remote sensing

MDPI

Article

SegMarsViT: Lightweight Mars Terrain Segmentation Network for Autonomous Driving in Planetary Exploration

Yuqi Dai [1,2], Tie Zheng [1,2], Changbin Xue [1,*] and Li Zhou [1]

[1] National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: xuechangbin@nssc.ac.cn

Abstract: Planetary rover systems need to perform terrain segmentation to identify feasible driving areas and surround obstacles, which falls into the research area of semantic segmentation. Recently, deep learning (DL)-based methods were proposed and achieved great performance for semantic segmentation. However, due to the on-board processor platform's strict comstraints on computational complexity and power consumption, existing DL approaches are almost impossible to be deployed on satellites under the burden of extensive computation and large model size. To fill this gap, this paper targeted studying effective and efficient Martian terrain segmentation solutions that are suitable for on-board satellites. In this article, we propose a lightweight ViT-based terrain segmentation method, namely, SegMarsViT. In the encoder part, the mobile vision transformer (MViT) block in the backbone extracts local–global spatial and captures multiscale contextual information concurrently. In the decoder part, the cross-scale feature fusion modules (CFF) further integrate hierarchical context information and the compact feature aggregation module (CFA) combines multi-level feature representation. Moreover, we evaluate the proposed method on three public datasets: AI4Mars, MSL-Seg, and S5Mars. Extensive experiments demonstrate that the proposed SegMarsViT was able to achieve 68.4%, 78.22%, and 67.28% mIoU on the AI4Mars-MSL, MSL-Seg, and S^{5}Mars, respectively, under the speed of 69.52 FPS.

Keywords: Mars terrain segmentation; semantic segmentation; planetary exploration

check for updates

Citation: Dai, Y.; Zheng, T.; Xue, C.; Zhou, L. SegMarsViT: Lightweight Mars Terrain Segmentation Network for Autonomous Driving in Planetary Exploration. *Remote Sens.* **2022**, *14*, 6297. https://doi.org/10.3390/rs14246297

Academic Editors: Jiaojiao Li, Qian Du, Wei Li, Bobo Xi, Jocelyn Chanussot, Rui Song and Yunsong Li

Received: 17 November 2022
Accepted: 9 December 2022
Published: 12 December 2022

1. Introduction

Intelligent environmental perception is a necessity for planetary rovers toward autonomous driving, which provides crucial semantic information, e.g., identifying feasible driving areas and surrounding obstacles. For such a panoptic perception mission, terrain segmentation is the most critical procedure, which also can be viewed as a semantic segmentation task. Semantic segmentation is a widely used perception method for self-driving vehicles on earth that can assign a separate predefined class label to each pixel of an image [1] it is the foundation of many high-level tasks that need to infer relevant semantic information from images for subsequent processing. This applies on self-driving vehicles on Mars as well. Therefore, this study explored the task of terrain segmentation on the Martian surface, aiming to characterize semantic information from rover images. As shown, the Figure 1a shows the Tianwen-1 Zhurong rover, China's first Mars rover, which is undergoing its fantastic exploration on the red planet. RGB sample images of the Mars surface and the corresponding terrain segmentation annotation are depicted in Figure 1b,c, respectively. It can be observed that semantic segmentation is a pixel-level dense prediction task, which requires an in-depth understanding of the semantics of the entire scene and is in some ways more challenging than those image-level prediction tasks.

Early image segmentation approaches dedicated to divide images into regions based on little more than basic color and low-level textual information [2,3]. With the rapid development of deep learning techniques in the 2010s, deep convolutional neural networks

(CNNs) became dominant in automatic semantic segmentation technology due to their tremendous modeling and learning capabilities, which strive to boost algorithm accuracy on the strength of massively parallel GPUs and large labelled datasets [4,5]. Long et al. [6] first proposed a fully convolutional network (FCNet), which is a revolutionary work and the majority of following state-of-the-art (SOTA) studies are extensions of the FCN architecture. One of the most pioneering works is UNet presented by Ronneberger et al. [7] for biomedical image segmentation, which adopts the influential encoder–decoder architecture and proved to be very useful for other types of image data [8–11]. Meanwhile, inspired by the high precision that CNNs achieved in semantic segmentation, many CNNs-based approaches were proposed for the Martian terrain segmentation (MTS) task. Rothrock et al. [12] proposed a soil property and object classification (SPOC) system based on DeepLab for visually identifying terrain types as well as terrain features (e.g., scarps, ridges) on a planetary surface. They also presented two successful applications to Mars rover missions, including the landing site traversability analysis and slip prediction. Iwashita et al. [13] proposed TU-Net and TDeelLab robust to illumination changes via data fusion from visible and thermal images. Liu et al. [14] proposed a hybrid attention-based terrain segmentation network called HASS for unstructured Martian images. Claudet et al. [15] employed advanced semantic segmentation algorithms to generate binary safety maps for the spacecraft safe planetary landing problem. Furthermore, several existing studies attempted to resolve the terrain segmentation issue by using wear-supervised techniques. Wang et al. [16] adopted the element-wise contrastive learning technique and proposed a semi-supervised learning framework for Mars imagery classification and segmentation through introducing online pseudo labels on the unlabeled areas. Goh et al. [17] proposed another semi-supervised Mars terrain segmentation algorithm with contrastive pretraining techniques. Zhang et al. [18] proposed a novel hybrid representation learning-based framework, which consists of a self-supervised pre-training stage and a semi-supervised learning phase for sparse data. Li et al. [19] introduced a stepwise domain adaptation Martian terrain segmentation network, which effectively alleviates covariate shift through unifying the color mapping space to further enhance the segmentation performance.

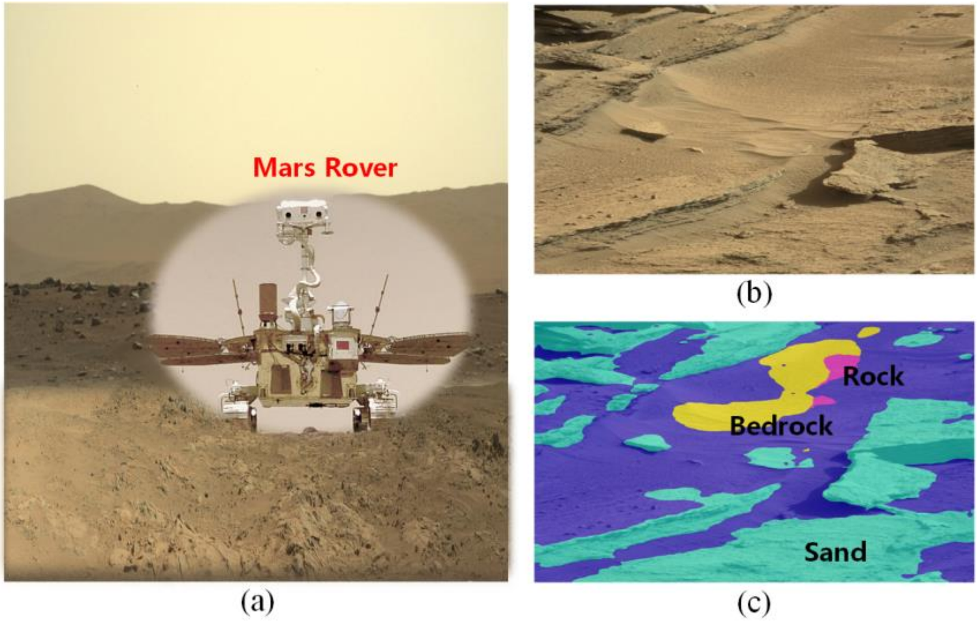


Figure 1. Planetary rover on Mars (**a**) and a sample image (**b**) along with its segmentation annotation for terrain types (**c**).

Furthermore, data-driven deep learning generally refers to learning directly through sufficient experience data. The level of success for deep learning applications is to a great extent determined by the quality and the depth of the data being used for training. In this respect, Mars terrain segmentation is currently attracting more and more attention, and scientific interest for deep learning-based segmentation datasets is growing rapidly. Several large-scale 2D image sets were established for the Mars terrain segmentation problem, the relevant information of which is listed in Table 1. Swan et al. [20] built the first large-scale dataset, AI4Mars, for the task of Mars terrain classification and traversability assessment, of which labels were obtained through a crowdsourcing approach and consisted four classes: soil, bedrock, sand, and big rock. Li et al. [19] extensively released a Mars terrain dataset annotated finely with nine classes, named Mars-Seg. Liu et al. [14] established a panorama semantic segmentation dataset for Mars rover images, named MarsScapes, which provides pixel-wise annotations for eight fined-grained categories. Zhang et al. [18] presented a high-resolution Mars terrain segmentation dataset, S^{5}Mars, annotated with pixel-level sparse labels for nine categories. The Martian surface condition is complicated and the corresponding annotation process is challenging. Hence, we thank all the above dataset creators that enabled us to conduct the research for this paper.

Table 1. To the best of our knowledge, there are already four public datasets established for MTS task up to now.

