EFFICIENT SEMANTIC SEGMENTATION OF MULTISPECTRAL LAND COVER IMAGES USING MASK2FORMER

Pablo Canosa[1], *Álvaro Ordóñez*[1,2], *Dora B. Heras*[1,2], *Francisco Argüello*[2]

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain.
[2] Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Spain.

ABSTRACT

Semantic segmentation for EO is a process that involves assigning a specific label or category to each pixel in an image, enabling precise analysis for land cover applications such as environmental conservation, urban planning or disaster management. Deep learning-based segmentation models have proliferated in recent years, but they often are not well adapted to the unique properties of multi and hyperspectral images, frequently used in remote sensing. Mask2Former is a universal segmentation model based on the concept of masked attention and employs a pretrained classification model as backbone to create intermediate representations. This article presents a preliminary adaptation of Mask2Former for the segmentation of multispectral remote sensing images. This adaptation includes modifying the backbone to accept multispectral inputs and adapting the data processing pipelines to leverage all available spectral bands effectively. The computational cost of the method has also been analyzed as an initial assessment of potential scalability and efficiency for large-scale applications. Experimental results using the FiveBillionPixels dataset reveal a notable improvement in segmentation accuracy when incorporating multispectral bands, outperforming RGB-only performance without a relevant increase in computational cost.

***Index Terms*—** land cover, transformer, semantic segmentation, multispectral, computational cost.

1. INTRODUCTION

Multispectral images capture details beyond the visible spectrum, allowing for enhanced analysis of materials and environmental conditions. In remote sensing, image segmentation is widely used for object identification, visual interpretation, and analysis. It is particularly effective for detecting and classifying land cover classes. The goal is to assign a label to every pixel, effectively segmenting the image into different regions based on the objects or areas they represent. Depending on the semantics involved, segmentation tasks are generally classified into three types: instance, semantic, and panoptic.

Recent advances in multispectral segmentation [1] resulted in the adaptation of widely used architectures such as U-Net [2]. More recent developments explore dual-stream networks for multispectral fusion [3] and transformer models, such as SWIN, tailored for multispectral data [4].

In this context, Mask2Former [5] emerges as powerful transformer-based model for universal segmentation of images with 3 bands (RGB), enabling instance, semantic, and panoptic segmentation with a single pretrained unified model. This article presents an initial approach to adapting the Mask2Former model for semantic segmentation of multispectral images, and analyzing the computational cost of the method.

2. MASK2FORMER

In this section the original Mask2Former model is described. Mask2Former, illustrated in Fig.1, consists of three main components: backbone, pixel decoder, and transformer decoder.

The first component of Mask2Former is the backbone. This is the initial feature extraction network that processes the raw input data and generates a set of low-resolution feature representations for further processing by the transformer layers. The backbone is pretrained over ImageNet-1k [6]. For the experiments in this paper, two different residual network models have been selected as backbones: ResNet50 and ResNet101 [7]. In both cases a four-stage structure is considered. At the end of each stage, intermediate representations are stored. After completing the four stages, the backbone produces four sets of features with resolutions $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$ of the input image size. The difference between the two ResNet models lies in their depth. ResNet101 incorporates a higher number of residual blocks per stage, making

This work was supported in part by grants PID2022–141623NB–I00 and TED2021–130367B–I00 funded by MCIN/AEI/10.13039/501100011033 and by "European Union NextGenerationEU/PRTR". It was also supported by Xunta de Galicia - Consellería de Cultura, Educación, Formación Profesional e Universidades [Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04 and Reference Competitive Group accreditation, ED431C-2022/16], and by "ERDF/EU".

