Remote-Sensing Images. *Mathematics* **2024**, *12*, 765. https://doi.org/10.3390/math12050765

Academic Editor: Radu Tudor Ionescu

Received: 8 February 2024
Revised: 27 February 2024
Accepted: 29 February 2024
Published: 4 March 2024

1. Introduction

With the rapid development of remote-sensing technology, a large number of RS images are captured daily by satellites, airplanes, or drones. Understanding the content in RS images has become an increasingly urgent practical need. In the field of computer vision, natural images (e.g., COCO [1,2], ADE20K [3], Cityscapes [4] and Mapillary Vistas [5]) are

captured within a local region at or near the ground level for specific purposes, which render a visual center, and have spatially contiguous distributions for the same categories. In contrast, RS images (e.g., ISPRS [6] and LoveDA [7]) are taken from a high-altitude perspective, causing ground objects to be scattered in various corners of the image. In addition, objects of different sizes appear simultaneously in RS images. This situation can be witnessed in urban scenes where some occupy large areas, and some have only small regions. For instance, road surfaces usually occupy a large area, while cars only occupy a minimal space. Recently, the techniques for understanding natural images have become rich in the field of computer vision. However, practices have demonstrated that directly applying those existing models to complex RS images could not yield satisfactory results due to the significant difference between their visual appearances.

Semantic segmentation is a fundamental image understanding method to determine specific class labels at the pixel level. It is a delicate yet challenging task, especially for those sophisticated high-resolution RS images with rich ground details and many multi-scale objects. Recently, the techniques for this task have been primarily advanced by following those in the field of deep learning for natural images. Since the fully convolutional network [8] was first proposed, convolution has been used as the most common basic operation to construct the neuron layer for semantic segmentation models for quite a long time, such as the well-known PSPNet [9], UNet [10], and DeepLabv3+ [11], and so on. However, the emergence of vision transformer [12] has changed this paradigm. A large amount of transformer-based work has become the state-of-the-art semantic segmentation models, such as MaskFormer [13], and Mask2Former [14], and so on. Most notably, MaskFormer [13] rethinks the per-pixel classification as the mask classification with learnable queries facilitated by a DETR-like [15] architecture. Later, Mask2Former [14] has further incorporated together the masked attention mechanisms and multi-scale features, rendering a powerful capability for visual representation. This observation motivates us in this study to investigate its ability with an improved model to segment complex RS images.

More specifically, the purpose of this study is to design an improved model by exploiting the query of the Mask2Former [14] to capture well the characteristics of remote-sensing images for achieving higher segmentation performance. In this study, the query will be improved in three pathways to adapt to the RS scenes, which yields a new model given neural architectural design. First, we propose a query scenario module. Considering the complexity of RS scenes, such as those in the clusters of buildings, field landscapes, and road scenarios, it is intuitively suggested that different scenarios should be associated with various queries. Additionally, the substantial number of classes in natural scenes inevitably results in many learnable queries, which increases the computational load and the number of parameters in the model in the original Mask2Former [14]. Such an increase could be more significant in query due to the various scenarios of multi-scale objects scattered in the RS images. Therefore, we design the QSM to enhance the model's adaptability to distinguish various scenarios adaptively. Technically, this module can decrease the number of queries, thereby reducing the computational load and the number of parameters in the model. Simultaneously, it can select the queries suitable for different scenarios and thus help adjust the scene adaptability of the model.

Second, we introduce a query position module. The motivation for designing this module is to consider the variations in the spatial information of the targets in RS images. For instance, in natural scene images, cars and pedestrians are typically located in the bottom half of the image, while the sky and trees are generally found in the upper region. In contrast, in RS images, the target positions are not fixed; for example, the car may appear at any position in the image. Consequently, compared to natural images, the spatial information of the target holds greater significance for transformer-based models in remote-sensing scenarios. However, in conventional transformer-based models, only the simple encoding like the cosine position encoding is incorporated into the image features [12], while a learnable embedding is added later to the queries. For the queries to be learned, this could result in insufficient position awareness of the images. To this end, the QPM is

proposed to address this issue, which integrates the cosine position encoding of the input image feature together into the queries. This enhancement aims to increase the model's position sensitivity in RS images, thereby improving the segmentation performance.

Finally, we propose a query attention module since mining effective information from visual features, such as those in the channel or spatial domain, has been demonstrated to be a practical approach for addressing a variety of visual tasks [16–20]. Inspired by the ODConv [16], we believe that the model's performance can be enhanced by explicitly incorporating learnable feature-attention modules into the learnable queries of the transformer-based model [14]. Based on this justification, the QAM is developed to spatially position those discriminative features between the duplicate transformer decoder layers. As a result, QAM can ensure the comprehensive utilization of the supervisory information and the exploitation of those fine-grained details, and thus help extract the pertinent query features. From the model performance perspective, by introducing the QAM, our model can be trained to improve the representation quality of the learnable queries in the initial stages of the transformer decoder layers. In the subsequent transformer decoder layers, the QAM can effectively calibrate and accumulate pertinent information from queries, thereby enhancing the model's response to varying image features.

The key of our model lies in the innovation of the queries in the Mask2Former [14], taking into account the differences between RS scenes and natural ones. The main contributions and primary work can be summarized as follows:

- We introduce the query scenario module. Considering the diversity of scenarios and the finite categories within RS datasets, we adaptively select effective queries as the input for the transformer decoder layer. This approach aims to enhance the model's performance while reducing the number of model parameters and computational load.
- We introduce the query position module. Regarding the complexity of positions in remote-sensing images, we incorporate the position encoding of image features into queries. This strategy is intended to further enhance the model's capacity to perceive targets.
- We propose the query attention module. We incorporate attention modules between duplicate transformer decoder layers to better mine valuable information from learnable queries. This approach is specifically designed to augment the extraction of valid query features.
- The performance of IQ2Former for segmenting RS images has been assessed on three challenging public datasets, including the Vaihingen dataset, the Potsdam dataset, and the LoveDA dataset. The comprehensive experimental results and ablation studies demonstrate the effectiveness of the proposed model, including numerical scores and visual segmentation.

The remaining chapters are structured as follows: Section 1 describes the background information, motivation, objectives, and hypotheses of this study. Section 2 reviews the related works. Details of the proposed method are provided in Section 3. Experimental results are presented in Section 4. Discussions can be found in Section 5, followed by conclusions in Section 6.

2. Related work

2.1. Transformer for Semantic Segmentation

Semantic segmentation is a fundamental computer vision task that performs pixel-level classification. Currently, there are two mainstream mechanisms for semantic segmentation: CNN-based and Transformer-based semantic segmentation. In the early research on semantic segmentation with the CNN-based mechanism, researchers took a technical roadmap along the fully convolutional networks (FCNs) [8] as the dominant approach and focused on aggregating the long-range context in the feature map. This mechanism has enriched the methods for image semantic segmentation. Some famous models were developed, including the DeconvNet [21], RefineNet [22], PSPNet [9], UNet [10], OCRNet [23], and DeepLabv3+ [11], HRNet series [24–26], and so on. These models were all developed based on the encod-

ing and decoding frameworks with convolution operations. Subsequently, image semantic segmentation based on these frameworks has been greatly developed and widely used in many applications.

With the proposal of the transformer [27], the transformer-based model became the mainstream mechanism of deep learning technology. As for semantic segmentation, mainstream methods [28–30] gradually replaced those traditional CNN-based backbones with transformer-based architectures. Segmenter [28] was the first architecture to extend the vision transformer [12] into the semantic segmentation task. Later, SETR (segmentation transformer) [30] utilized the self-attention mechanism of the transformer to establish global contextual relationships. SegFormer [29] employs the transformer to construct the encoder to enhance the feature representation ability for semantic segmentation. As a whole, the transformer employs a multi-layer perceptron to aggregate information from different layers, which renders a new mechanism that fully utilizes global and local attention to increase the ability of representation learning. However, the above methods only focus on replacing the backbone with the transformer-based architecture and lack the targeted research on segmentation task with the transformer-based mechanism.

Interestingly, MaskFormer [13] and Mask2Former [14] present a mask classifier with learnable queries and specialized designs for mask prediction under transformer-based semantic segmentation. Later, based on the Mask2Former [14], OneFormer [31] presented a universal image segmentation framework that unified segmentation with a multi-task train-once design. Recently, Segment Anything (SAM) [32] first explored transformer-based interactive open-world segmentation. Semantic SAM [33] further extended SAM into open vocabulary segmentation, and Faster Segment Anything [34] applied the SAM to mobile applications by replacing the heavyweight transformer encoder with a lightweight one. However, the above methods only focus on natural scenes, lack consideration for RS scenes, and cannot effectively perform segmentation in specific RS images. Given that semantic segmentation is one of the essential sub-tasks for RS images, these methodologies have significantly contributed to our work.

2.2. Semantic Segmentation for RS Images

For ordinary natural images taken at or near the ground level, researchers have devoted tremendous efforts and achieved remarkable success in semantic segmentation. While for RS-type semantic segmentation, there is still much worthy of research efforts to achieve higher superior performance to RS images [35].

Regarding chronological order, the first surging method is based on CNN models. FCN [8] is the pioneering work in semantic segmentation in the era of deep learning, followed by [9–11,23,36]. With the usage of encoder/decoder backbones, some variants have been constructed for this issue [37–41]. For example, the cascaded network with context information fusion is developed to extract confusing human-made objects [40], and the shuffling network is employed to enhance the feature learning ability [38]. Later, Guo et al. [41] adopted a learnable gate mechanism during feature fusion, further improving the fitting ability of the model. Additionally, a multi-level aggregation network [42] extracts deep global features by learning the inter-relationships of all positions in the context and filters redundant channel information as well enhances the model's ability to recover detailed information. These research studies have primarily enhanced the performance of the semantic segmentation for RS images.

Subsequently, more complex models have been considered for segmenting RS images. Specifically, Diakogiannis et al. [43] developed an encoder–decoder method with multi-tasking inference sequentially on object boundary, segmentation masks, and the reconstruction of the input. Zhang et al. [44] employed the high-resolution network with different branches to extract features at both the local and global levels. Xu et al. [45] constructed a high-resolution context extraction network to fuse multi-scale contextual information. Later, attention modules in different views [46–48] were designed for fine

segmentation. These models achieved good segmentation in different ways for local, global, and multi-scale feature fusion.