Dataset	Year	Classes	RGB
AI4Mars	2021	4	1.6 k
Mars-Seg	2021	8	~4.1 k
S^{5}Mars	2022	9	6 k
MarsScapes	2022	8	~18.5 k

In comparison to natural scene images, the Martian images have their particular characteristics. Objects on the surface of Mars exhibit unstructured characteristics with rich textures, ambiguous boundaries and diverse sizes, such as rocks and gravel [21]. Understanding unstructured scenes quite heavily depend on modeling the connection between the target pixel and its relevant surrounding content to a certain extent. Therefore, a limited receptive field is hard-pressed to meet demand, and several acquired rare target instances available for training are in small numbers. Class imbalance remains a problem in the MTS task. The above difficulties make it unreliable to directly apply the semantic segmentation methods designed for natural images on Martian terrain segmentation tasks.

On the other hand, CNN-based semantic segmentation methods always made brilliant achievements at the expense of high computational costs, large model size, and inference latency. This situation prevented recent state-of-the-art methods from being applied to real-world applications. Real on-board applications have a strong demand of semantic segmentation algorithms to run on resource-constrained edge devices in a timely manner. Therefore, deep models for the Mars terrain segmentation task should be efficient and accurate. Considering the performance limitations of spacecraft equipment, it is essential to develop efficient networks for accurate Mars terrain segmentation.

Toward this end, this paper proposes a novel lightweight Martian terrain segmentation model, named SegMarsViT. In the encoder part, the mobile vision transformer (MobileViT) backbone is leveraged to extract local–global spatial and capture high-level multiscale contextual information concurrently. An effective layer aggregation decoder (ELAD) is designed to further integrate hierarchical feature context information and generate powerful representations. Moreover, we evaluate the proposed method on three public datasets: AI4Mars, MSL-Seg, and S^{5}Mars. Extensive experiments demonstrate that the proposed SegMarsViT achieves comparable accuracy as the state-of-the-art semantic segmentation method. In the meantime, SegMarsViT has much less computation burden with a smaller model size.

The main contributions of this work can be summarized as follows:

(1) To the best of our knowledge, this is the first effort toward introducing the lightweight semantic segmentation model into the field of Martian terrain segmentation. We evaluate several representative semantic segmentation models and conduct enough comparable experiments. This is expected to facilitate the development and benchmarking of terrain segmentation algorithms in Martian images.

(2) We investigate a novel vision transformer-based deep neural network SegMarsViT for real-time and accurate Martian terrain segmentation. In the encoder, we employ a lightweight MobileViT backbone to capture a hierarchical feature. Notably, the proposed SegMarsViT is the first transformer-based network for the Martian terrain segmentation task. In the decoder part, a cross-scale feature fusion (CFF) module and a compact feature aggregation (CFA) technique are designed to strengthen and merge the multi-scale context feature.

(3) We conduct extensive experiments on AI4Mars, S5Mars, and MSL-Seg datasets. The results validate the effectiveness and efficiency of the proposed model, which can obtain competitive performance with 68.4%, 78.22%, and 67.28% mIoU, respectively. In the meantime, SegMarsViT has much less computation burden with smaller model size.

The remainder of this article is organized as follows: In Section 2, we will briefly introduce some previous work related to lightweight semantic segmentation and vision transformer. Section 3 describes the proposed method in detail. Section 4 provides overall performance and comparison results of the proposed method with analysis and discussion, and Section 5 concludes this study.

2. Related Work

2.1. Lightweight Semantic Segmentation

In real-world applications, such as robotics [22] and land resource monitoring [23], it is hard to deploy high-precision, high-complexity, and time-consuming semantic segmentation models for real-time inference speed in need. Hence, lightweight semantic segmentation networks came into being. Several research works were proposed to address the challenge of real-time semantic segmentation.

The standard convolution layer is the basic building layer in CNNs, which is computationally expensive. Real-time semantic segmentation pursues the fast data processing capability of the network. In order to meet the requirements of real-time inference performance and ensure high-quality prediction as much as possible, efficient convolution operations are generally used. For example, DABNet [24] introduced the depth-wise asymmetric bottleneck module, which increases efficiency through the combination of depth-wise separable and asymmetric factorized convolutions. ESPNet [25] proposed an efficient spatial pyramid module utilizing 1 × 1 grouped convolution to reduce dimension complexity and parallel dilation convolution modules to increase the effective receptive field, which results in a very compact and significant network. In addition, many other segmentation models, e.g., RTSeg [26] and EACNet [27], straightly employ the lightweight backbone networks designed for classification tasks as the feature extractor to improve the inference speed.

In addition to commonly used techniques for decreasing the latency and model size, designing novel and lightweight architectures is another effective solution. BiseNetV1 [28] is a two-branch architecture to reserve spatial feature information and enlarge the receptive field, which consists of a context path based on Xception architecture and a spatial branch based on strided convolution layers. Attention refinement modules (ARM) are applied to encode global context. The improved version, BiseNetV2 [29], further simplifies the architecture through utilizing the inverted bottleneck blocks of MobileNetv2 and efficient convolutions and obtains more favorable performance. The real-time general purpose semantic segmentation network (RGPNet) introduces a novel adapter module and a lightweight asymmetric encoder–decoder architecture. The adaptor module intermediates between encoder and decoder through the combination of features of three different levels. The strategy of integrating multi-scale context information results in excellent segmentation performance and the optimized progressive resizing training scheme makes RGPNet achieve an effective balance between speed and accuracy.

2.2. Vision Transformer-Based Semantic Segmentation

In spite of the exceptional representational power, CNN-based approaches generally exhibit limitations for modeling explicit long-range relations, due to the intrinsic local connectivity mechanism of convolution operations. Recently, transformer became a "hotspot" in the computer vision community, which was initially designed for sequence-to-sequence prediction and was powerful at modeling global contexts [30]. To overcome the limitation of the local receptive field of CNN, the latest efforts were focused on adapting transformer models into the computer vision sector [31,32], named vision transformer (ViT). Many scholars introduced the ViT mechanism into the semantic segmentation task. The two most common ways to do this are applying ViTs in conjunction with CNNs and developing pure ViTs. Wang et al. [33] proposed PVT, a pyramid vision transformer for dense prediction tasks, which is a natural extension of ViT with pyramid structures. Zheng et al. [34] proposed SETR, which is a hierarchical transformer from a sequence-to-sequence learning perspective, and it shows that good results can still be obtained without relying on the convolution operation. Huang et al. [35] designed a scale-wise intra-scale transformer, named ScaleFormer, of which the elaborate hybrid CNN-transformer backbone can effectively extract intra-scale local features and global information. Shi et al. [36] took the idea of the SwinTransformer [37] and presented the hierarchical SSformer with an elaborate and simple MLP decoder for semantic segmentation. Xie et al. [38] proposed SegFormer, which comprises a novel hierarchically structured transformer encoder and a lightweight all-MLP decoder, yielding great results. Hatamizadeh et al. proposed the UNetFormer [39] with a 3D SwinTransformer [40]-based encoder and a hybrid CNN-transformer decoder, which can achieve a trade-off performance between efficiency and accuracy for medical image segmentation. Similarly, there are UNETR [41] and nnFormer [42] in the same vein. Motivated by the astounding achievements of ViT, this paper presents the first study to explore the potential of ViT and fulfill the local–global semantics research gap in the context of Martian terrain segmentation.

3. Methodology

In this section, we first provide an overview of our method in the Section 3.1. Then, we introduce the lightweight encoder and effective decoder in the Sections 3.2 and 3.3, respectively. Finally, we present the loss function in the Section 3.4.

3.1. Framework Overview

The overall structure of the proposed SegMarsViT is illustrated in Figure 2. This paper is dedicated to the encoder–decoder segmentation architecture through ViT modules. The whole SegMarsViT is a novel combination of CNN and transformers to some extent, which has the local advantage of CNN and the long-range dependency merit of a transformer. The proposed network utilizes MobileViT backbone to extract corresponding features of five stages (stage1~stage5), whose outputs are denoted as F_1, F_2, F_3, F_4 and F_5, with scales of $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$, respectively. In other words, the output feature maps F_i after each stage are down-sampled with strides of 2^i. After the backbone, we perform an efficient stage-wise layer aggregation decoder, named ELAD, to generate segmentation outputs. The novel ELAD is designed to make multiscale features more distinguishable to learn representative features for SegMarsViT. In ELAD, a series of cross-scale feature fusion (CFF) modules are proposed to further enhance the context modeling and boost the cross-scale communication, which are built upon the top-down pathway. After obtained, we introduce a compact feature aggregation (CFA) module to ensure that feature maps extracted from different stages can be well merged. As shown in Figure 2, the proposed SegMarsViT is asymmetric and the contracting path is deeper than the expansion path. In what follows, we describe all the structures of the above-mentioned modules in detail.