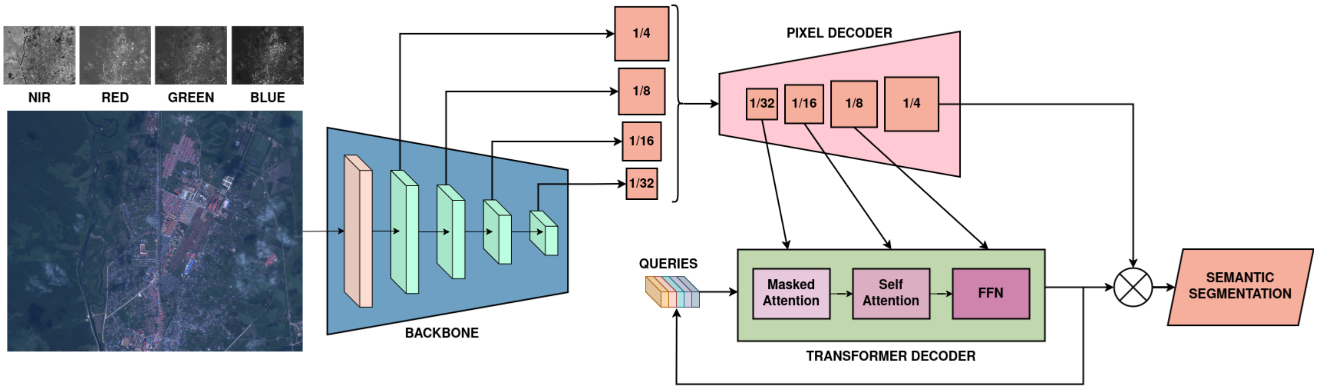

Fig. 1. Diagram of the Mask2Former model.

it a deeper and more complex model compared to ResNet50. This increased complexity allows ResNet101 to capture more intricate patterns at the cost of higher computational requirements.

The second component in Fig. 1 is the pixel decoder, whose objective is producing high-resolution per-pixel embeddings. Its architecture has been maintained as in the original Mask2Former and consists of a Feature Pyramid Network (FPN). At each level of the pyramid, the multi-scale deformable attention encoder [8] is employed, enhancing pixel-level feature processing. The pixel decoder utilizes the low-resolution feature maps generated by the backbone to produce a set of multiscale features. These features effectively capture both fine-grained details and global context.

The last component of Mask2Former, also shown in Fig. 1, is the transformer decoder, which is designed to process learnable queries and generate the final segmentation output. This component is divided into three blocks: the masked attention block, the self attention block and a Feed Forward Network (FFN). The transformer decoder operates using a fixed number of object queries, which are randomly initialized as learnable embeddings. Each query represents a potential object, region, or mask that the model aims to detect or segment. These object queries offer flexibility by enabling each query to focus on a specific region or object within the image. Moreover, they provide scalability, as the number of queries can be adjusted according to the complexity or nature of the images being processed, making the framework adaptable to a wide range of segmentation tasks.

The complexity of the transformer decoder, and consequently its computational cost, is influenced by two factors: the number of queries and the number of sequential layers (L). A layer, represented by the green box in Fig. 1, processes the object queries through the three main blocks: masked attention, self-attention, and the FFN. Depending on the number of layers, several blocks like those shown in the figure are sequentially executed.

Inside each layer of the transformer decoder the masked attention block plays a crucial role in the segmentation, as it introduces masked attention. Instead of attending to the entire feature map, the transformer decoder uses intermediate mask predictions to restrict attention to relevant spatial locations. This improves computational efficiency and ensures that each query focuses on the most informative regions, leading to more precise segmentation mask predictions. The three lower-resolution embeddings from the pixel decoder are processed through the transformer decoder, where object queries refine their representations. The output is then combined with the highest resolution embedding from the pixel decoder to produce the final segmentation mask.

3. MASK2FORMER FOR MULTISPECTRAL IMAGES

This section describes the adaptation of Mask2Former for semantic segmentation of multispectral images to operate using the Near-Infrared (NIR) band in addition to the RGB bands of the original model. Mask2Former was adapted to handle the data formats and metadata of multispectral images, which required modifications in dataset handling, model initialization, and training configuration. Specialized data mappers were developed to preprocess the multispectral inputs accurately, ensuring that the model receives consistent and relevant information.