There are a few works on the semantic segmentation of RS images under the neural architecture search (NAS) frameworks in the literature. Typically, Zhang et al. [49] employed a directed acyclic graph with tricks of Gumbel-max operations under a differentiable searching framework. RS images include various scenes such as cities, rural areas, urban villages, farms, etc., which are suitable for domain adaptation models to save annotation resources. Accordingly, Shi et al. [50] tackled the domain shift problem by employing adversarial learning to tune the semantic segmentation network to obtain similar outputs for input images from different domains adaptively. Later, Wang et al. [51] proposed the decoupling NAS framework with a hierarchical search space at the path level, connection level, and cell level for RS objects. Broni-Bediako et al. [52] developed an evolutionary NAS method for this task. In their framework, gene expression programming and cellular encoding are employed to represent the encoding scheme for block-building. More recently, Guo et al. [26] used HRNet [24] as the supernet and accelerated proximal optimization algorithm in the search to obtain a better-performing pruning model. However, although these approaches achieved a good performance, high computational complexity degrades their real-world applications.

Recently, the models based on the transformer have also achieved great success in RS images. Wang et al. [53] introduced the Swin Transformer [54] as the backbone to extract the context information and design a novel decoder of the densely connected feature aggregation module to restore the resolution and produce the segmentation map. Subsequently, the transformer was also employed as the backbone in this task [53,55]. Later, Ye et al. [56] constructed modules to segment the different scales of RS objects with the transformer and multi-scale feature representation.

Technically, the work that combines CNN and transformer can use both to complement each other's strengths and weaknesses. He et al. [57] embedded the Swin Transformer [54] into the classical CNN-based UNet [10]. The encoder of [58] was used to extract features to achieve a better long-range spatial dependencies modeling, and the decoder was used to draw on some effective blocks and successful strategies of CNN-based models in RS image segmentation.

In summary, architecturally, almost all of the methods developed for RS images inherit the general framework of semantic segmentation for natural images by modifying or adding some modules that can utilize the knowledge about the RS scenes or the objects themselves. Although these semantic segmentation approaches for natural images have achieved a good performance in terms of accuracy, directly applying them to RS images could not yield satisfactory results. This is largely due to the complexity of the RS images, which contain multi-scale objects with different visual appearances and spatial resolutions.

3. Methods

The section will first revisit the procedure of MaskFormer [13] and the core innovation of Mask2Former [14] in Section 3.1. Then, the overall framework of our methods will be presented in Section 3.2, which is explicitly designed for semantic segmentation of RS images. Architecturally, our model comprises a query scenario module, a query position module, and a query attention module, which will be introduced, respectively, in Sections 3.2.1–3.2.3.

3.1. Preliminary

Transformer [12] has demonstrated a powerful learning ability for visual representation. Some variants of the transformer have been developed to perform specific visual tasks [13,29,30,54,59]. Here, we briefly introduce the highly realized MaskFomer and Mask2Former for clarity.

Rather than formulating semantic segmentation as a per-pixel classification task, MaskFormer [13] predicts a set of binary masks, each corresponding to a single global class label prediction. Under this framework, the query in MaskFormer [13] plays a crucial role

in successfully segmenting images, ensuring that the order of binary masks is always the same as the order of the predicted class results. As a result, the queries are also associated with the intermediate results, which are then taken as the input of the next decoder layer in the transformer hierarchically for further abstract representation.

A recent study [59] has shown that cross attention could require much training time to learn to pay attention to local object regions. To remedy this drawback, Mask2Former [14] was developed in terms of masked attention. Methodologically, it is actually a variant of cross-attention that only focuses on the foreground region of the predicted mask for each query. Mathematically, the masked attention is achieved by performing the following operation:

$$\text{Masked-Attention}(Q, K, V) = \text{softmax}\left(\mathcal{M} + \frac{Q\,K^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where $Q \in \mathbb{R}^{N\times D}$ refers to N D-dimensional query features. $K, V \in \mathbb{R}^{HW\times d_k}$ represents HW d_k-dimensional keys and value features, respectively. In Equation (1), H and W denoted the spatial resolution of the image features. The mask $\mathcal{M}$ at the feature location (x, y) is then determined as follows:

$$\mathcal{M}(x,y) = \begin{cases} 0, & if\ \mathcal{M}(x,y) = 1 \\ -\infty, & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathcal{M} \in \{0,1\}^{N\times H_l W_l}$ represents the binarized output of the resized mask prediction of the previous transformer decoder layer. It is worth noting that Equation (1) turns out to be the standard cross attention when omitting $\mathcal{M}$.

3.2. Overview of Our Method

The overall framework of our model is depicted in Figure 1, which is named IQ2Former for convenience. Architecturally, it inherits the encoder/decoder meta-architecture in the Mask2Former [14]. In Mask2Former [14], the image first passes through the image encoder and pixel decoder in sequence, and then the first three feature maps generated by the pixel decoder are sent to the transformer decoder to generate the mask and class corresponding to each query. Then, the last feature maps of the mask and pixel decoder are multiplied to obtain the foreground feature maps, and each foreground feature map and class are multiplied to obtain the final segmentation result. With this backbone, we propose three improvements to the query to achieve higher performance in RS images.

First, to improve the adaptability of our model to various RS scenes, the query scenario module is designed in Section 3.2.1. This module aims to select an appropriate number of queries and perceive several common scenarios for RS images. Second, to intensify the sensitivity of our model to the positions of the objects in RS images, the query position module was developed in Section 3.2.2. This module integrates together the cosine position encoding and the input image feature into the queries. Finally, to further enhance the ability of our model to extract the valid features of the queries, the query attention module is proposed in Section 3.2.3. Being positioned between the duplicate transformer decoder layers, this module aimed to help strengthen the extraction of the pertinent query features and weaken the impact of the redundant features.

As a whole, the three queries obtained from the above three independent modules are formally named $Query^*_{QSM}$, $Query^*_{QPM}$, and $Query^*_{QAM}$ in the following subsections. Because all improvements are based on queries, our model is named IQ2Former, and the abbreviation of Mask2Former with improved query.

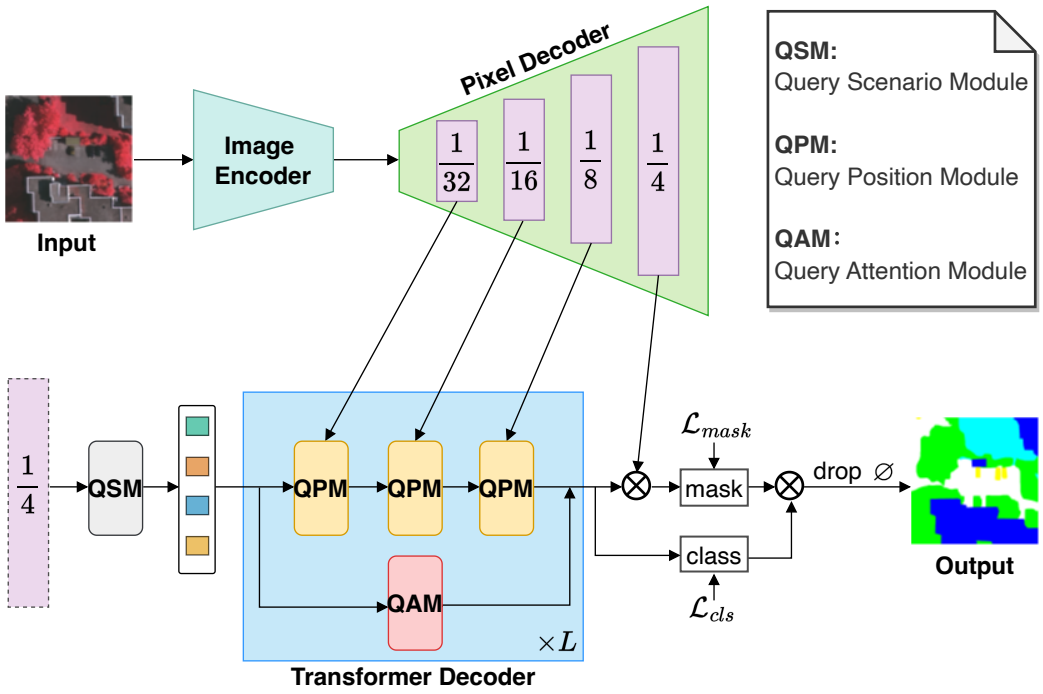


Figure 1. The overview of the IQ2Former. The top-right corner is the legend, which shows the full names of our three improvement modules.

3.2.1. Query Scenario Module

In contrast to traditional transformer models such as SegFormer [29], the first query in the MaskFormer [13,14] is randomly initialized and not derived from image features. To remedy this problem, we take one layer feature of the pixel decoder as the input of the query scenario module to utilize the image's information fully. As illustrated in Figure 1, assuming $X \in \mathbb{R}^{C \times H \times W}$ is the input image, we can obtain the consecutive four output features $Feature_i$, $i = 1, 2, 3, 4$ in the pixel decoder. The size of the feature $Feature_i$ is $C_i \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$, where C_i represents the dimensions of the i-th stage output features. As shown in Figure 2, the last layer $Feature_1 \in \mathbb{R}^{C_1 \times \frac{H}{4} \times \frac{W}{4}}$ with high resolution is employed as the input of the QSM since it contains the most image detail information in pixel decoder.

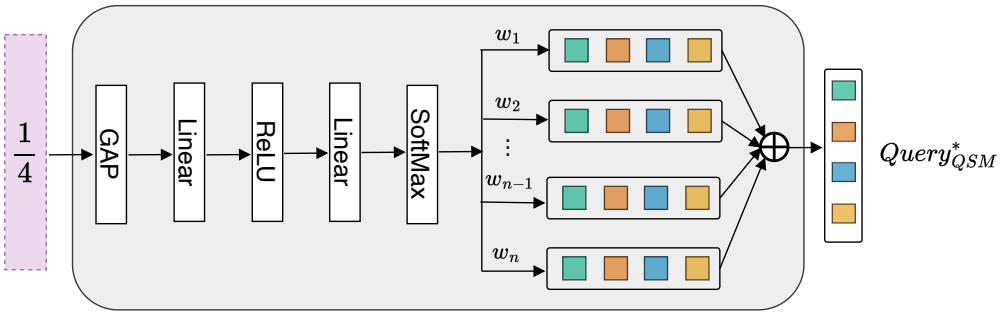


Figure 2. The details of the query scenario module. The block marked as 1/4 is the last layer from the pixel decoder.