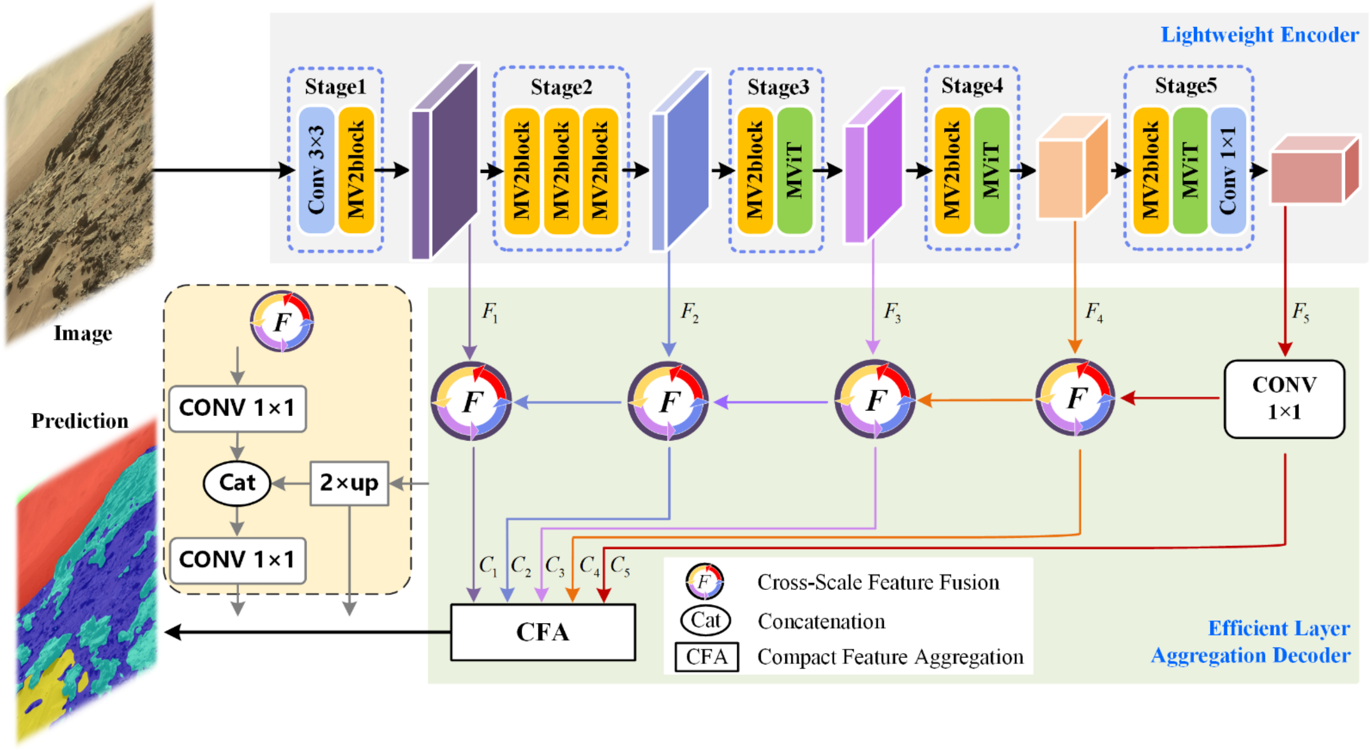


Figure 2. Framework overview of the proposed SegMarsViT.

3.2. Lightweight MViT-Based Encoder

Context modeling in not yet proven to be critical for segmentation and the encoder progressively reduces the spatial resolution and learns more abstract visual concepts with larger receptive fields. However, the encoder is always the most vital part of the whole framework and accounts for the dominant proportion of model size and computational budget.

Considering the strict complexity limitations on the spaceborne payload hardware, we use MobileViT as the backbone to accelerate feature extraction and improve the real-time performance of the proposed method. MobileViT is a lightweight and general-purpose neural network architecture introduced by Apple ML researchers. We removed the last pooling layers and all fully connected layers for image-to-image semantic segmentation prediction. With a special perspective to encode both local and global representations effectively, MobileViT is a hybrid network with both CNN and ViT-like properties. MobileViT improves its stability and performance through incorporating spatial inductive biases of CNN in ViT. As can be seen in Figure 2, the architecture of MobileViT contains the initial fully convolution layer, followed by several MV2 blocks and MViT blocks. Figure 3 visually depicts the design of the two main modules. The MV2 blocks (Figure 3a) come from MobileNetv2 [43] and are mainly responsible for down-sampling in the backbone. Even more to the point, unlike conventional ViTs, the elaborate MViT block (Figure 3b) can learn local and global information with an effective receptive field of $H \times W$.

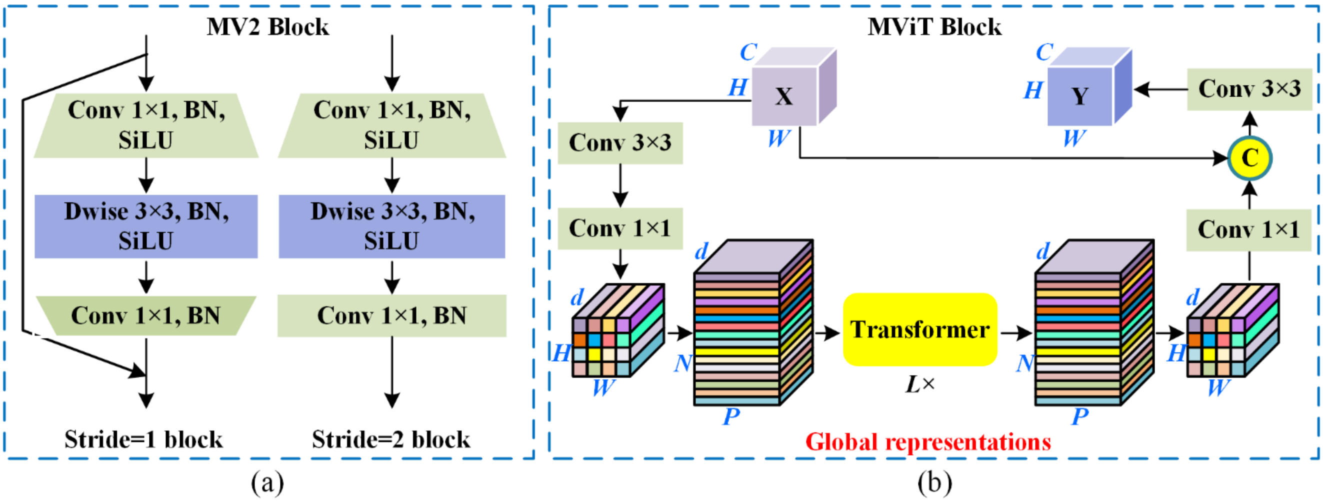


Figure 3. Two types of building blocks in MobileViT backbone: (**a**) MV2 block; (**b**) MViT block.

The first two layers are a standard 3 × 3 convolution layer and a 1 × 1 point-wise expansion layer, where the given input tensor $\mathbf{X} \in \mathbf{R}^{H \times W \times C}$ is projected to $\mathbf{X}_L \in \mathbf{R}^{H \times W \times d}(d > C)$. As is the first step of all the ViTs, $\mathbf{X}_L$ is then split into N non-overlapping patches $\mathbf{X}_U \in \mathbf{R}^{P \times N \times d}$. Next, the standard transformer blocks of multi-headed self-attention (MHA) [44] is applied to model long-range non-local dependencies as:

$$\mathbf{X}_G(p) = \text{Transformer}(\mathbf{X}_U(p)), 1 \leq p \leq P. \tag{1}$$

Then $\mathbf{X}_G \in \mathbf{R}^{P \times N \times d}$ will be folded to obtain $\mathbf{X}_F \in \mathbf{R}^{H \times W \times d}$ as the order of unfolding process. In the end, $\mathbf{X}_F$ will be projected to low C-dimensional space with a point-wise contraction layer and integrate with the raw input tensor $\mathbf{X}$ via concatenation and convolution operations.

The detailed configurations of the MobileViT model are shown in Table 2. The MobileViT models provide three different network sizes (s: small, xs: extral small, and xxs: extra extra small). To obtain multi-scale terrain information, the hierarchical output of five stages will be forwarded into the following decoder module.

Table 2. Detailed architecture of the lightweight backbone used in our SegMarsViT.

	Layer	Output Size	Repeat	Channel		
				xxs	xs	s
Stage 1	Conv 3 × 3	$\frac{H}{2} \times \frac{W}{2}$	1	16	16	16
	MV2 block	$\frac{H}{2} \times \frac{W}{2}$	1	16	32	32
Stage 2	MV2 block	$\frac{H}{2} \times \frac{W}{2}$	1	24	48	64
	MV2 block	$\frac{H}{4} \times \frac{W}{4}$	2	24	48	64
Stage 3	MV2 block	$\frac{H}{4} \times \frac{W}{4}$	1	48	64	96
	MViT block	$\frac{H}{8} \times \frac{W}{8}$	1	48	64	96
Stage 4	MV2 block	$\frac{H}{8} \times \frac{W}{8}$	1	64	80	128
	MViT block	$\frac{H}{16} \times \frac{W}{16}$	1	64	80	128
Stage 5	MV2 block	$\frac{H}{16} \times \frac{W}{16}$	1	80	96	160
	MViT block	$\frac{H}{16} \times \frac{W}{16}$	1	80	96	160
	Conv 1 × 1	$\frac{H}{32} \times \frac{W}{32}$	1	320	384	512

Through leveraging the vision transformer to focus on modeling the global context at all stages, the proposed SegMarsViT can better establish long-range semantic relationships between feature representation. The capacity to model local–global context relationships of images would benefit to learn more abstract semantic visual concepts through enlarging the receptive field. Moreover, the mobile vision transformer is a lightweight and low-latency architecture, which can meet the requirements of accuracy and model complexity in practical satellite missions. The lightweight encoder, therefore, makes the network suitable for real-time applications, as it provides rich semantic information.