Mask2Former processes images in the PIL Image format. To effectively use the multispectral data, the image loading methods were modified to directly utilize the RAW data, thus avoiding unnecessary decompression overheads. Moreover, Detectron2 [9], the modular library for object detection and segmentation used by Mask2Former, requires the dataset metadata in a specific format. To fullfill this requirement, an initial registration stage that iterates through each image in the dataset is carried out.

In terms of model architecture, the input handling of the backbone was modified. ResNet50 and ResNet101 [7] were selected as backbone networks, as they have demonstrated good performance and can be adapted to multispectral inputs. The size of the input was modified from the initial $3 \times height \times width$ to $4 \times height \times width$ in order to introduce an additional spectral band. In particular, the initial STEM block of the backbone, responsible for applying a 7×7 convolution with 64 filters and a stride of 2, was adjusted to accommodate the extra spectral channel. This approach enables the use of pretrained weights with minimal architectural changes, leveraging the information in the additional spectral band effectively.

The backbone models used, ResNet50 and ResNet101, were pretrained on the ImageNet-1k dataset [6]. This dataset consists of 1,281,167 training images, 50,000 validation images and 100,000 test images. Each image belongs to one of the 1,000 classes present in the dataset. Both pretrained models can be accessed through the Detectron2 repository.

4. EXPERIMENTS

This section evaluates the segmentation capabilities of Mask2Former adapted for multispectral images, alongside its memory requirements and the execution times using 1 and 2 GPUs.

4.1. Experimental Setup

The Five Billion Pixels dataset [10] is used to train and evaluate the proposed adaptation of Mask2Former for multispectral remote sensing images. This dataset comprises images captured by the Gaofen-2 satellite, provided in two formats: 16-bit non-quantized and 8-bit quantized, both featuring 4 spectral bands (red, green, blue, and near-infrared). For training, all images were divided into quarters, yielding a resolution of 3454×3650 pixels, as detailed in Table 1.

Table 1. Five Billion Pixels [10] dataset information (original dataset and adapted version).

	Original Five Billion Pixels	Used in this work
Categories	24 + Unlabeled	24 + Unlabeled
Image Size	6908 × 7300 pixels	3454 × 3650 pixels
Training	120 images	480 images
Test	30 images	120 images

The segmentation evaluation metrics are the same as those used in Common Objects in Context (COCO) [11], a widely recognized benchmark dataset and framework for object detection, segmentation, and captioning tasks. The segmentation performance was assessed using: Mean Intersection over Union (mIoU), Frequency Weighted IoU (fwIoU), Mean Accuracy (mACC), and Overall Accuracy (OA). These metrics were obtained by training over 60,000 iterations with a batch size of 2.

To evaluate computational cost, the FVCore library was employed to measure the number of model parameters and GFLOPs (Giga Floating Point Operations). Additionally, training times, measured as Training Time Per Epoch (TTPE), were calculated over 1,200 iterations with a batch size of 8, corresponding to a total of 20 epochs.

The models were trained on a PowerEdge R730 server running Ubuntu 22.04.4 LTS, equipped with two quad-core Intel Xeon E5-2623 v4 CPUs operating at 2.6 GHz and 128 GB of RAM. The system includes two Tesla P40 GPUs, each with 23 GB of memory, and utilizes the CUDA Toolkit 12.1. For the implementation, FVCore 0.1.5 was employed in combination with Detectron 0.6.

4.2. Experimental Results

The experiments evaluate the segmentation models utilizing all 4 spectral bands comparing them with the original Mask2Former model, which relies solely on the 3 RGB bands. Two backbones pretrained on the ImageNet-1k [6] dataset, ResNet50 and ResNet101, were evaluated.

Table 2 presents segmentation metrics as well as computational cost metrics (number of parameters and GFLOPs) for the 4 best models using ResNet50 or Resnet101 as backbone and 3 or 4 spectral bands. Execution times in terms of TTPE for the cases of execution using only 1 GPU and 2 GPUs with Distributed Data Parallel (DDP) are also shown, as well as the speedup achieved by using 2 GPUs. All models were trained using Automatic Mixed Precision (AMP) which combines both 16 bit and 32 bit float operations.