For the natural image database ADE20K-Full [3], where there are a total of 847 classes, as demonstrated in Mask2Former [14], each query needs to memorize 8.47 classes on average. The substantial number of classes in natural scenes inevitably leads to a large number of learnable queries, thereby increasing the computational load and the number of parameters. On the contrary, remote scenes typically have fewer classes. Theoretically, this allows us to employ fewer queries for computation reduction. However, using fewer

queries may directly result in a decrease in performance, as demonstrated in [14]. Based on the observation of different kinds of datasets and the analysis of the trial results mentioned above, we innovatively split a large number of queries in Mask2Former [14] into several groups of a small number of queries. As shown in Figure 2, the operation can be formatted as follows:

$$\boldsymbol{w} = \text{SoftMax}(\text{Linear}_2(\text{ReLU}(\text{Linear}_1(\text{GAP}(Feature_1))))), \tag{3}$$

where GAP stands for global average pooling, and is responsible for transforming $Feature_1$ from a spatial tensor to a vector with a length equal into the number of channels C_1. Assuming the output dimension of Linear_2 is n, we can obtain the selection weights $\boldsymbol{w} = \{w_1, w_2, \ldots, w_n\}$ ($\sum_{i=1}^{n} w_i = 1$) through the SoftMax function, and then weight learnable sub-queries $\boldsymbol{q} = \{q_1, q_2, \ldots q_n\}$ with fewer numbers using $\boldsymbol{w}$. The ultimate query is expressed by:

$$Query^*_{QSM} = \sum_{i=1}^{n} \boldsymbol{wq} = w_1 q_1 + w_2 q_2 + \ldots + w_n q_n, \tag{4}$$

where $q_i \in \mathbb{R}^{N \times D}$, $i \in 1, 2, \ldots, n$. Here, N refers to the number of each sub-query, and D represents the dimension of each query. The final query obtained by the query scenario module is $Query^*_{QSM}$.

It should be emphasized that the strategy of the query grouping is only performed in the first round of querying. Each remaining transformer decoder adopts the same query passing strategy of the Mask2Former [14], but the query length is still N. In summary, our QSM can technically adapt to various scenarios with fewer queries according to the characteristics of different RS scenarios. The discussion in Section 5 also confirms our hypothesis in detail.

3.2.2. Query Position Module

Natural images are captured using vanilla mobile devices for specific purposes, usually with a visual center, and the distribution of the same categories in the images is often contiguous. On the contrary, RS satellites capture all the information on ground appearance from a high-altitude perspective, causing ground objects to be scattered in various corners of RS images.

Compared with natural images, significant differences in target positions in RS scenes can increase the difficulty of semantic segmentation. Therefore, we construct the query position module to strengthen the IQ2Former's sensitivity to the position of the target in the remote scene. This is achieved by incorporating an additional image and spatial position information into the learnable query, as depicted in Figure 3. Specifically, we define f_Q, f_K, and f_V as a linear transformation of the given queries $Query$ and the visual features $Feature_i$, $i = 2, 3, 4$ outputted by the pixel decoder. Consequently, $Q = f_Q(Query)$, $K = f_K(Feature_i)$, and $V = f_V(Feature_i)$, $i = 2, 3, 4$, represent the query, key, and value features, respectively.

To further incorporate additional image features into the learnable query, we first obtain the mixed intermediate features, which contain the image features. This is achieved through the dot product of the query feature Q and key feature K. The formula is presented as follows:

$$W_{attention} = QK^T = f_Q(Query_b + POS_{Query_b}) \cdot f_K(Feature_i + POS_{Feature})^T, \quad i = 2, 3, 4, \tag{5}$$

where POS_{Query_b} represents a learnable query position embedding with the same dimensions as $Query_b$, and $POS_{Feature}$ denotes a cosine position encoding with the same dimensions as $Feature_i$. The intermediate attention weight of the mask-attention is expressed as $W_{attention} \in \mathbb{R}^{N \times \frac{HW}{2^{2i+2}}}$, where N is the number of queries.

Additionally, to inject the cosine position encoding into the learnable query, we then multiply the position encoding of the input image features with the attention weight $W_{attention}$. The ultimate query can be represented as follows:

$$Query^{*}_{QPM} = W_{attention} \cdot POS_{Feature} + Query_{b+1}, \quad b = 1, 2, \ldots, B-1, \tag{6}$$

where B is the total number of blocks in the pixel decoder. The final query obtained by the query position module is $Query^{*}_{QPM}$.

The QPM was subtly conceived, where $W_{attention}$ records patches that each query is interested in. After multiplying by $POS_{Feature}$, the position encoding of the patch of interest in each query is recorded. By coincidence, the shape of the final product is the same as the query, which is the basis for the addition of Equation (6).

As can be seen from the formulation in Equation (6), the position information is explicitly embedded into the learning process. In this way, the location diversity of the target in RS scenes could help improve segmentations with high quality. As a result, after the above QPM is performed, the query learned by the IQ2Former could be more sensitive to the mask features in different locations.

Notably, $W_{attention}$ is a part of Equation (1) that already exists, and $POS_{Feature}$ is also a variable that was already used. Therefore, there is no increase in the number of parameters in our QPM, and the convergence speed of the mask attention was unaffected.

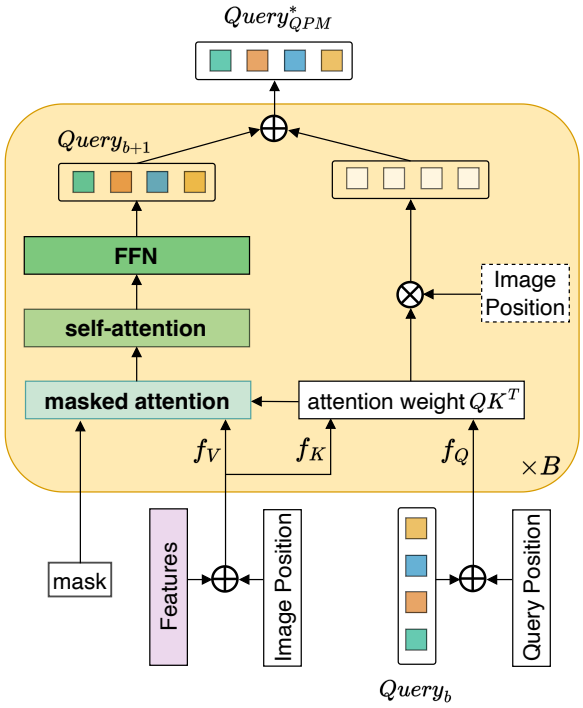


Figure 3. The details of query position module. The image position in the dashed box is identical to that in the solid box, drawn this way to minimize the mess caused by overlapping lines.

3.2.3. Query Attention Module

Enormous and tiny objects appear simultaneously in RS images, which poses a challenge for the segmentation of distinct objects. For example, road surfaces usually occupy a large area, while cars only occupy a minimal region. Therefore, low-level image features that contain rich detail information are actually crucial for fulfilling the segmentation of RS images. However, Mask2Former [14] feeds the successive feature maps from the pixel decoder directly into the transformer decoder in a round-robin fashion. For example, in the original Mask2Former, query learning in the transformer–decoder layer takes the visual features with a resolution of 1/8 as its last input features in the current transformer–decoder layer while employing those with 1/32 as the first input features in the next transformer–decoder layer. Thus, the large-scale span in visual information between the repeated transformer–decoder layers could lead to difficulty in capturing the details of the visual information. Furthermore, previous research [13] has demonstrated that a single-layer

transformer decoder can also deliver competitive performance. This fact indicates that it is unnecessary to design highly hierarchical layers to fulfill this task.

Based on the above reasons, we introduce the query attention module, designed to augment the model's capacity to extract valid features from the queries while weakening the influence of superfluous features. Specifically, as demonstrated in Figure 1, the QAM will be positioned parallel between the duplicated transformer–decoder layers to help extract more fine-grained visual features. As a result, the information transformation is always kept within the features of high resolution without being transformed from low resolution to high resolution.

However, traditional convolution is not appropriate to process queries. In addition, due to the different heights and widths of images, the two dimensions of the query are totally different in essence. Drawing inspiration from the ODconv [16], our QAM introduces a multi-dimensional attention mechanism, employing a parallel strategy to learn diverse attentions for the learnable queries from different dimensions. Furthermore, the sigmoid operation is capable of ensuring that the weight assigned to the previous transformer decoder layer is between 0 and 1, which means that the current transformer decoder layer dominates more in the fused query. As illustrated in Figure 4, the weight along each direction is computed by GAP + Linear + ReLU + Linear + Sigmoid operations. Considering that learnable queries $Query \in \mathbb{R}^{N \times D}$ have two dimensions, the query fusion is then designed as follows:

$$\begin{aligned} Query^{*}_{QAM} = &(\alpha_1 Query_l^{1,:} + \alpha_2 Query_l^{2,:} + \ldots + \alpha_N Query_l^{N,:}) \quad \odot \\ &(\beta_1 Query_l^{:,1} + \beta_2 Query_l^{:,2} + \ldots + \beta_D Query_l^{:,D}) + Query_{l+1}, \\ &l = 1, 2, \ldots, L-1, \end{aligned} \tag{7}$$

where $\odot$ denotes the multiplication operations along different dimensions of the $Query$. Two dimensions of $Query_l^{i,j}$ are dimension $i \in [1, N]$ and dimension $j \in [1, D]$. The attentions introduced in Equation (7) are $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]$ and $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_D]$. In addition, L is the total number of layers in the transformer–decoder. Thus, in this way, the final query obtained by the QAM can be taken as $Query^{*}_{QAM}$ for further query fusion.

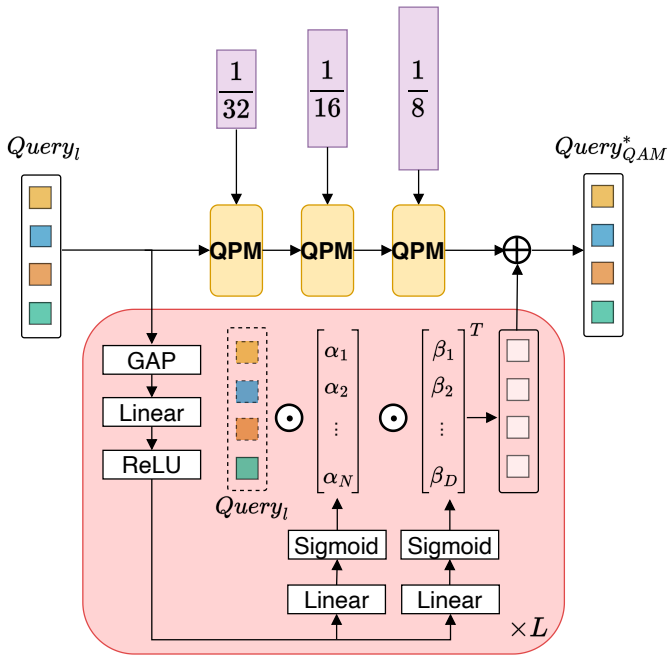


Figure 4. The details of the query attention module. The query in the dashed box is identical to $Query_l$ in the solid box, drawn this way to minimize the mess caused by overlapping lines.