3.3. Efficient Layer Aggregation Decoder

In order to further model and fuse multi-level information from the feature encoder, we design an efficient layer aggregation decoder (ELAD) consisting of two primary elements: cross-scale feature fusion (CFF) module and compact feature aggregation (CFA) module in SegMarsViT, as shown in Figure 4. In ELAD, CFF modules are designed to interact multiscale information and strengthen the feature representation learning of lightweight backbone network, and the CFA module is conducted to efficiently aggregation multi-scale deep features and obtain the final segmentation results.

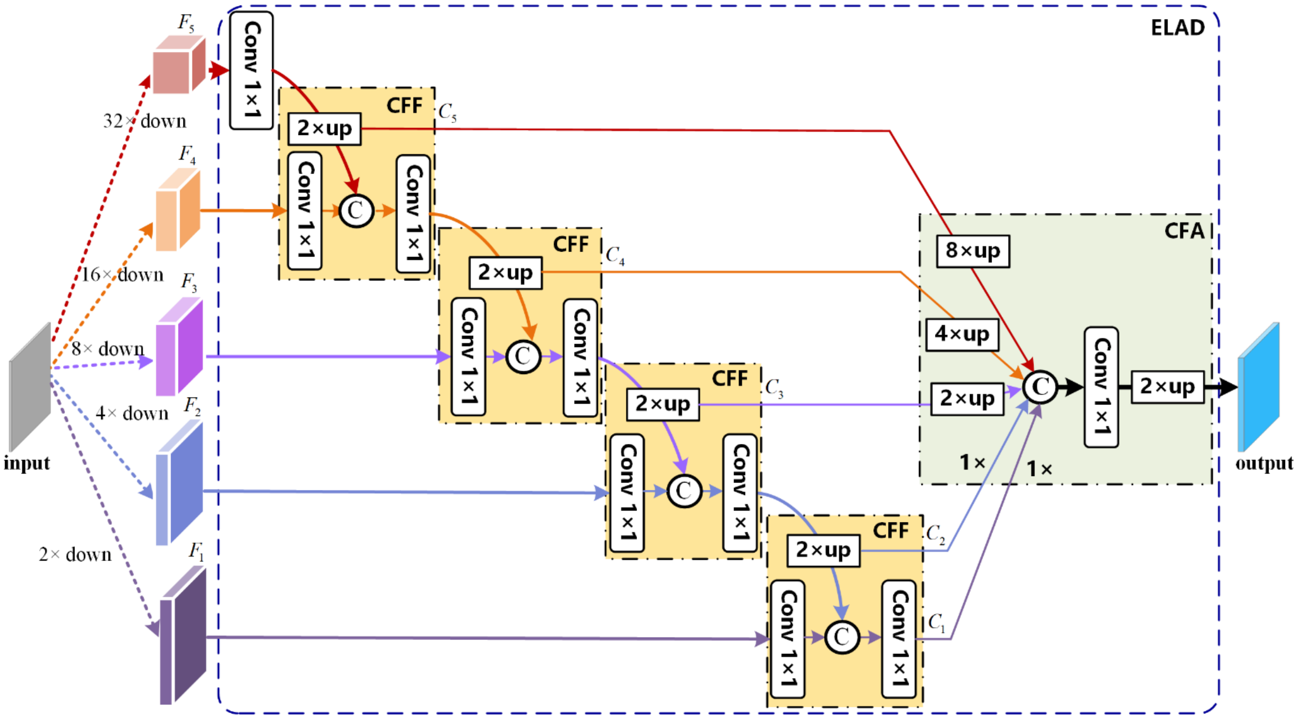


Figure 4. Illustration of the proposed Efficient Layer Aggregation Decoder.

- Cross-Scale Feature Fusion: The utilization of our CFF modules allows high-level context information to be delivered to multi-scale feature maps at different pyramid levels, each of which contains four sub-branches. As can be seen, following the top-down pathway, the input feature maps with coarser resolutions are firstly up-sampled by a factor of 2 to obtain C_i. Meanwhile, we utilize the 1×1 convolution layer for the feature maps in the lower level F_i to unify the channel dimension. Then the up-sampled feature maps C_i are concatenated and fused to the F_i. A 1×1 convolution layer is attached after fusion. Specifically, we have $C_1 = \text{Conv}_{1\times1}(\text{Conv}_{1\times1}(F_1) \oplus C_2)$, where $\text{Conv}_{1\times1}(\cdot)$ represents a 1×1 convolution and $\oplus$ denotes the concatenation operation. In this way, the proposed CFF modules assist our model to enlarge the receptive field through enabling each spatial location to view the local context in different scale spaces.

- Compact Feature Aggregation: After CFF, we perform multi-level feature integration for predicting segmentation maps with fine details. To accomplish multi-level feature fusion, we construct the compact feature aggregation (CFA) module. The output of CFF consists of five fusion maps. We first reshape the high-level feature maps $\{C_1, C_2, C_3\}$ to the same size as C_4 and C_5. Then, all these feature maps in the same spatial resolution are concatenated and followed by a 1 × 1 convolution for feature fusion. By this means, our lightweight decoder merges multi-level features from top to bottom till the segmentation map of size equal to input image is reconstructed.

3.4. Loss Function

We continue by introducing our loss function for optimizing the proposed SegMarsViT. Our loss function combines the weighted intersection over union (IoU) loss and the weighted binary cross-entropy (BCE) loss:

$$\mathcal{L}oss = \mathcal{L}^{\omega}_{\text{IoU}} + \mathcal{L}^{\omega}_{\text{BCE}} \tag{2}$$

where $\mathcal{L}^{\omega}_{\text{IoU}}$ and $\mathcal{L}^{\omega}_{\text{BCE}}$ represent the weighted IoU loss and BCE loss for the global restriction and local (pixel-level) restriction. Different from the standard IoU loss, which was widely adopted in segmentation tasks, $\mathcal{L}^{\omega}_{\text{IoU}}$ increases the weights of hard pixels to highlight their importance. In addition, compared with the standard BCE loss, $\mathcal{L}^{\omega}_{\text{BCE}}$ pays more attention to hard pixels rather than assigning all the pixels equal weights. The definitions of these losses are the same as in [45,46], and their effectiveness was validated in the field of semantic segmentation.

4. Results and Analysis

In this section, we first provide the experimental setup in the Section 4.1. Then the Section 4.2 presents the results achieved with our model and a comparison made with other segmentation models. In Section 4.3, we conduct comprehensive ablation studies.

4.1. Experimental Settings

4.1.1. Datasets

In order to demonstrate the proposed network's performance, we extensively evaluate our SegMarsViT on three publicly available MTS datasets, including AI4Mars-MSL, MSL-Seg, and S^5Mars. These three datasets consist of 17,030, 4155, and 6000 real images of Mars with corresponding pixel-level labels. We offer a brief view in Table 3.

- AI4Mars-MSL: AI4Mars is the first large-scale semantic segmentation dataset build for terrain-aware autonomy on Mars contains 17,000 images with a spatial resolution of 1024 × 1024, which consists of 3-band RGB images taken by NASA's Mars Science Laboratory (MSL). It contains four classes: Soil, Bedrock, Sand and Big Rock.
- MSL-Seg: The MSL-Seg dataset contains 4184 images with the size of 560 × 500, which consists of 3-band RGB images from the mars32k dataset (available at https://dominikschmidt.xyz/mars32k/ (accessed on 20 February 2022)). It contains eight categories: Martian soil, Sands, Gravel, Bedrock, Rocks, Tracks, Shadows, and Background.
- S^5Mars: The S^5Mars dataset contains 6000 images with a spatial resolution of 1200 × 1200, which are collected by the color mast camera (Mastcam) from the Curiosity rover on Mars. Different from AI4Mars-MSL and MSL-Seg, the overall annotations in S^5Mars are employed in a sparse style, which only the pixels with enough human confidence are labeled. It contains nine classes: Sky, Ridge, Soil, Sand, Bedrock, Rock, Rover, Trace, and Hole.

Table 3. Statistics of experimental datasets in this research.

Dataset	Classes	Annotated Images	Image Size	Split
AI4Mars-MSL	4	17,030	1024 × 1024	16,064:322:322
MSL-Seg	8	4155	560 × 500	2893:827:414
S^5Mars	9	6000	1200 × 1200	5000:200:800

4.1.2. Implementation Details

We implement our experiments with the MMSegmentation [47] open source toolbox and Pytorch [48] accelerate training via NVIDIA GPUs. During training, we applied data augmentation operations through random mirror, random resize with ratio 0.5–2.0, random horizontal flipping, random rotation between -10 and 10 degrees and random Gaussian blur for all datasets. Particularly, we random crop to 512 × 512 for AI4Mars, S^5Mars, and MSL-Seg datasets. The proposed model was trained for 400 epochs with a mini-batch size of 16 over 4 GPUs RTX2080Ti. We use the SGD optimizer with the initial learning rate (LR) $1e^{-2}$. The polynomial LR policy [49] was used to update the learning rate and help the model in faster convergence for improving performance.

4.1.3. Evaluation Metrics

For all experiments, we run the same training recipe three times and report several widely used metrics, such as the mean of class-wise intersection over union (mIoU), pixel-wise accuracy (pixelACC), the mean of F1 score (mFscore), and the mean of precision value (mPrecision).