As shown in the table, R50 NirRGB (the model using a ResNet50 and the 4 bands Nir, R, G, and B) consistently outperforms the other models in segmentation performance metrics, indicating that the addition of the NIR band significantly improves the segmentation performance. Moreover, the best-performing models in terms of segmentation quality are all based on ResNet50, suggesting that it may generalize better than the ResNet101 backbone. This is particularly noteworthy given that ResNet50, in particular, the configuration using NirRGB has a lower TTPE than ResNet101 NirRGB as it has significantly fewer parameters (43.959×10^6 vs. 62.899×10^6) and requires lower GFLOPs (174.53 vs. 225.81).

More in detail, the R50 NirRGB model outperforms its RGB counterpart across all accuracy metrics, with higher OA (81.47% vs. 79.72%), mAcc (52.07% vs. 48.87%), and mIoU (41.30% vs. 39.98%). This highlights the significant performance increase provided by the inclusion of the NIR band without increasing computational complexity.

Regarding execution time, switching from RGB to multispectral images while using the same backbone does not, on average, increase the training time per epoch (TTPE), as shown in Table 3, but it results in improved segmentation performance metrics. The execution time for the R101 model is higher than for R50 in all the cases, being these results compatible with the higher values in GFLOPS and number of parameters of R101.

Table 2. Performance metrics and computational cost of the adapted Mask2Former model.

	Performance metrics (%)				**Computational cost**				
Model	**OA**	**mAcc**	**mIoU**	**fwIoU**	**Parameters (10^6)**	**GFLOPs**	**TTPE 1 GPU (s)**	**TTPE 2 GPUs (s)**	**Speedup**
R50 RGB	79.72	48.87	39.98	66.50	43.956	173.99	150.576	92.574	1.627
R50 NirRGB	**81.47**	**52.07**	**41.30**	**70.43**	**43.959**	**174.53**	**150.570**	**90.408**	**1.665**
R101 RGB	77.20	49.31	38.37	63.40	62.896	225.26	170.274	102.474	1.662
R101 NirRGB	76.40	42.62	33.67	63.67	62.899	225.81	170.076	99.510	1.709

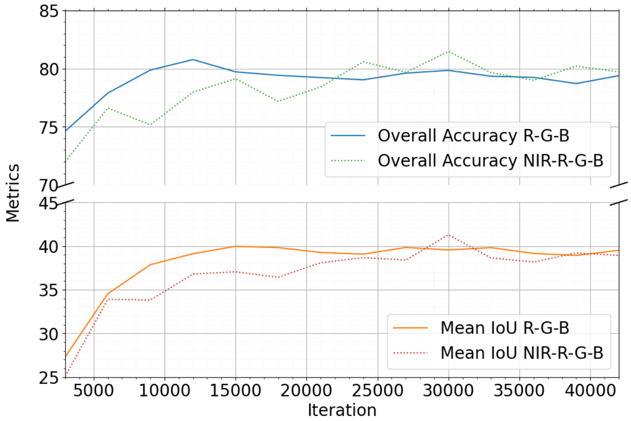

Fig. 2. Metrics comparison between the multispectral model (dotted line) and the RGB model (continuous line) using the ResNet50 backbone.

Table 3. Metrics comparison with previous work.

Method	**OA (%)**	**mAcc (%)**	**mIoU (%)**
CADR [12]	75.49	47.83	36.63
DPA [12]	75.36	48.57	37.77
CASSSS [12]	75.97	**53.53**	40.68
DeepLabv3+ [10]	**79.87**	-	**42.12**
U-Net [10]	**80.35**	-	**44.51**
Proposed R50 RGB	79.72	48.87	39.98
Proposed R50 NirRGB	**81.47**	**52.07**	**41.30**
Proposed R101 RGB	77.20	**49.31**	38.37
Proposed R101 NirRGB	76.40	42.62	33.67

To analyze the evolution in the segmentation process carried out by Mask2Former, Fig. 2 illustrates the evolution of the metrics for the R50 RGB and R50 NirRGB models during training. Initially, the RGB model performs better until around iteration 23,000. Beyond this point, the multispectral model incorporating the NIR band begins to outperform the RGB model. This improvement reflects the gradual adaptation of the model to the additional information contributed by the NIR band.