As can be seen from the formulation in Equation (7), the QAM could mine effective information from the latent visual features, such as those in the channel or spatial domain. Therefore, our QAM can ensure the comprehensive utilization of the supervisory information and the exploitation of those fine-grained details, and thus help extract the pertinent query features for a higher performance. Furthermore, we take zero initialization for the linear in QAM inspired by ControlNet [60]. To this end, our QAM can identify the relevant features within the query in a gradual process without extra supervisory information.

4. Experiment

4.1. Data Description

The performance of the proposed IQ2Former has been evaluated on three public challenging benchmark datasets, which all focus on the semantic segmentation of remote sensing images. The details about these datasets are described in Table 1. The Vaihingen [6] and Potsdam [6] datasets ceased updating in 2018, forming the appearance we see now. The paper of LoveDA dataset [7] was published in 2021. The ground truth of the Vaihingen [6] and Potsdam [6] datasets encompasses six categories: impervious surface, building, low vegetation, tree, car, and cluster/background. We omit the segmentation results of meaningless cluster/background. The ground truth of the LoveDA dataset consists of seven categories: building, road, water, barren, forest, agriculture, and background.

Table 1. Comparison of the dataset used. # represents numbers.

Dataset	Year	GSD	DSM	# Classes	# Images	# Pixels	# Croped Pixels	# Traning	# Testing
Vaihingen [6]	2018	9 cm	71.02	5	33	2494×2064	512×512	344	398
Potsdam [6]	2018	5 cm	71.48	5	38	6000×6000	512×512	3456	2016
LoveDA [7]	2021	30 cm	78.23	7	5987	1024×1024	1024×1024	2522	1669

4.2. Baseline Model Description

- **FCN**: Fully convolutional network [8] is the pioneering work in semantic segmentation in the deep learning era. FCN replaces fully connected layers with convolutional layers, enabling the network to process input images of arbitrary sizes and generate output with the exact spatial dimensions. Nowadays, FCN is the most crucial baseline in semantic segmentation tasks.
- **PSPNet**: Pyramid scene parsing network [9] utilizes a pyramid pooling module that gathers contextual information from the diverse areas of an image, enabling the network to have a holistic understanding of the scene. This module effectively captures both local and global contexts by hierarchically partitioning the input feature map and performing spatial pyramid pooling operations.
- **DeepLabV3+**: DeepLabv3+ [11] uses dilated convolution to effectively enlarge the receptive field of filters, allowing the network to capture more contextual information. Additionally, the feature pyramid network [61] is introduced to combine features at various spatial resolutions. DeepLabV3+ was almost the most advanced algorithm in semantic segmentation before the advent of transformer.
- **OCRNet**: Object-contextual representations network [23] obtains coarse segmentation results from a general backbone and object region representation from gathering pixel embeddings in it. The second step is computing the relationship between each pixel and each object region. The final acceptable segmentation result was obtained by enhancing the expression of each pixel with the object-contextual representation.
- **UPerNet**: Unified perceptual parsing network [62] mimics the human recognition of multiple levels of the visual world and unifies the datasets containing various scenes, objects, parts, materials, and textures. Using a feature pyramid network [61] and different detection heads, UPerNet can be applied to multi-task learning in addition to semantic segmentation.
- **MaskFormer**: Rather than predicting the class of each pixel point, MaskFormer [13] is first proposed to predict a set of binary masks associated with a single global class label

prediction. Under the supervision of both mask loss and category loss, MaskFormer has achieved excellent performance in both semantic and panoptic segmentation tasks under the supervision of both mask and category loss.

- **Mask2Former**: The critical component of Mask2Former [14] is masked attention, which extracts local features by confining cross-attention within the predicted mask regions. In this way, the research effort is reduced by at least three times while improving performance by a significant margin. Mask2Former is capable of addressing all image segmentation tasks, including the panoptic, instance, or semantic ones.

4.3. Experiment Details

Backbone: Our IQ2Fomer is compatible with any backbone architecture. For a fair comparison, the standard convolution-based ResNet with 50 layers, ResNet with 101 layers, and the transformer-based Swin-transformer (Swin-L) are used as our visual backbone. All backbones are pre-trained on the ImageNet-1K [63] if not stated otherwise.

Pixel decoder: As shown in Figure 1, four different resolution outputs of the pixel decoder are expressed as F_i, where $i = 1, 2, 3, 4$. They are feature maps with resolutions $1/32$, $1/16$, $1/8$, and $1/4$, respectively, in our experiments. Similarly to Mask2Former [14], the same multi-scale deformable attention transformer (MSDeformAttn) [64] is utilized as our pixel decoder.

Transformer decoder: Totally, there are three consecutive transformer decoders in our IQ2Former (i.e., nine layers as a whole). the QSM is the input of the first transformer decoder of our IQ2Former. Note that the number of queries in the Vaihingen dataset and Potsdam dataset is set to 20, while the number of queries in the LovaDA dataset is taken as 40. The reason for this phenomenon is discussed in Section 5. Each transformer decoder contains three QPMs and one QAM. Among them, the role of the QPM is to enhance the query's perception ability of image position. An auxiliary loss is added to every intermediate transformer decoder layer, and a competitive query can be obtained. Therefore, the proposed QAM is used to explore the query output capability of the previous layer of the transformer decoder further.

Losses: The comprehensive training loss has two components: the classification loss and the mask loss. The classification loss is formulated by a cross-entropy loss, denoted by $L_{cls} = L_{ce}$. The mask loss integrates the binary cross-entropy loss and the dice loss [65], which is depicted as $L_{mask} = L_{bce} + L_{dice}$ for clarity. The overall training loss can be expressed as $L = \lambda_{cls} L_{cls} + \lambda_{mask} L_{mask}$, where λ_{cls} and λ_{mask} are the hyper-parameters. According to the default settings of Mask2Former [14], in our study, we also set the balance weight of the overall training loss to be $\lambda_{cls} = 2.0$ and $\lambda_{mask} = 5.0$, respectively.

Inference: Assuming that $O^{cls} \in \mathbb{R}^{N \times (K+1)}$ and $O^{mask} \in \mathbb{R}^{N \times H \times W}$ are predicted per-pixel embeddings and binary masks, respectively. Here, K represents the total number of object classes, and N is the number of object queries. $O = O^{cls} \times O^{mask}$, $O \in \mathbb{R}^{(K+1) \times H \times W}$ performs matrix multiplication and sums on the dimension of the query. In this way, the dimension of the query is eliminated, and a probability distribution is obtained for each pixel of the output feature. The ultimate segmentation results are $argmax\ O$, without considering the no-object class $\varnothing$.

Batch size and learning rate: In our experiments, the batch size is set to 8, and all models are trained for 80 k iterations. The AdamW [66] and the poly [67] learning rate schedule with an initial learning rate of e^{-4} and a weight decay of 0.05 were adopted for both ResNet and Swin-transformer backbones.

Data augmentation: During the stage of training, random scale jittering (between 0.5 and 2.0), random horizontal flipping, random cropping, as well as random color jittering are used to perform data augmentation. During the stage of testing, test-time augmentation (TTA) was utilized in our experiments. Specifically, it creates multiple augmented copies for each image in the test set, allowing the model to make predictions for each image, and then returns the set of these predictions and the final results with the highest number of votes. Random flip and multi-scale testing are adopted in this paper.

Metric: IoU (intersection over union) is the quotient of the intersection and union between the predicted segmentation and annotation regions. **mIoU** (i.e., mean intersection over union) is the mean IoU of all classes. **OA** (overall accuracy, also known as pixel accuracy) is all correctly classified pixels divided by all pixels.

Environment: All models are trained with four A100-PCIE graphics cards with a memory capacity of 40 GB. The conda environmental configuration is as follows: Python 3.8.17, NumPy 1.24.3, PyTorch 2.0.0, TorchVision 0.15.0, CUDA version 11.7, MMCV [68] 2.0.1, and MMSegmentation [69] 1.1.0.

4.4. Experiment Results

To provide a thorough comparison of the models, we list the numerical scores of the OA and mIoU obtained by the eight models on the Vaihingen, Potsdam, and LoveDA datasets in Tables 2–4, respectively. All the values are obtained from the corresponding test images and are then averaged across all categories. Note that the two methods are adopted to output the final performance. The first method is to directly calculate OA and mIoU on the images in the test dataset. The second method is to calculate OA and mIoU via test-time augmentation, which augments multiple copies for each testing image by randomly flipping and resizing.

As can be seen from the results of the comparisons in these tables, our IQ2Former model outperforms other baseline models to a large extent, both in the ResNet [70] and Swin [54] backbones.

Table 2. Results of quantitative comparison on the Vaihingen testing set. The numbers are the percent scores (%). † indicates that the scores are acquired via the flip and MS testing. Bold font means the highest performance of that class.

Backbone	Models	Imp. Surf.	Building	Low Veg.	Tree	Car	OA	mIOU	OA †	mIOU †
ResNet-50	MaskFormer [13]	84.67	90.64	70.01	79.37	69.52	89.50	78.84	90.29	80.82
	Mask2Former [14]	85.49	90.69	**71.54**	**80.08**	73.49	90.01	80.26	**90.86**	82.05
	IQ2Former(ours)	**85.92**	**92.00**	71.24	79.99	**77.32**	**90.22**	**81.29**	90.76	**82.23**
ResNet-101	FCN [8]	84.99	91.31	70.31	79.57	76.18	89.76	80.47	90.34	81.87
	PSPNet [9]	85.83	91.49	71.37	79.90	75.14	90.16	80.75	90.76	82.39
	DeepLabV3+ [11]	86.20	91.61	71.43	79.74	75.18	90.23	80.83	90.81	82.19
	OCRNet [23]	84.70	90.57	69.74	78.83	66.71	89.36	78.11	90.23	79.99
	UPerNet [62]	85.81	91.66	70.94	79.96	76.46	90.14	80.97	90.79	82.48
	MaskFormer [13]	85.38	91.26	71.16	79.96	70.08	89.96	79.57	90.63	81.11
	Mask2Former [14]	85.77	91.24	70.95	79.74	75.23	90.03	80.59	90.60	81.66
	IQ2Former(ours)	**86.71**	**91.79**	**72.63**	**80.43**	**76.51**	**90.57**	**81.61**	**91.18**	**82.98**
Swin-L	UPerNet [62]	86.98	92.32	72.56	80.51	78.70	90.74	82.21	91.21	83.28
	MaskFormer [13]	85.58	91.62	71.91	78.96	76.44	90.03	80.90	91.39	82.88
	Mask2Former [14]	86.47	92.00	72.36	80.59	78.61	90.56	82.01	90.94	82.68
	IQ2Former(ours)	**87.09**	**92.71**	**73.13**	**81.03**	**81.74**	**90.99**	**83.14**	**91.44**	**83.59**

Note: (1) Imp. surf. is the abbreviation for impervious surface; (2) Low veg. is the abbreviation for low vegetation.