4.2. Comparison with SOTA Methods

In this paper, we compared the proposed SegMarsViT with existing lightweight semantic segmentation methods. We evaluate SegMarsViT against eight SOTA natural image semantic segmentation methods, including FCN [10], DeepLabV3+ [50], Segmenter [51], PSPNet [52], PSANet [53], SegFormer [38], and FPN-PoolFormer [54].

4.2.1. Results on AI4Mars-MSL

Table 4 summarizes our results including parameters, FLOPS and other accuracy metrics of different lightweight semantic segmentation methods on the AI4Mars-MSL dataset. Red, blue, and green denote the best, the second-best, and the third-best results, respectively. For AI4Mars-MSL, there is a relatively small amount of labeled terrain types. With the computing power constraint of available GPUs, we mainly report the results trained with a lightweight backbone. From the results, in comparison with several SOTA approaches, our proposed SegMarsViT outperforms most of them. As shown, on AI4Mars, SegMarsViT yields 68.4% mIoU using only 8.54 M parameters and 5.6 G FLOPs, achieving competitive results in contrast to all other real-time counterparts in terms of parameters and efficiency comprehensively. For instance, compared to SegFormer (MIT-B0), SegMarsViT keeps 0.66% better mIoU.

Table 4. Comparison with state-of-the-art methods on AI4Mars-MSL.

Method (PubYear)	Encoder	pixelACC	mIoU	FLOPs (G)	Params (M)
Segmenter (2021)	ViT-s	92.04	66.85	17.93	26.03
SegFormer (2021)	MIT-B0	92.76	67.74	6.39	3.72
FPN-PoolFormer (2022)	S12	92.72	67.79	30.69	15.64
FCN (2016)	MobileNetv2	92.27	67.12	39.6	9.8
PSPNet (2018)	MobileNetv2	92.41	66.58	52.94	13.72
DeepLabV3+ (2018)	MobileNetv2	91.17	62.04	69.4	15.35
PSANet (2018)	ResNet50	86.83	54.6	194.8	54.07
SegMarsViT (Ours)	MobileViT-s	92.46	68.4	8.54	5.61

4.2.2. Results on S^5Mars

Here, we show both quantitative and qualitative results on S^5Mars. Table 5 shows the comparative results on the test set of S^5Mars. We achieve 78.22% in terms of mIoU, with the standard MobileViT structure as the backbone. The depicted results demonstrate that our model outperforms most of current real-time semantic segmentation works. Figure 5 shows the visual comparison of Martian terrain segmentation methods on five examples from the S^5Mars dataset. The examples include a diverse scene context and backgrounds. The proposed methodology can achieve better or comparable performance in Martian terrain segmentation. What should be noted is that the samples Figure 5a,b are of scenarios in which rough and scattered terrains coexist. From the visual results, the proposed SegMarsViT has less false-positive detection. As for the samples (c) and (d), unstructured scene properties particularly stand out in the images. The proposed method can model contextual information well under the circumstance of unstructured scenes on Mars, which benefit from that ViT-based self-attention technique is applicable to explore spatial correlations. While other competitors may not detect the whole semantic objects or even not find some semantic objects in difficult scenarios, SegMarsViT can segment semantic regions with more accurate results. Especially in the boundary part, the loss of spatial details leads to the loss of accuracy. However, when the difference among foreground objects is relatively small, such as the Figure 5e, some missed detections occur in the results and there is still room for improvement.

Table 5. Comparison with state-of-the-art methods on S^5Mars.

Method	pixelACC	mIoU	mFscore	mPrecision	FLOPs(G)	Params (M)	FPS
Segmenter—ViT-s	90.99	77.15	84.21	85.4	17.93	26.03	48.44
SegFormer—MIT-B0	91.99	79.05	85.74	85.61	6.39	3.72	59.21
FPN—PoolFormer-s12	91.74	76.82	83.3	84.28	30.69	15.64	37.35
FCN—MobileNetv2	86.53	56.57	64.46	72.7	39.6	9.8	58.17
PSPNet—MobileNetv2	90.68	74.64	82.32	82.26	52.94	13.72	50.21
DeepLabV3+—MobileNetv2	89.64	69.5	78.22	80.94	69.4	15.35	37.64
FCN—HRNetv2-w18s	87.71	63.68	73.41	79.71	9.6	3.94	49.53
PSANet—ResNet50	89.11	72.18	80.78	81.96	194.8	54.07	17.64
SegMarsViT—MobileViT-s (Ours)	92.15	78.22	84.74	85.86	8.54	5.61	69.52

4.2.3. Results on MSL-Seg

Table 6 summarizes our results including FLOPS, frame per seconds (FPS), and other four metrics to evaluate the segmentation accuracy on the MSL-Seg dataset. Compared with other latest methods, the proposed SegMarsViT exhibited significant improvement of 2.96% and 5.17% in terms of the pixelACC and mIoU, respectively. We further analyze the classwise segmentation performance of the proposed SegMarsViT on eight classes, we obtain classwise IoU on the test dataset, and is shown in Table 7. It can be observed that IoU for few classes is low, e.g., the Martian soil and bedrocks. This is because the notion of these classes is ambiguous in MSL-Seg. Their low IoU score is due to the low pixel count of these objects in the training data.

Figure 6 shows the ground truth segmentation maps of five sample images along with their predicted segmentation maps. It can be observed that the proposed SegMarsViT has much better comparative results in scenes. As shown in the last two rows of Figure 6, our method can work well on several kinds of complex scenarios with noisy information, while others may fail in such scenarios. It can be seen from the overall detection effect that the main Martian terrain feature can be extracted. However, missed detections and error detections of some objects existed, and the segmentation accuracy needs to be further improved.

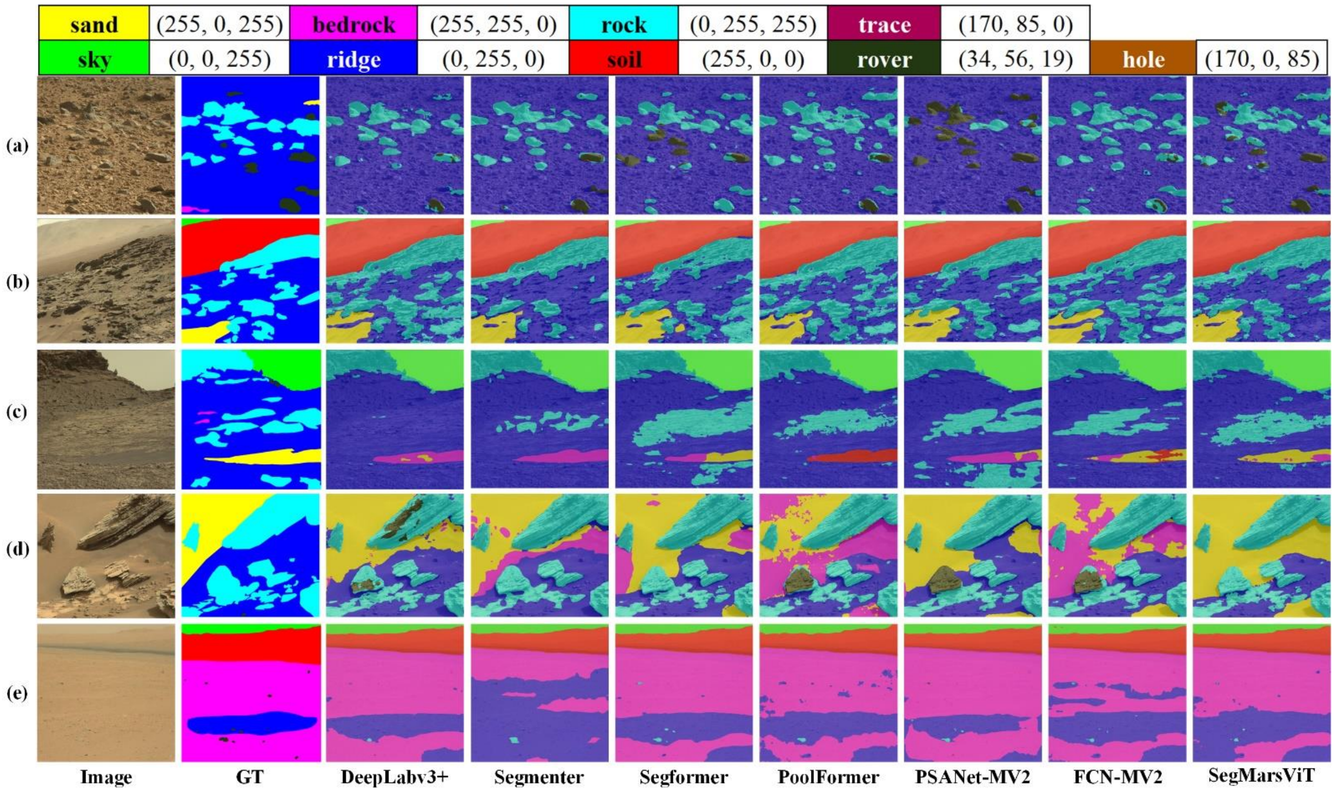


Figure 5. Qualitative comparison on S^{5}Mars test set. (**a**–**e**) show five experimental samples.

Table 6. Comparison with state-of-the-art methods on MSL-Seg.