A comparative analysis against recent state-of-the-art methods for semantic segmentation was also conducted. The segmentation results over the Five Billion Pixels dataset using NirRGB (4-band) images are summarized in Table 3. The evaluation emphasizes metrics that provide deeper insights into the robustness of the segmentation process: OA, mAcc, and mIoU.

The proposed R50 NirRGB model performs better than the other proposed methods as it is shown by the higher OA values, indicating its ability to correctly classify the majority of pixels. The high values obtained for mAcc demonstrate balanced performance across categories. The results in terms of mIoU show efficiency in segmenting regions with minimal overlap errors.

Compared to the state-of-the-art models such as U-Net and DeepLabv3+, which also achieve high OA, the R50 NirRGB model offers a good balance between segmentation precision and per-class accuracy. It improves methods such as Advent and CASSSS in terms of OA while maintaining competitive mAcc and mIoU scores. The inclusion of NIR data in the R50 NirRGB model proves beneficial for distinguishing between visually similar or complex regions.

5. CONCLUSIONS

In this work, a first approach to adapt the Mask2former segmentation model to multispectral images has been presented and evaluated over images for land cover applications. The adaptation involves modifying Mask2Former to accommodate extra bands of the the multispectral. The objective is to ensure that the model fully utilizes the available spectral information to improve the segmentation accuracy. A first analysis of the computational cost of the new approach is also carried out.

It has been found that the addition of the NIR band improves the model learning process, leading to faster convergence, and better utilization of data compared to models trained on RGB images. The evaluation was performed by using pretrained ResNet50 and ResNet101 networks as backbone. The proposed model, based on ResNet50 and incorporating the NIR band, achieved competitive results compared to the state-of-the-art models at the cost of only a small increase in the number of parameters.

Given the competitive performance of this model, we propose the following future work directions: fine-tuning the backbone on multispectral land cover data to better adapt the feature extraction to this data type, modifying the pixel decoder to utilize higher-resolution embeddings adapted to high spatial resolution images, and testing over several spatial resolutions for the input images. Additionally, other alternatives will be explored to reduce the computational cost by exploiting parallelism thereby enabling the efficient handling of larger datasets and more complex models.

6. REFERENCES

[1] Leo Thomas Ramos and Angel D. Sappa, "Multispectral semantic segmentation for land cover classification: An overview," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 14295–14336, 2024.

[2] Halil Mertkan Sahin, Tajul Miftahushudur, Bruce Grieve, and Hujun Yin, "Segmentation of weeds and crops using multispectral imaging and CRF-enhanced U-Net," *Computers and Electronics in Agriculture*, vol. 211, pp. 107956, 2023.

[3] Yujia Fu, Xiangrong Zhang, and Mingyang Wang, "DSHNet: A semantic segmentation model of remote sensing images based on dual stream hybrid network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 4164–4175, 2024.

[4] Xuanyu Zhou, Lifan Zhou, Shengrong Gong, Shan Zhong, Wei Yan, and Yizhou Huang, "Swin transformer embedding dual-stream for semantic segmentation of remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 175–189, 2024.

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," 2021.

[9] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[10] Xin-Yi Tong, Gui-Song Xia, and Xiao Xiang Zhu, "Enabling country-scale land cover mapping with meter-resolution satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 178–196, 2023.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár, "Microsoft COCO: Common Objects in Context," 2015.

[12] Runmin Dong, Lichao Mou, Mengxuan Chen, Weijia Li, Xin-Yi Tong, Shuai Yuan, Lixian Zhang, Juepeng Zheng, Xiaoxiang Zhu, and Haohuan Fu, "Large-scale land cover mapping with fine-grained classes via class-aware semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 16783–16793.