Table 3. Results of quantitative comparison on the Potsdam testing set. The numbers are the percent scores (%). † indicates that the scores are acquired via the flip and MS testing. Bold font means the highest performance of that class.

Backbone	Models	Imp. Surf.	Building	Low Veg.	Tree	Car	OA	mIOU	OA †	mIOU †
ResNet-50	MaskFormer [13]	85.78	91.61	75.85	78.46	90.84	89.90	84.51	90.65	85.66
	Mask2Former [14]	86.32	92.65	76.07	78.58	92.63	90.22	85.25	90.94	86.32
	IQ2Former(ours)	**87.97**	**93.98**	**77.71**	**80.11**	**93.54**	**91.17**	**86.66**	**91.57**	**87.34**
ResNet-101	FCN [8]	87.00	93.58	75.77	78.90	92.44	90.44	85.54	90.82	86.23
	PSPNet [9]	87.44	**94.03**	76.64	79.33	93.07	90.80	86.10	91.29	86.81
	DeepLabV3+ [11]	87.40	93.82	76.60	79.28	93.07	90.80	86.03	91.27	86.72
	OCRNet [23]	85.17	90.22	75.31	76.96	89.83	89.33	83.50	90.21	84.92
	UPerNet [62]	87.33	93.61	76.62	79.58	92.39	90.78	85.91	91.38	86.88
	MaskFormer [13]	86.46	92.89	76.32	78.82	91.43	90.36	85.18	90.97	86.21
	Mask2Former [14]	86.50	93.33	76.51	79.08	92.36	90.48	85.56	91.19	86.64
	IQ2Former(ours)	**87.61**	93.65	**78.19**	**80.61**	**93.61**	**91.18**	**86.74**	**91.66**	**87.49**

Table 3. *Cont.*

Backbone	Models	Imp. Surf.	Building	Low Veg.	Tree	Car	OA	mIOU	OA †	mIOU †
Swin-L	UPerNet [62]	**88.15**	94.50	78.47	80.60	93.20	91.42	86.98	91.75	87.55
	MaskFormer [13]	87.41	94.71	78.62	80.16	93.11	91.32	86.80	91.66	87.21
	Mask2Former [14]	87.88	94.42	78.97	81.01	93.11	91.60	87.08	91.85	87.58
	IQ2Former(ours)	87.88	**94.82**	**79.68**	**81.17**	**94.33**	**91.67**	**87.58**	**91.92**	**87.89**

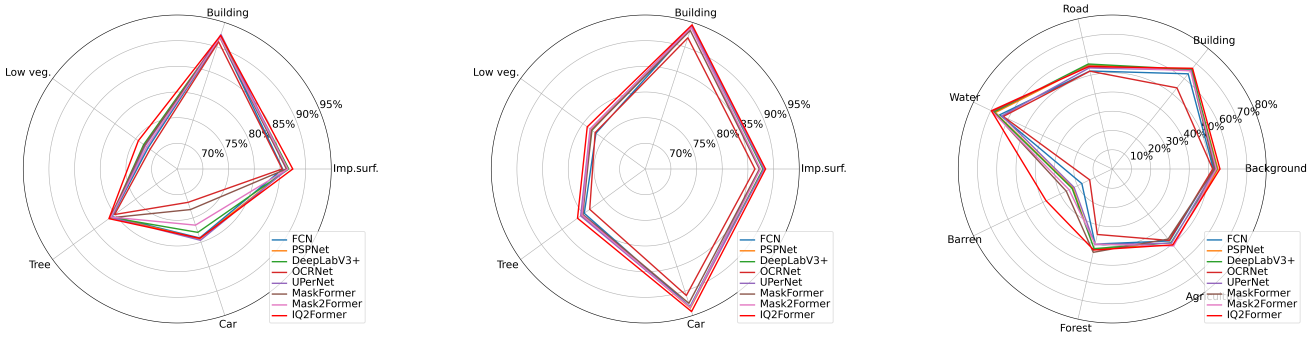
Note: (1) Imp. surf. is the abbreviation for Impervious surface. (2) Low veg. is the abbreviation for Low vegetation.

Table 4. Results of quantitative comparison on the LoveDA testing set. The numbers are the percent scores (%). † indicates that the scores are acquired via the flip and MS testing. Bold font means the highest performance of that class.

Backbone	Model	Back.	Build.	Road	Water	Barren	Forest	Agri.	OA	mIOU	OA †	mIOU †
ResNet-50	MaskFormer [13]	52.87	62.42	53.68	**68.47**	28.08	41.97	48.47	68.71	50.85	69.42	51.37
	Mask2Former [14]	53.67	**65.13**	**55.92**	66.24	24.58	40.57	50.41	69.33	50.93	69.54	50.97
	IQ2Former(ours)	**54.95**	62.80	53.86	65.54	**31.53**	**42.70**	**51.93**	**69.91**	**51.90**	**70.19**	**52.19**
ResNet-101	FCN [8]	52.75	62.63	53.62	66.06	22.38	38.97	49.54	68.11	49.42	67.47	48.31
	PSPNet [9]	**55.14**	64.24	55.54	68.03	27.01	41.56	**51.53**	70.27	51.86	69.98	51.34
	DeepLabV3+ [11]	54.19	64.39	55.67	68.14	27.17	41.44	49.29	69.50	51.47	69.60	51.32
	OCRNet [23]	53.10	51.79	54.56	59.71	23.70	35.69	46.99	66.56	46.51	65.54	45.21
	UPerNet [62]	54.09	64.31	54.51	65.27	25.85	40.02	50.52	69.29	50.65	68.69	49.74
	MaskFormer [13]	53.03	63.40	54.86	**70.80**	29.14	**43.96**	46.28	68.80	51.64	68.83	51.39
	Mask2Former [14]	53.88	64.62	55.57	69.52	27.15	40.27	51.37	69.91	51.77	69.94	51.39
	IQ2Former(ours)	54.88	**66.08**	**56.08**	70.78	**36.67**	42.70	49.58	**70.34**	**53.82**	**70.95**	**54.11**
Swin-L	UPerNet [62]	54.25	67.40	56.74	**72.92**	29.75	44.31	52.81	71.01	54.03	71.47	54.13
	MaskFormer [13]	53.88	66.99	57.53	72.29	28.94	**44.47**	54.32	71.18	54.06	71.76	54.66
	Mask2Former [14]	54.07	**68.50**	58.31	71.73	29.83	41.19	**58.03**	**71.94**	54.52	72.49	54.90
	IQ2Former(ours)	**54.71**	68.43	**59.16**	72.80	**35.08**	40.37	54.69	71.50	**55.03**	**72.65**	**56.31**

Note: (1) Back. is the abbreviation for background. (2) Build. is the abbreviation for building. (3) Agri. is the abbreviation for agriculture.

Figure 5 shows the radar plots using test-time augmentation on the three datasets to further compare our model with baseline models, category by category. The points in these plots represent the corresponding mIoU scores, which are obtained via test-time augmentation tricks on the testing dataset. From these plots, it can be seen that the curves obtained by our IQ2Former are always located in the external region, indicating that it achieves a higher performance compared to the baseline models.



(**a**) Vaihingen. (**b**) Potsdam. (**c**) LoveDA.

Figure 5. Category-by-category comparisons on the three datasets via the Radar chart. The digits are the *mIoU* scores, obtained via the flip and MS testing. For the sake of fairness, the backbone of all models is ResNet-101.

For the convenience of comparison, Figures 6–8 show the visual segmentation results of sample images obtained by all models, including the FCN [8], PSPNet [9], DeepLabv3+ [11], OCRNet [23], UPerNet [62], MaskFormer [13], Mask2Former [14], IQ2Former. For the sake of fairness, the backbone of all models is ResNet-101 [70].

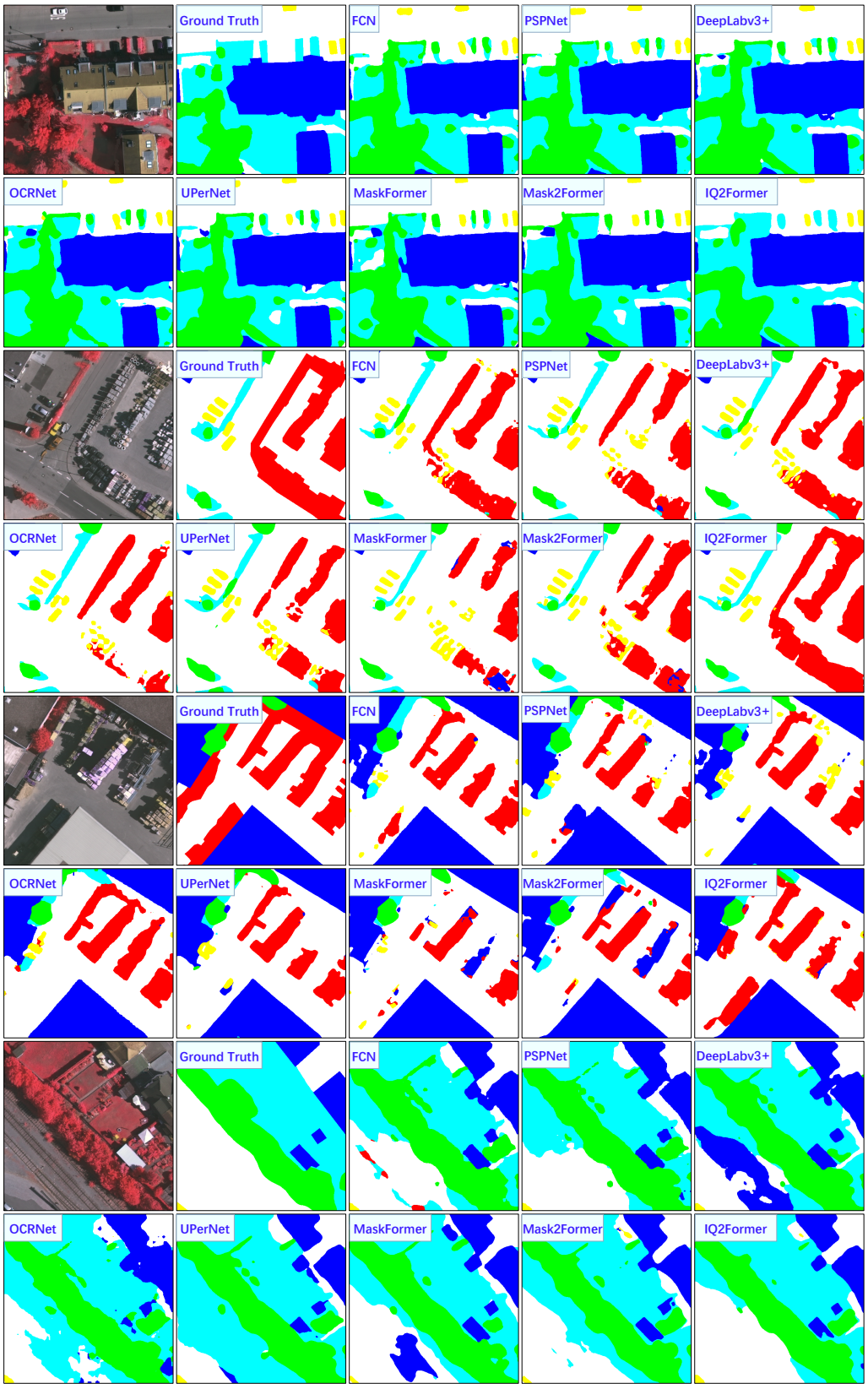


Figure 6. Visual segmentation comparisons between our model and other related models on the Vaihingen dataset. For the sake of fairness, the backbone of all models is ResNet-101. The label includes six categories: impervious surface (white), building (blue), low vegetation (cyan), tree (green), car (yellow), and clutter/background (red).