Method	pixelACC	mIoU	mFscore	mPrecision	FLOPs (G)	Params (M)	FPS
Segmenter—ViT-s	84.39	66.2	78.32	76.4	17.93	26.03	48.44
SegFormer—MIT-B0	83.84	64.37	76.99	74.74	6.39	3.72	59.21
FPN—PoolFormer-s12	83.9	63.41	76.56	76.03	30.69	15.64	37.35
FCN—MobileNetv2	81.67	54.96	68.32	77.72	39.6	9.8	58.17
PSPNet—MobileNetv2	82.32	60.62	74.2	71.1	52.94	13.72	50.21
DeepLabV3+—MobileNetv2	82.47	59.08	72.78	70.56	69.4	15.35	37.64
FCN—HRNetv2-w18s	82.59	58.57	71.98	75.44	9.6	3.94	49.53
PSANet—ResNet50	83.23	62.43	75.41	73.47	194.8	54.07	17.64
SegMarsViT—MobileViT-s (Ours)	86.05	67.28	78.69	78.75	8.54	5.61	69.52

Table 7. Classwise IoU of the proposed SegMarsViT on MSL-Seg dataset.

Martian Soil	Sands	Gravel	Bedrock	Rocks	Tracks	Shadows	Unknown	mIoU
41.3	77	82.43	47.91	74.39	54.22	89.23	71.78	67.28

To further analyze the model efficiency, we summarize the efficiency-related metrics on the three datasets mentioned above and state them in Figure 7. As shown, the proposed SegMarsViT has the fewest parameters among all the models. These metrics are crucial for Martian terrain segmentation on satellite, which has limited storage. Here, frames per second (FPS) is an average speed that per second with size 512 × 512. Data and parameters load time is not considered, and the employed single GPU is NVIDIA 3070Ti with 8-G storage. The time spent per image of our SegMarsViT is less than other semantic segmentation methods. In conclusion, our method achieves the state-of-the-art performance in Martian terrain segmentation and meanwhile is much more efficient than methods with comparable accuracy.

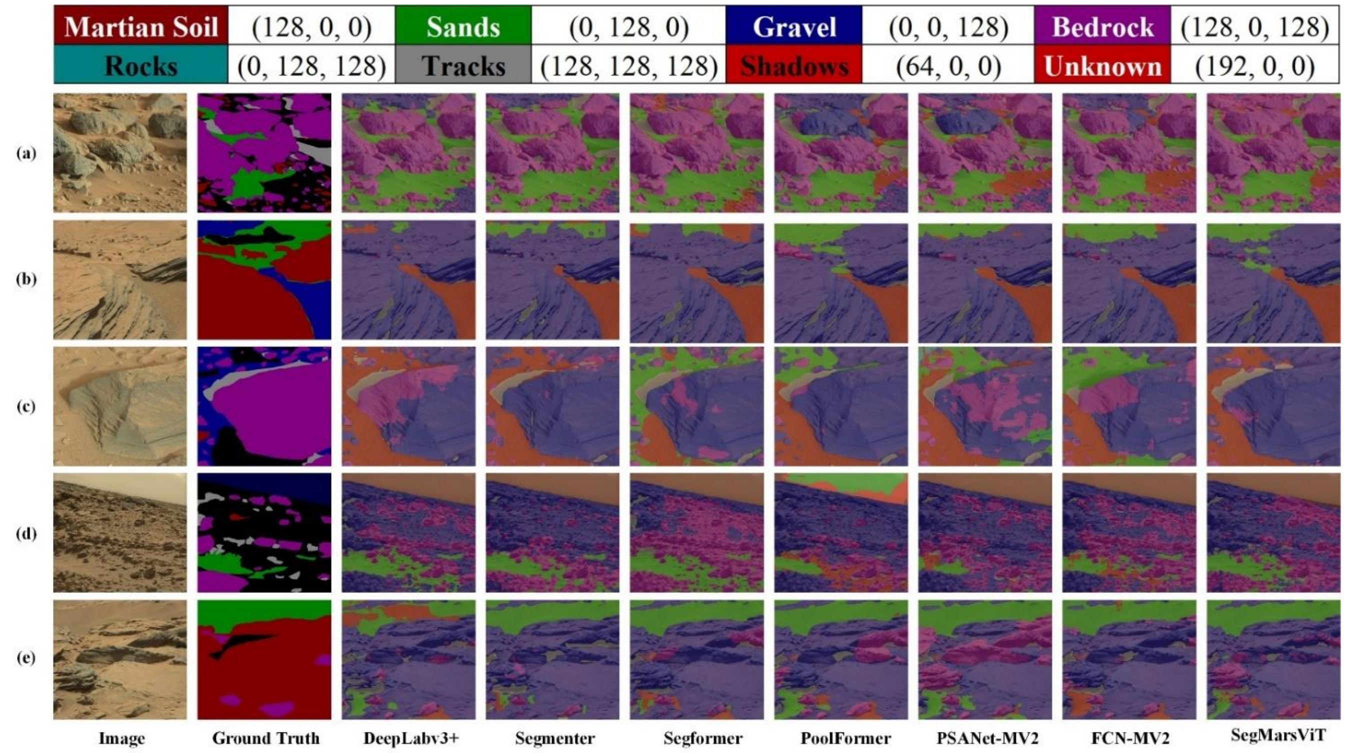


Figure 6. Visualization examples on MSL-Seg test set. (**a**–**e**) represent five experimental samples.

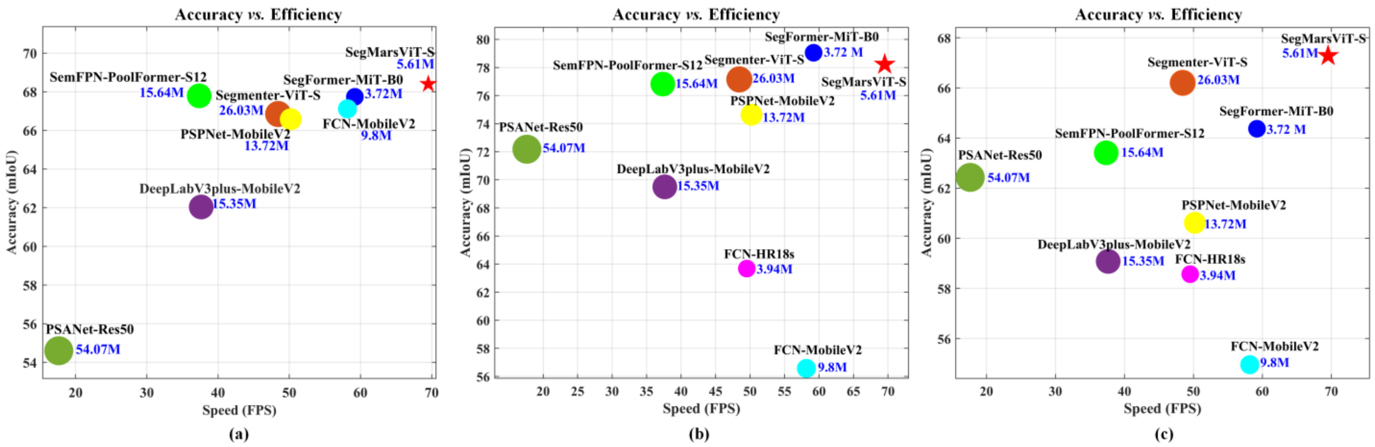


Figure 7. Speed and accuracy comparison on three datasets, (**a**) AI4Mars; (**b**) S^{5}Mars; (**c**) MSL-Seg. Compared with other regular semantic segmentation methods, the proposed SegMarsViT is competitive.

4.3. Ablation Studies

4.3.1. Effect of Backbone

We first analyze the effect of increasing the size of the encoder on the performance and model efficiency. MobileViT-xxs, MobileViT-xs, and MobileViT-s are the series of mobile transformer encoders with the same architecture but different sizes (as illustrated in Table 2 of Section 3.2). Table 8 summarizes the comparison results for three datasets. It can be observed that both the largest model SegMarsViT-s and the super small model SegMarsViT-xxs achieve close to or exceeding SOTA performance. Furthermore, the super small model SegMarsViT-xxs has good performance, and the number of parameters and FLOPs are 1.84 M and 1.16 G, along with 66.81%, 74.83%, and 65.80% mIoU on the three datasets, respectively. Because the model parameters of the backbone network are smaller and the structure is compact, the pressure on computing resources is smaller. Hence, the proposed model can be better applied to engineering.

Table 8. Evaluation of encoder with different model sizes for SegMarsViT.

Encoder	**Complexity**			**AI4Mars**		**S^5Mars**		**MSL-Seg**	
	FLOPs	**Params**	**FPS**	**pixelACC**	**mIoU**	**pixelACC**	**mIoU**	**pixelACC**	**mIoU**
MobileViT-xxs	1.16 G	1.84 M	110.1	91.92	66.81	89.34	74.83	83.17	65.80
MobileViT-xs	2.23 G	4.61 M	80.3	91.99	67.73	91.58	76.32	84.35	66.83
MobileViT-s	8.54 G	5.61 M	69.5	92.46	68.4	92.15	78.22	86.05	67.28

4.3.2. Effect of ELAD

In this subsection, we test the proposed ELAD with different decoders. As mentioned earlier, the proposed efficient layer aggregation decoder (ELAD) consists of stagewise CFF modules and one CFA module in a nutshell, which are constructed in the way shown in Figure 8. In addition, we select two representative decoders (Figure 9a,b) for test: the All-MLP decoder, termed AMD, first proposed in SegFormer [38], and the classic decoder of the U-shaped network [11], termed UNetD. In practice, we use the official code provided by the authors to implement our experiments. From Table 9, with the same encoder, e.g., MobileViT-s encoder, we find that the proposed ELAD produces higher performance.