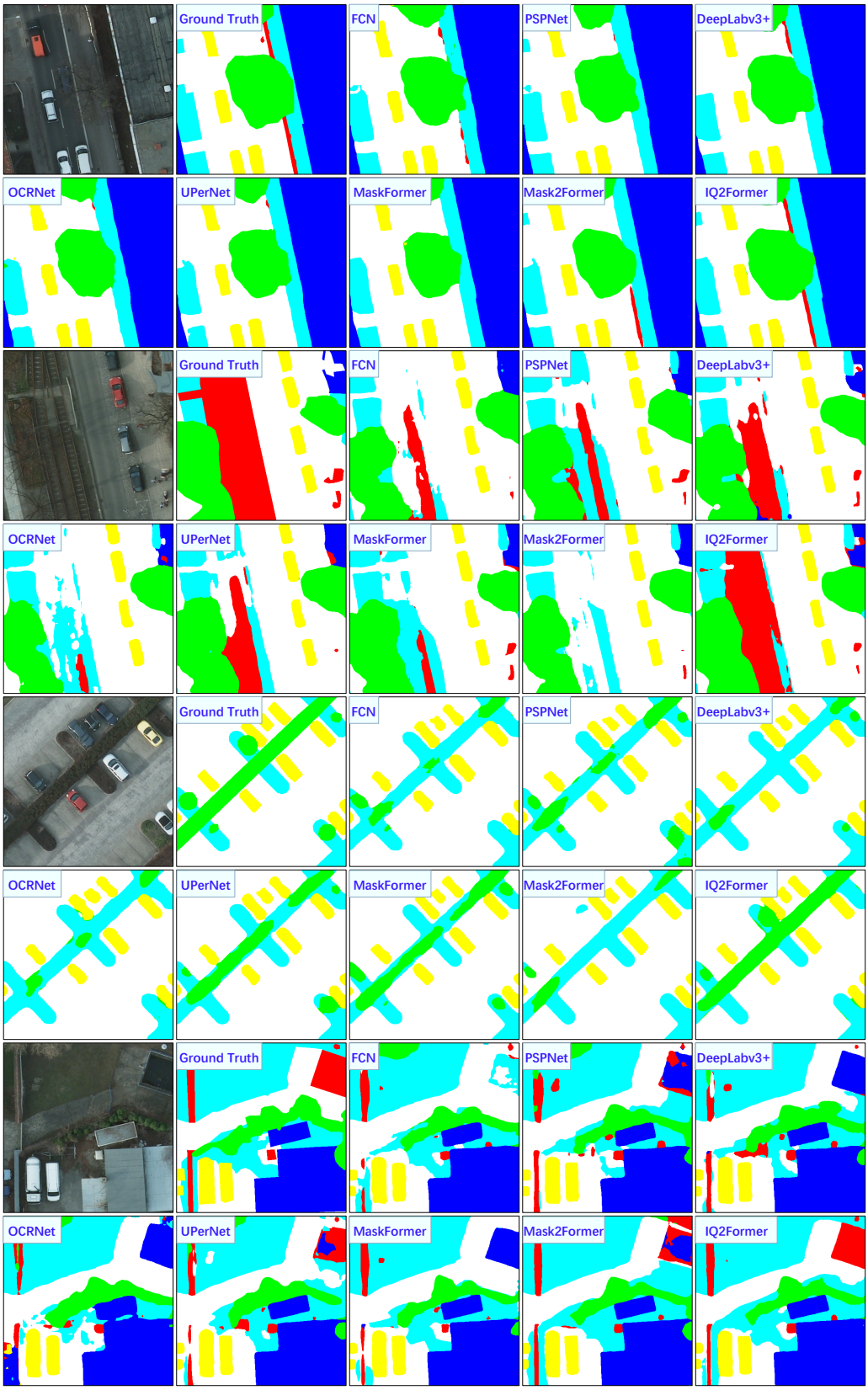


Figure 7. Visual segmentation comparisons between our model and other related models on the Potsdam dataset. For the sake of fairness, the backbone of all models is ResNet-101. The label includes six categories: impervious surface (white), building (blue), low vegetation (cyan), tree (green), car (yellow), and clutter/background (red).

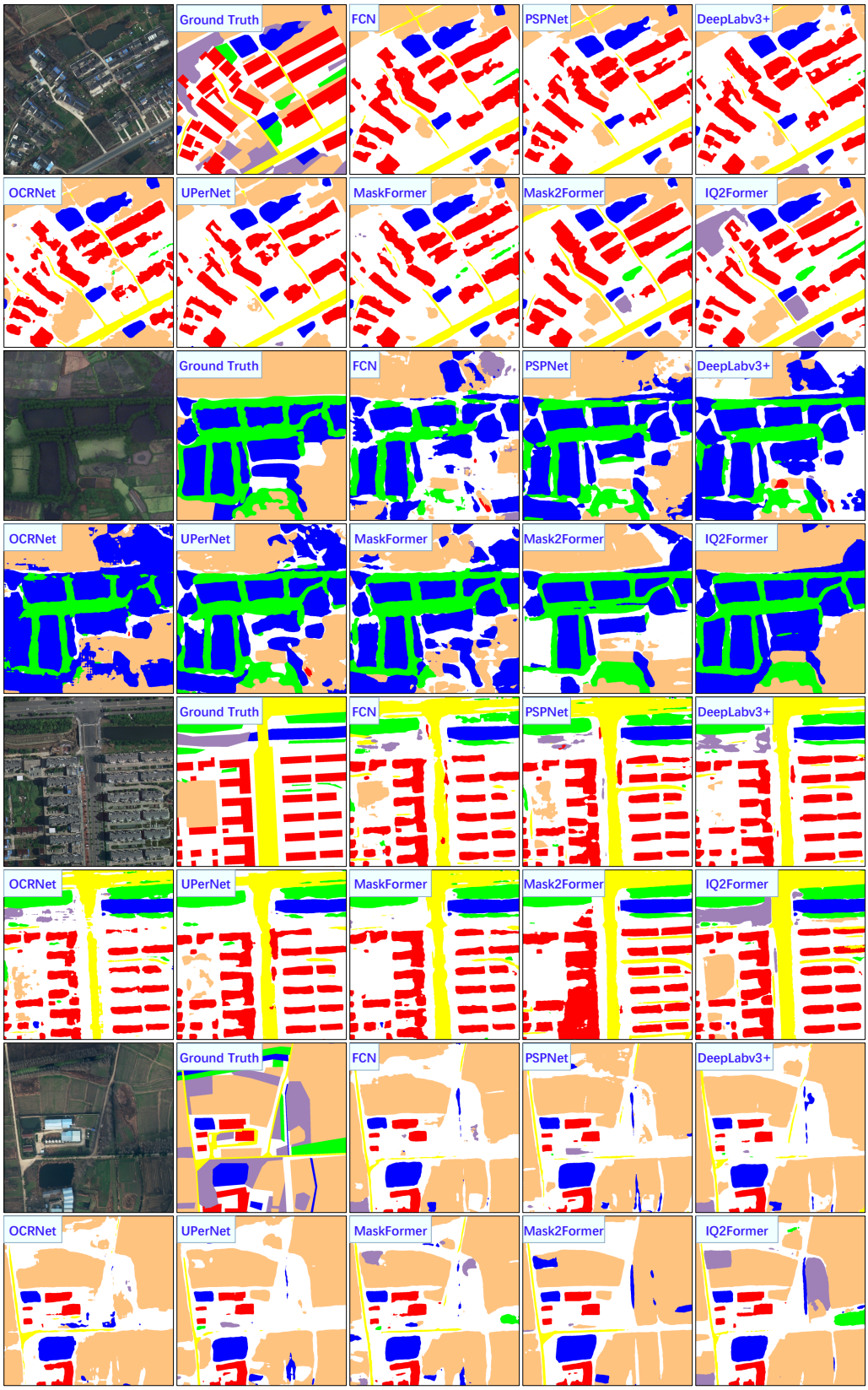


Figure 8. Visual segmentation comparisons between our model and other related models on the LoveDA dataset. For the sake of fairness, the backbone of all models is ResNet-101. The label includes six categories: background (white), building (red), road (yellow), water (blue), barren (plum), forest (green), and agriculture (orange).

To summarize, the above comparisons demonstrate that our IQ2Former is capable of successfully segmenting RS images with high resolution.

4.5. Ablation Study

This subsection describes the ablation experiments to evaluate the effectiveness of the three components proposed in our model. Table 5 illustrates the validity of the query scenario module in Section 3.2.1, query position module in Section 3.2.2, and query attention module in Section 3.2.3. It can be seen that performing both fundamental components helps enhance the performance. It is worth pointing out that there are no increases in the number of parameters in QPM.

Table 5. Effectiveness analysis of each module. "–" represents the removal of this module. It can be seen that our method performs better than other variant methods. Bold font means the highest performance of that class.

Module	Imp. Surf.	Building	Low Veg.	Tree	Car	mIoU	OA	# Params.	FLOPS
IQ2Former(ours)	**86.71**	**91.79**	**72.63**	80.43	76.51	**81.61**	**90.57**	63.122 M	80.698 G
– QSM	85.67	91.76	71.02	80.12	76.21	80.96	90.16	63.102 M	80.693 G
– QPM	86.48	91.03	72.27	80.81	**76.54**	81.43	90.51	63.122 M	80.677 G
– QAM	86.19	91.67	72.37	**80.89**	75.10	81.25	90.50	63.107 M	80.698 G
– all 3 modules above	85.77	91.24	70.95	79.74	75.23	80.59	90.03	62.996 M	85.608 G

Note: (1) Imp. surf. is the abbreviation for impervious surface. (2) Low veg. is the abbreviation for low vegetation. (3) Params. is the abbreviation for parameters.

For the query scenario module, we already verified that the performance of our IQ2Former is higher than the Mask2Former [14]. Tables 6 and 7 serve as the foundation for selecting hyper-parameters in QSM for the Vaihingen and Potsdam datasets. Table 8 is the basis for assigning hyper-parameters in QSM for the LoveDA dataset. The factors related to the computational efficiency are listed in Table 9, including the number of the parameters and the number of the floating-point operations (FLOPs) with giga multiplier accumulators (GMACs) in the model.

Table 6. The choice of four query groups is better than others in terms of performance. The experiments are conducted using the Vaihingen dataset as an example. The number of queries is 20 here. Bold font means the highest performance of that class.

Number of Groups	Imp. Surf.	Building	Low Veg.	Tree	Car	mIoU	OA	# Params.	FLOPS
# 1	85.67	91.76	71.02	80.12	76.21	90.16	80.96	63.102 M	80.693 G
# 2	86.21	91.58	71.48	80.06	76.08	90.26	81.08	63.112 M	80.698 G
# 4	**86.71**	**91.79**	**72.63**	80.43	76.51	**90.57**	**81.61**	63.122 M	80.698 G
# 8	86.17	91.37	72.02	**80.65**	**77.59**	90.34	81.56	63.142 M	80.698 G

Note: (1) Imp. surf. is the abbreviation for impervious surface. (2) Low veg. is the abbreviation for low vegetation. (3) Params. is the abbreviation for parameters.

Table 7. The choice of twenty queries is better than others in terms of performance. The experiments are conducted using the Vaihingen dataset as an example. The number of query groups is four here. Bold font means the highest performance of that class.

Number of Queries	Imp. Surf.	Building	Low Veg.	Tree	Car	mIoU	OA	# Params.	FLOPS
# 10	86.30	**91.99**	72.07	80.58	76.57	81.50	90.51	63.108 M	80.074 G
# 20	**86.71**	91.79	**72.63**	80.43	76.51	**81.61**	**90.57**	63.122 M	80.698 G
# 50	85.82	91.51	72.20	**80.78**	76.80	81.42	90.38	63.163 M	82.573 G
# 100	85.63	91.93	71.54	80.06	**78.00**	81.43	90.21	63.232 M	85.717 G

Note: (1) Imp. surf. is the abbreviation for impervious surface. (2) Low veg. is the abbreviation for low vegetation. (3) Params. is the abbreviation for parameters.

Table 8. Performance comparison between 20, 40, and 100 queries in LoveDA datasets. † indicates that the scores are acquired via the flip and MS testing. Bold font means the highest performance of that class.

Number of Queries	Back.	Build.	Road	Water	Barren	Forest	Agri.	OA	mIOU	OA †	mIOU †
# 20	53.57	64.01	**57.23**	70.65	29.94	41.33	49.00	69.53	52.25	70.00	52.94
# 40	54.88	**66.08**	56.08	**70.78**	**36.67**	**42.70**	49.58	70.34	**53.82**	**70.95**	**54.11**
# 100	**55.92**	65.53	53.95	68.13	30.10	40.43	**53.69**	**71.13**	52.54	70.65	52.04

Note: (1) Back. is the abbreviation for background. (2) Build. is the abbreviation for building. (3) Agri. is the abbreviation for agriculture.

Table 9. Comparison of computational efficiency among different models, including the floating-point operations and the total number of parameters. For the sake of fairness, the backbone of all models is ResNet-101. The experiments are conducted using the Vaihingen dataset as an example.

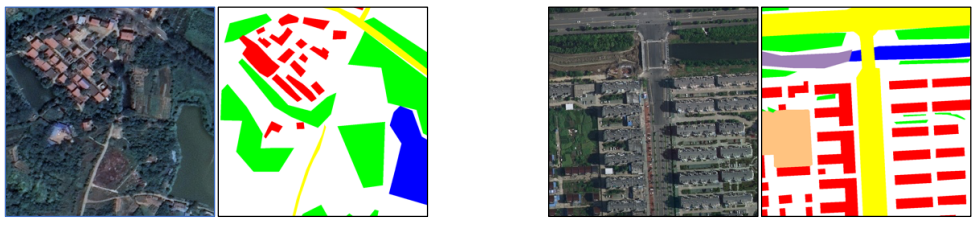
Method	# Params.	FLOPs	mIoU (%)
FCN [8]	68.48 M	275.38 G	80.47
PSPNet [9]	67.96 M	256.14 G	80.75
DeepLabV3+ [11]	62.57 M	253.93 G	80.83
OCRNet [23]	55.51 M	230.57 G	78.11
UPerNet [62]	64.04 M	236.99 G	80.97
MaskFormer [13]	60.26 M	70.32 G	79.57
Mask2Former [14]	63.00 M	85.61 G	80.59
IQ2Former (ours)	63.12 M	80.70 G	81.61

Note: (1) Params. is the abbreviation for parameters. (2) FLOPs is the abbreviation for floating-point operations.

5. Discussions

5.1. Discussion about the Number of Queries in the IQ2Former

In our IQ2Former, the number of queries in the Vaihingen and Potsdam datasets is set to 20, while the number of queries in the LovaDA dataset is taken as 40. There are three reasons to explain this setting. First, there are two significant categories of scenarios in the LoveDA dataset, namely rural and urban, which render large differences in the visual appearances. Second, the amount of data in LoveDA is much more enormous than the first two datasets, which record many more scenes from different regions. Finally, the Vaihingen and Potsdam datasets only require the segmentation of five classes of objects, while LoveDA needs to achieve the segmentation of seven classes of objects. As shown in Figure 9, urban storied buildings and rural one-story houses are both considered as building in annotations, the winding paths in rural areas and the wide roads in urban areas are both considered as road in annotations, and the trees along urban streets and the vast forests in rural areas are both considered as forest in annotations.



(a) Rural example image and annotation. (b) Urban example image and annotation.

Figure 9. A comparative example of the LoveDA dataset for rural and urban scenarios. The building is depicted in red, the road in yellow, and the forest in green.

Therefore, we suppose that more queries for the LoveDA are beneficial to our model. To validate our hypothesis, we randomly picked two images of the countryside and two of the city for the query comparison. For clarity, the query in the QSM is visualized by the normalization of the $Query^{*}_{QSM}$ for intuitive visualization, which is expressed as follows:

$$Query\ (QSM) = \frac{Query^{*}_{QSM} - \min(Query^{*}_{QSM})}{\max(Query^{*}_{QSM}) - \min(Query^{*}_{QSM})}, \tag{8}$$

where $Query^{*}_{QSM}$ records the selected queries calculated by Equation (4). Figure 10 illustrates the visualization. As shown in Figure 10, the queries for the same scenarios are more similar to each other (up to about cosine similarity 99.6% for urban scenarios and 99.7% for rural scenarios). In contrast, the queries for different scenarios are more different from each other (with an average cosine similarity of 69.4% for various scenarios). The experimental results in Table 8 also support our hypothesis. The above comparative evaluation indicates that our IQ2Former has enough flexibility and capability for setting an approximate number of queries, conveying a powerful learning ability for both simple and complex scenes. This facilitates its usage for real-world applications.

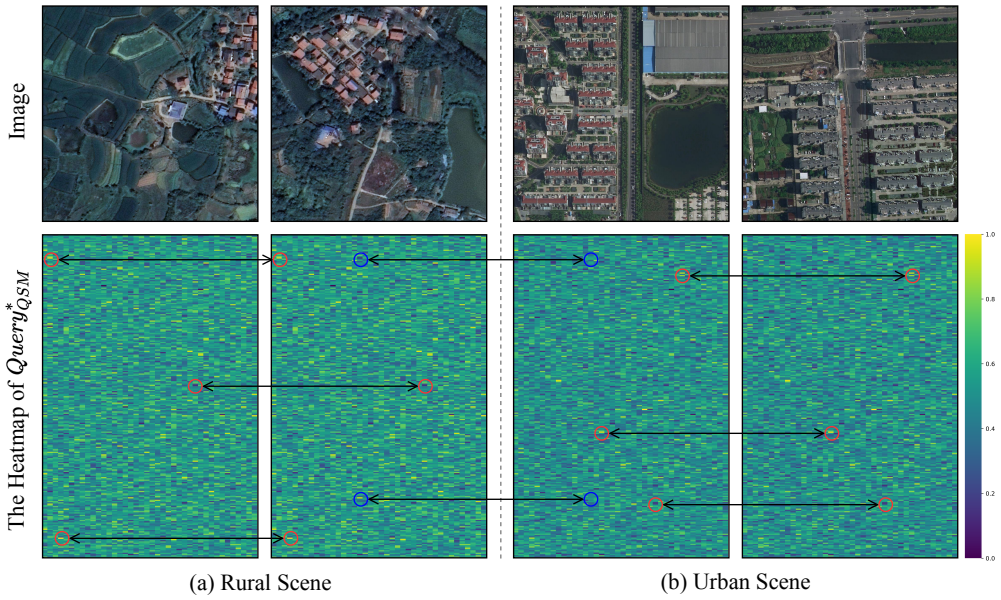


Figure 10. A comparison of $Query^{*}_{QSM}$ heatmaps for rural and urban scenarios in the LoveDA dataset. The ◯ indicates the similar features, while the ◯ represents various query features.

In addition, the original MaskFormer [13] claimed that a total of 100 queries could consistently perform the best across general image scene datasets and suggested that it is necessary to adjust the number of queries concerning the number of categories or the volume of datasets. However, this study indicates that, at least on three RS datasets, it experimentally proves that 100 queries are not the best and unique choice. In other words, one can set the number of queries by comprehensively considering the following aspects, including the complexity of the scenarios, the difference between the scenes recorded in the images, the volume of the dataset for training, and the number of categories to be segmented, and so on.