Compared with AMD, the proposed ELAD achieves through introducing the CFF modules to build internal communications between the adjacent feature stages. The experimental results in Table 9 verify the significant effect of our CFFs on better fusing feature maps at different scales from another perspective, and this is exactly the common point of ELAD and UNetD. Both consist of an information fusion path for modeling a more representative and robust context. The key difference is the way they implement feature fusion across adjacent stages. The comparison results on Table 9 show that our ELAD has the least FLOPs with comparable parameters, which are the vital part of the construction for the lightweight segmentation network.

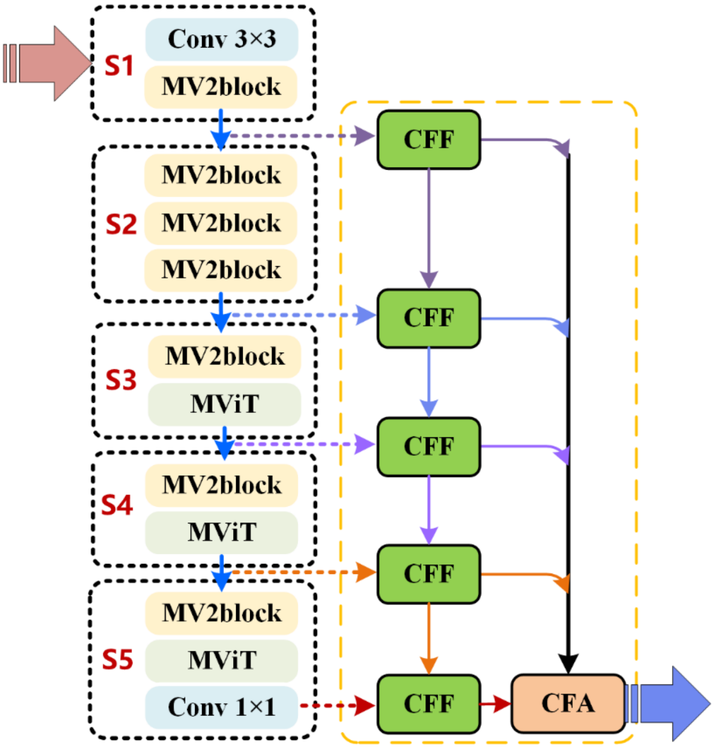
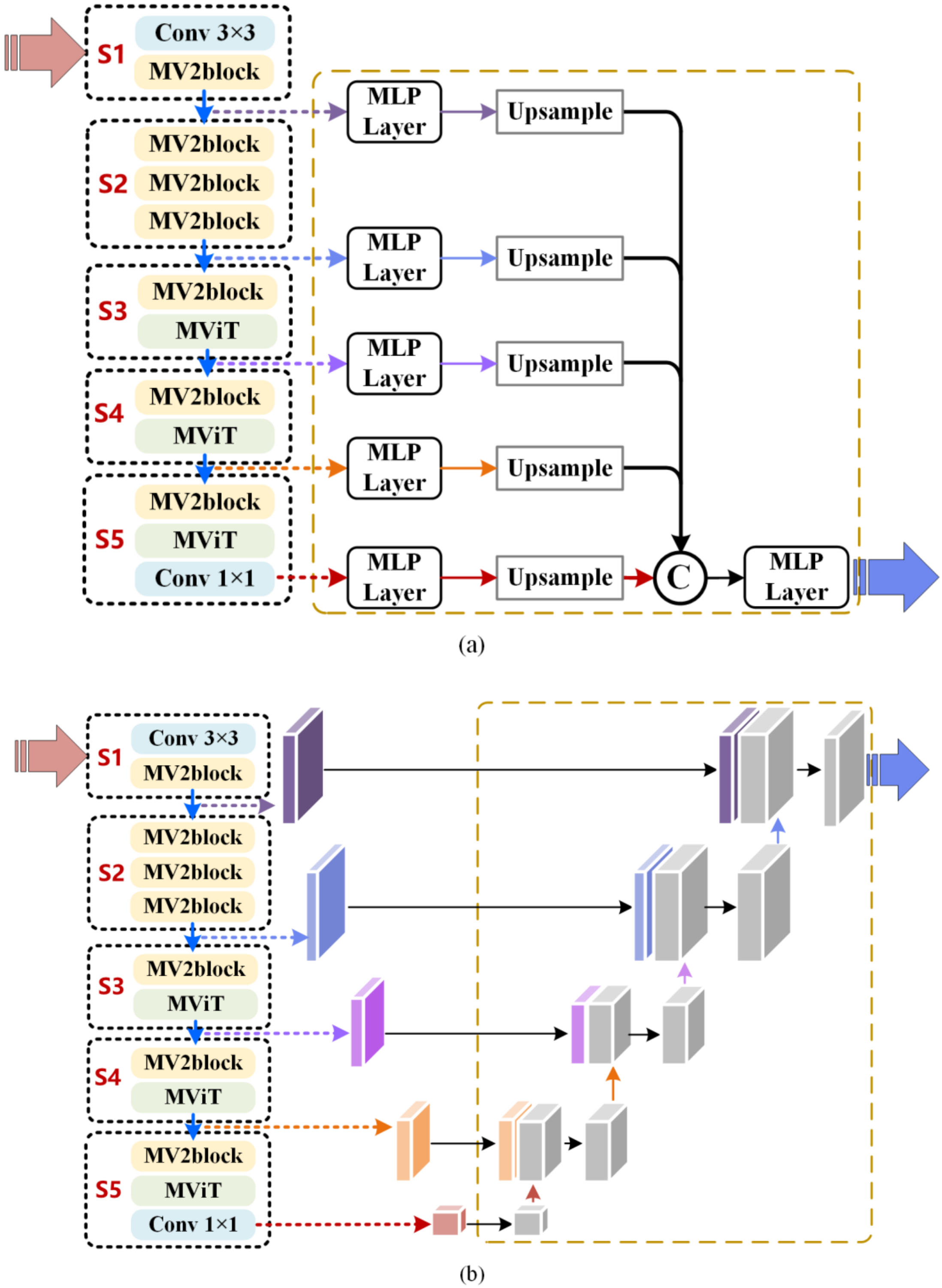


Figure 8. Illustrating the ELAD architecture of the proposed SegMarsViT.



(a)

(b)

Figure 9. Illustration of different decoder architectures: (**a**) All-MLP Decoder, (**b**) UNet Decoder.

Table 9. Ablation studies for decoder on MSL-Seg dataset.

Encoder	Decoder	pixelACC	mIoU	FLOPs(G)	Params (M)
MobileViT-s	AMD	85.66	66.69	13.14	5.09
MobileViT-s	UNetD	85.48	66.83	12.85	6.23
MobileViT-s	ELAD	86.05	67.28	8.54	5.61

4.3.3. Effect of Components in Decoder

In this subsection, we design ablation experiments on SegMarsViT to examine the validity of CFF and CFA modules.

(1) Baseline: MobileViT-s + ELAD (Without CFA).
(2) SegMarsViT: MobileViT-s + ELAD (CFF + CFA).

Table 10 reports the ablation studies of the baseline and our model on the MSL-Seg test set. We can see that incorporating both the CFF and CFA modules results in consistent and significant increase over the baseline. In particular, when compared with the baseline model, mIoU and PixelACC of the SegMarsViT with both CFF and CFA blocks integrated are improved by 4.06% and 2.62%, respectively. The great improvement of SegMarsViT proves the gain effect of their combination.

Table 10. Ablation results on MSL-Seg dataset.

Method	**Modules**		**pixelACC**	**mIoU**	**FLOPs(G)**	**Params (M)**
	CFF	**CFA**				
Baseline	-	✓	83.43	63.22	7.54	5.0
SegMarsViT	✓	✓	86.05	67.28	8.54	5.61

5. Conclusions

In this paper, we propose SegMarsViT, a lightweight network for the real-time Martian terrain segmentation task. We adopt a deployment-friendly MobileViT backbone to extract discriminative local–global context information from multi-scale feature space. Further, an effective cross-scale feature fusion module was designed to encode context information in the multi-level features, with a cross-scale feature fusion mechanism applied to help further aggregate feature representations. In the end, a compact prediction head is used to aggregate hierarchical features and help enhance feature learning, yielding run-time efficiency. Empirical results validate the superiority of the proposed SegMarsViT over mainstream semantic segmentation methods. The ablation study verifies the effectiveness of each module. More specifically, MobileViT helps obtain the semantic properties of terrain objects in terms of morphology and distribution, while the compact decoder can lead to both high efficiency and performance. Through the comparison of parameters, FLOPs and FPS, the SegMarsViT further demonstrates its advantages in terms of space and computation complexity. All of the results fully demonstrate the capability of the SegMarsViT in efficient and effective Martian terrain segmentation, which provides significant potential for the further development of MTS task.

One potential limitation is that there's an enormous gap between high-end GPU and a low-memory spacecraft device. Our future work will experiment on a realistic hardware platform to evaluate the model efficiency. Energy consumption and practical performance will be our primary focus. Moreover, we will proceed to refine our approach and be committed to investigate MTS methods for more challenging cases, e.g., multi-source heterogeneous data and a multi-task perception system.