5.2. Implications and Limitations

In this study, we conducted experimental evaluations on the three publicly challenging datasets. Our IQ2Former mainly aims to improve the model's performance for semantic segmentation in RS images with the improved query. Specifically, the main contribution of the original Mask2Former [14] is to design a Transformer-based architecture for mask classification, while for the query, it only performs a randomly initialized learnable query, lacking the design of the query. Based on this, we design three improved query modules, namely the query scenario module in Section 3.2.1, which is used to implement adaptive weighted queries according to different RS scenarios; the query position module in Section 3.2.2, which enhances the query's sensitivity to the location information for the

different positions of objects in the same RS scene; and the query attention module in Section 3.2.3 for any RS image, to perform additional attention weighting of the query to enhance the performance of our IQ2Former. Technically, the IQ2Former model we proposed is mainly an improvement to query for effective representation learning. However, it has several limitations, which are described as follows:

- The QSM uses multiple groups of the query to be adaptively weighted according to different RS scenarios to obtain a better performance and achieve fewer calculations. However, since the number of initial query groups of the model is fixed, the number of parameters is not reduced to a significant degree, as shown in Table 9. In subsequent research, designing dynamic group numbers of the query for different RS scenarios is a direction worthy of further research.
- The QAM mainly introduces additional attention to learn for querying. However, in RS semantic segmentation tasks, it is essential to make full use of multi-scale visual features. This implies that our model can introduce other attention mechanisms and thoroughly combine them with our querying mechanism used in this study, thereby expanding its expression ability and performance.
- The three query modules designed in this study can be used not only for RS segmentation tasks but also as pluggable modules for other query-based transformer architectures. Future research directions include applying our modules to additional visual tasks and other types of RS data (like hyper-spectral RS data), offering further experimental validation for architecture optimization about these query modules, for example, reducing their computational complexity while keeping their representation ability.

6. Conclusions

This paper has proposed an IQ2Former for semantic segmentation in RS images. Technically, we have improved the query capability of the model in three aspects. Such an improvement fulfilled in this study is due to the fact that the embedding of the querying mechanism largely determines the representational power of the MaskFormer-like models. For selecting different remote sensing image scenarios, the QSM is designed to learn to group the queries from feature maps, which serve to select different scenarios such as urban and rural areas, building clusters, and parking lots. For classifying small targets in complex RS images, the QPM is constructed to assign the image position information to the query without increasing the number of parameters. For utilizing lower features ignored by Mask2Former, the QAM is proposed to be positioned between the duplicate transformer decoder layers, which mainly utilizes the characteristics of ODConv to extract valuable information from the previous query. With our QAM, the supervisory information is fully utilized, and the find-grained information is further exploited to achieve high-quality segmentations.

Comprehensive experiments have been conducted on three publicly challenging RS image datasets. Our model achieves 91.44% OA and 83.59% mIoU on the Vaihingen dataset. Our model achieves 91.92% OA and 87.89% mIoU on the Potsdam dataset. Our model achieves 72.63% OA and 56.31% mIoU on the LoveDA dataset. Additionally, ablation experiments and visual segmentation figures all demonstrate the effectiveness and superiority of our IQ2Former. In the future, we would like to conduct the research in the following three directions. First, the design of the query mechanism developed in our model could be optimized with more powerful attention and lightweight tricks. Second, we could comprehensively evaluate the performances of our model on the datasets containing RS images with low resolutions, noise, or a percentage of distortions. Third, we would also like to extend the applications of our model in the multi-spectral RS images and hyper-spectral RS data.

Author Contributions: Conceptualization, S.G. and Q.Y.; methodology, S.G. and Q.Y.; software, Q.Y. and S.G.; validation, S.X., S.W. and X.W.; investigation, S.X., S.W. and X.W.; writing—original draft preparation, S.G. and Q.Y.; writing—review and editing, S.G. and Q.Y.; formal analysis, Q.Y. and S.X.; visualization, Q.Y. and S.G.; data curation, Q.Y., S.X. and X.W.; supervision, S.X. and X.W.; resources, S.X. and X.W.; project administration, S.X. and X.W.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Research Program of Frontier Sciences, CAS (grant number: ZDBS-LY-DQC016), National Key Research and Development Program of China (grant number: 2022YFF1301803).

Data Availability Statement: Three public datasets (i.e., the Vaihingen, Potsdam, and LoveDA datasets) were included in this study. Both the Vaihingen dataset and the Potsdam dataset were obtained via the official website: https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx (accessed on 26 April 2023). The LoveDA dataset was downloaded from the webpage: https://github.com/Junjue-Wang/LoveDA (accessed on 26 April 2023).

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

References

1. Caesar, H.; Uijlings, J.R.R.; Ferrari, V. COCO-Stuff: Thing and Stuff Classes in Context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1209–1218.
2. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
3. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene parsing through ade20k dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 633–641.
4. 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.
5. Neuhold, G.; Ollmann, T.; Rota Bulo, S.; Kontschieder, P. The mapillary vistas dataset for semantic understanding of street scenes. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4990–4999.
6. Rottensteiner, F.; Sohn, G.; Gerke, M.; Wegner, J.D. ISPRS Test Project on Urban Classification, 3D Building Reconstruction and Semantic Labeling. 2013. Available online: https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx (accessed on 1 February 2024).
7. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. *arXiv* **2021**, arXiv:2110.08733.
8. 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.
9. 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.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
11. 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.
12. 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.
13. 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.
14. 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.
15. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
16. Li, C.; Zhou, A.; Yao, A. Omni-Dimensional Dynamic Convolution. *arXiv* **2022**, arXiv:2209.07947.
17. Yang, B.; Bender, G.; Le, Q.V.; Ngiam, J. Condconv: Conditionally parameterized convolutions for efficient inference. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.

18. Ma, N.; Zhang, X.; Huang, J.; Sun, J. Weightnet: Revisiting the design space of weight networks. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 776–792.
19. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic convolution: Attention over convolution kernels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11030–11039.
20. Zhang, Y.; Zhang, J.; Wang, Q.; Zhong, Z. DyNet: Dynamic Convolution for Accelerating Convolutional Neural Networks. *arXiv* **2020**, arXiv:2004.10694.
21. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
22. 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.
23. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part VI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 173–190.
24. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
25. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5386–5395.
26. Guo, S.; Yang, Q.; Xiang, S.; Wang, P.; Wang, X. Dynamic High-Resolution Network for Semantic Segmentation in Remote-Sensing Images. *Remote Sens.* **2023**, *15*, 2293. [CrossRef]
27. 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*.
28. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
29. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
30. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
31. Jain, J.; Li, J.; Chiu, M.T.; Hassani, A.; Orlov, N.; Shi, H. Oneformer: One transformer to rule universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2989–2998.
32. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Vancouver, BC, Canada, 17–24 June 2023; pp. 3992–4003.
33. Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Yang, J.; Li, C.; Zhang, L.; Gao, J. Semantic-sam: Segment and recognize anything at any granularity. *arXiv* **2023**, arXiv:2307.04767.
34. Zhang, C.; Han, D.; Qiao, Y.; Kim, J.U.; Bae, S.H.; Lee, S.; Hong, C.S. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *arXiv* **2023**, arXiv:2306.14289.
35. Jiang, B.; An, X.; Xu, S.; Chen, Z. Intelligent image semantic segmentation: A review through deep learning techniques for remote sensing image analysis. *J. Indian Soc. Remote Sens.* **2023**, *51*, 1865–1878. [CrossRef]
36. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
37. Alshehhi, R.; Marpu, P.R.; Woon, W.L.; Dalla Mura, M. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 139–149. [CrossRef]
38. Chen, K.; Fu, K.; Yan, M.; Gao, X.; Sun, X.; Wei, X. Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 173–177. [CrossRef]
39. Chen, G.; Zhang, X.; Wang, Q.; Dai, F.; Gong, Y.; Zhu, K. Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1633–1644. [CrossRef]
40. Liu, Y.; Fan, B.; Wang, L.; Bai, J.; Xiang, S.; Pan, C. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 78–95. [CrossRef]
41. Guo, S.; Jin, Q.; Wang, H.; Wang, X.; Wang, Y.; Xiang, S. Learnable gated convolutional neural network for semantic segmentation in remote-sensing images. *Remote Sens.* **2019**, *11*, 1922. [CrossRef]
42. Chen, B.; Xia, M.; Qian, M.; Huang, J. MANet: A multi-level aggregation network for semantic segmentation of high-resolution remote sensing images. *Int. J. Remote Sens.* **2022**, *43*, 5874–5894. [CrossRef]
43. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]

44. Zhang, J.; Lin, S.; Ding, L.; Bruzzone, L. Multi-scale context aggregation for semantic segmentation of remote sensing images. *Remote Sens.* **2020**, *12*, 701. [CrossRef]
45. Xu, Z.; Zhang, W.; Zhang, T.; Li, J. HRCNet: High-resolution context extraction network for semantic segmentation of remote sensing images. *Remote Sens.* **2020**, *13*, 71. [CrossRef]
46. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 426–435. [CrossRef]
47. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607713. [CrossRef]
48. Zhao, Q.; Liu, J.; Li, Y.; Zhang, H. Semantic segmentation with attention mechanism for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5403913. [CrossRef]
49. Zhang, M.; Jing, W.; Lin, J.; Fang, N.; Wei, W.; Woźniak, M.; Damaševičius, R. NAS-HRIS: Automatic design and architecture search of neural network for semantic segmentation in remote sensing images. *Sensors* **2020**, *20*, 5292. [CrossRef] [PubMed]
50. Shi, Q.; Liu, M.; Liu, X.; Liu, P.; Zhang, P.; Yang, J.; Li, X. Domain adaption for fine-grained urban village extraction from satellite images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1430–1434. [CrossRef]
51. Wang, Y.; Li, Y.; Chen, W.; Li, Y.; Dang, B. DNAS: Decoupling Neural Architecture Search for High-Resolution Remote Sensing Image Semantic Segmentation. *Remote Sens.* **2022**, *14*, 3864. [CrossRef]
52. Broni-Bediako, C.; Murata, Y.; Mormille, L.H.; Atsumi, M. Evolutionary NAS for aerial image segmentation with gene expression programming of cellular encoding. *Neural Comput. Appl.* **2022**, *34*, 14185–14204. [CrossRef]
53. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. [CrossRef]
54. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
55. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Transformer-based decoder designs for semantic segmentation on remotely sensed images. *Remote Sens.* **2021**, *13*, 5100. [CrossRef]
56. Ye, W.; Zhang, W.; Lei, W.; Zhang, W.; Chen, X.; Wang, Y. Remote sensing image instance segmentation network with transformer and multi-scale feature representation. *Expert Syst. Appl.* **2023**, *234*, 121007. [CrossRef]
57. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [CrossRef]
58. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408820. [CrossRef]
59. Sun, Z.; Cao, S.; Yang, Y.; Kitani, K.M. Rethinking transformer-based set prediction for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3611–3620.
60. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Vancouver, BC, Canada, 17–24 June 2023; pp. 3836–3847.
61. 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.
62. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
63. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
64. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable {DETR}: Deformable Transformers for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021.
65. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
66. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
67. 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]
68. Contributors, M. MMCV: OpenMMLab Computer Vision Foundation. 2018. Available online: https://github.com/open-mmlab/mmcv (accessed on 1 February 2024).
69. Contributors, M. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. 2020. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 1 February 2024).
70. 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.