Author Contributions: Conceptualization, Y.D., C.X. and L.Z.; funding acquisition, C.X.; investigation, Y.D., T.Z. and C.X.; methodology, Y.D.; writing—original draft, Y.D.; writing—review and editing, T.Z., C.X. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chinese Academy of Sciences Project, grant number: CXJJ-20S017.

Data Availability Statement: Not applicable.

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

References

1. Cakir, S.; Gauß, M.; Häppeler, K.; Ounajjar, Y.; Heinle, F.; Marchthaler, R. Semantic Segmentation for Autonomous Driving: Model Evaluation, Dataset Generation, Perspective Comparison, and Real-Time Capability. *arXiv* **2022**, arXiv:2207.12939.
2. Csurka, G.; Perronnin, F. A Simple High Performance Approach to Semantic Segmentation. In Proceedings of the BMVC, Leeds, UK, 1 September 2008; pp. 1–10.
3. Corso, J.J.; Yuille, A.; Tu, Z. Graph-Shifts: Natural Image Labeling by Dynamic Hierarchical Computing. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
4. Holder, C.J.; Shafique, M. On Efficient Real-Time Semantic Segmentation: A Survey. 19. *arXiv* **2022**, arXiv:2206.08605.
5. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A Review on Deep Learning Techniques Applied to Semantic Segmentation 2017. *arXiv* **2017**, arXiv:1704.06857.
6. 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.
7. Ronneberger, O.; Fischer, P.; Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*; Springer: Cham, Switzerland, 2015.
8. McGlinchy, J.; Johnson, B.; Muller, B.; Joseph, M.; Diaz, J. Application of UNet Fully Convolutional Neural Network to Impervious Surface Segmentation in Urban Environment from High Resolution Satellite Imagery. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 3915–3918.
9. Sun, J.; Shen, J.; Wang, X.; Mao, Z.; Ren, J. Bi-Unet: A Dual Stream Network for Real-Time Highway Surface Segmentation. *IEEE Trans. Intell. Veh.* **2022**, *15*. [CrossRef]
10. Chattopadhyay, S.; Basak, H. Multi-Scale Attention u-Net (Msaunet): A Modified u-Net Architecture for Scene Segmentation. *arXiv* **2020**, arXiv:2009.06911.
11. Chu, Z.; Tian, T.; Feng, R.; Wang, L. Sea-Land Segmentation with Res-UNet and Fully Connected CRF. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 3840–3843.
12. Rothrock, B.; Kennedy, R.; Cunningham, C.; Papon, J.; Heverly, M.; Ono, M. SPOC: Deep Learning-Based Terrain Classification for Mars Rover Missions. In Proceedings of the AIAA SPACE 2016, American Institute of Aeronautics and Astronautics, Long Beach, CA, USA, 13–16 September 2016.
13. Iwashita, Y.; Nakashima, K.; Stoica, A.; Kurazume, R. Tu-Net and Tdeeplab: Deep Learning-Based Terrain Classification Robust to Illumination Changes, Combining Visible and Thermal Imagery. In Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 28–30 March 2019; pp. 280–285.
14. Liu, H.; Yao, M.; Xiao, X.; Cui, H. A Hybrid Attention Semantic Segmentation Network for Unstructured Terrain on Mars. *Acta Astronaut.* **2022**, *in press*. [CrossRef]
15. Claudet, T.; Tomita, K.; Ho, K. Benchmark Analysis of Semantic Segmentation Algorithms for Safe Planetary Landing Site Selection. *IEEE Access* **2022**, *10*, 41766–41775. [CrossRef]
16. Wang, W.; Lin, L.; Fan, Z.; Liu, J. Semi-Supervised Learning for Mars Imagery Classification and Segmentation. *arXiv* **2022**, arXiv:2206.02180. [CrossRef]
17. Goh, E.; Chen, J.; Wilson, B. Mars Terrain Segmentation with Less Labels. *arXiv* **2022**, arXiv:2202.00791.
18. Zhang, J.; Lin, L.; Fan, Z.; Wang, W.; Liu, J. S^{5}Mars: Self-Supervised and Semi-Supervised Learning for Mars Segmentation. *arXiv* **2022**, arXiv:2207.01200.
19. Li, J.; Zi, S.; Song, R.; Li, Y.; Hu, Y.; Du, Q. A Stepwise Domain Adaptive Segmentation Network with Covariate Shift Alleviation for Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3152587. [CrossRef]
20. Swan, R.M.; Atha, D.; Leopold, H.A.; Gildner, M.; Oij, S.; Chiu, C.; Ono, M. AI4MARS: A Dataset for Terrain-Aware Autonomous Driving on Mars. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 19–25 June 2021; pp. 1982–1991.
21. Dai, Y.; Xue, C.; Zhou, L. Visual Saliency Guided Perceptual Adaptive Quantization Based on HEVC Intra-Coding for Planetary Images. *PLoS ONE* **2022**, *19*, e0263729. [CrossRef]
22. Tian, Y.; Chen, F.; Wang, H.; Zhang, S. Real-Time Semantic Segmentation Network Based on Lite Reduced Atrous Spatial Pyramid Pooling Module Group. In Proceedings of the 2020 5th International Conference on Control, Robotics and Cybernetics (CRC), Wuhan, China, 16 October 2020; pp. 139–143.
23. Liu, R.; Tao, F.; Liu, X.; Na, J.; Leng, H.; Wu, J.; Zhou, T. RAANet: A Residual ASPP with Attention Framework for Semantic Segmentation of High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3109. [CrossRef]
24. Li, G.; Yun, I.; Kim, J.; Kim, J. DABNet: Depth-Wise Asymmetric Bottleneck for Real-Time Semantic Segmentation. *arXiv* **2019**, arXiv:1907.11357.
25. Mehta, S.; Rastegari, M.; Caspi, A.; Shapiro, L.; Hajishirzi, H. Espnet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 552–568.
26. Siam, M.; Gamal, M.; Abdel-Razek, M.; Yogamani, S.; Jagersand, M. Rtseg: Real-Time Semantic Segmentation Comparative Study. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 1603–1607.

27. Li, Y.; Li, X.; Xiao, C.; Li, H.; Zhang, W. EACNet: Enhanced Asymmetric Convolution for Real-Time Semantic Segmentation. *IEEE Signal Process. Lett.* **2021**, *28*, 234–238. [CrossRef]
28. 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.
29. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
30. Yang, Y.; Jiao, L.; Liu, X.; Liu, F.; Yang, S.; Feng, Z.; Tang, X. Transformers Meet Visual Learning Understanding: A Comprehensive Review. *arXiv* **2022**, arXiv:2203.12944.
31. Ye, L.; Rochan, M.; Liu, Z.; Wang, Y. Cross-Modal Self-Attention Network for Referring Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10502–10511.
32. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2021**, arXiv:2010.04159.
33. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. *arXiv* **2021**, arXiv:2102.12122.
34. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. *arXiv* **2021**, arXiv:2012.15840.
35. Huang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.-H.; Chen, Y.-W.; Tong, R. ScaleFormer: Revisiting the Transformer-Based Backbones from a Scale-Wise Perspective for Medical Image Segmentation. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 23–29 July 2022; pp. 964–971.
36. Shi, W.; Xu, J.; Gao, P. SSformer: A Lightweight Transformer for Semantic Segmentation. *arXiv* **2022**, arXiv:2208.02034.
37. 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, Nashville, TN, USA, 19–25 June 2021; pp. 10012–10022.
38. 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**, *14*, 12077–12090.
39. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: An UNet-like Transformer for Efficient Semantic Segmentation of Remotely Sensed Urban Scene Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
40. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video Swin Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3202–3211.
41. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. Unetr: Transformers for 3d Medical Image Segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, New Orleans, LA, USA, 18–24 June 2022; pp. 574–584.
42. Zhou, H.-Y.; Guo, J.; Zhang, Y.; Yu, L.; Wang, L.; Yu, Y. Nnformer: Interleaved Transformer for Volumetric Segmentation. *arXiv* **2021**, arXiv:2109.03201.
43. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
44. 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 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
45. Wu, Y.-H.; Liu, Y.; Xu, J.; Bian, J.-W.; Gu, Y.-C.; Cheng, M.-M. MobileSal: Extremely Efficient RGB-D Salient Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 10261–10269. [CrossRef]
46. Wu, Y.-H.; Liu, Y.; Zhang, L.; Cheng, M.-M.; Ren, B. EDN: Salient Object Detection via Extremely-Downsampled Network. *IEEE Trans. Image Process.* **2022**, *31*, 3125–3136. [CrossRef]
47. Contributors, Mms. MMSegmentation: Openmmlab Semantic Segmentation Toolbox and Benchmark. 2020. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 18 May 2022).
48. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–12.
49. Mishra, P.; Sarawadekar, K. Polynomial Learning Rate Policy with Warm Restart for Deep Neural Network. In Proceedings of the TENCON 2019—2019 IEEE Region 10 Conference (TENCON), Kochi, India, 17–20 October 2019; pp. 2087–2092.
50. 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.
51. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 7262–7272.
52. 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.

53. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-Wise Spatial Attention Network for Scene Parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
54. Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; Yan, S. MetaFormer Is Actually What You Need for Vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.