Comparative Analysis of Semantic Segmentation Methods for Satellite Images Segmentation

Qing Bu
CETC Les Information System Co., Ltd
39020765@qq.com

Wei Wan
CETC Les Information System Co., Ltd
1271130252@qq.com

Elizaveta Savitskaya
Belarusian State University
Faculty of Applied Mathematics and Computer Science
Minsk, Belarus
veta.s.working@gmail.com

***Abstract* — This paper proposes a comparative analysis of different automatic semantic segmentation methods for satellite images segmentation on the Semantic Drone Dataset with 23 classes (paved-area, dirt, grass, gravel, water, rocks, pool, vegetation, roof, wall, window, door, fence, fence-pole, person, dog, car, bicycle, tree, bald-tree, ar-marker, obstacle, conflicting). We compare such models as U-net, U-net++, FPN, PAN, DeepLabV3, DeepLabV3+ and Transformer architecture model - SegFormer.**

Keywords* — *semantic segmentation, image segmentation, urban scenes, deep neural network, U-Net, CNN-based semantic segmentation, Transformers.

I. INTRODUCTION

In recent years, satellite image segmentation has garnered increasing attention due to its pivotal role in diverse applications such as agriculture, urban planning, disaster management, and environmental monitoring. Accurate and efficient satellite image segmentation is crucial for extracting meaningful information from high-resolution remote sensing data. To address this challenge, numerous deep learning architectures have been developed, each offering unique advantages and capabilities. As the demand for accurate and efficient satellite image segmentation continues to rise, selecting the most suitable architecture becomes an increasingly important decision.

In this comprehensive comparative analysis, we delve into the world of satellite image segmentation by examining seven cutting-edge architecture models: U-Net, U-Net++, Feature Pyramid Network (FPN), Path Aggregation Network (PAN), DeepLabV3, DeepLabV3+ and SegFormer. These architectures stand at the forefront of image segmentation research and have demonstrated remarkable performance in various computer vision tasks. Our aim is to provide an authoritative assessment of their strengths, weaknesses, and suitability for satellite image segmentation on the Semantic Drone Dataset with 23 different classes of objects.

Through an in-depth exploration of these architectures, we will evaluate their ability to handle the intricate details and complex features present in satellite imagery. We will consider several critical factors, including segmentation accuracy, computational efficiency, speed of learning.

II. DATA

For training and testing ware used The Semantic Drone Dataset provided by Institute of Computer Graphics and Vision, that focuses on semantic understanding of urban scenes and present large number of classes - 23 classes that are: paved-area, dirt, grass, gravel, water, rocks, pool, vegetation, roof, wall, window, door, fence, fence-pole, person, dog, car, bicycle, tree, bald-tree, ar-marker, obstacle and conflicting for area that is not specified as any of previous classes. The imagery depicts more than 20 houses from a nadir (bird's eye) view acquired at an altitude of 5 to 30 metres above ground. A high resolution camera was used to acquire images at a size of 6000x4000px (24Mpx). The training set contains 400 publicly available images and the test set is made up of 200 private images. http://dronedataset.icg.tugraz.at/ Data was divided into training validation and test sample in the following ratio: train : 288, validation : 32, test : 80. Examplea of original images and segmentation are presented on Fig. 1 and Fig. 2.

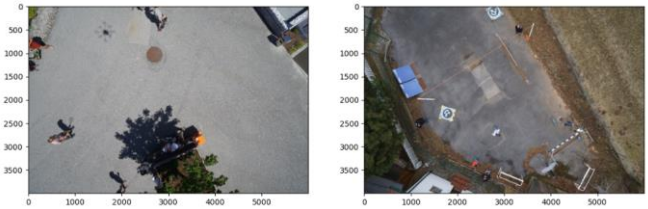

Fig. 1. Examples of original images from Semantic Drone Dataset

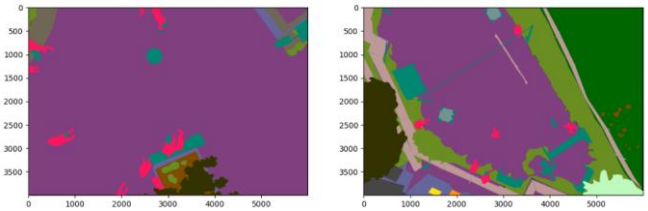

Fig. 2. Examples of original images from Semantic Drone Dataset

III. MODELS DESCRIPTIONS

Selecting the most appropriate model for satellite image segmentation is a critical decision in the field of remote sensing and geospatial analysis. The complexity and diversity of satellite imagery pose unique challenges that require tailored solutions. In this comparative study, the aim to compare performance of some of the most popular models used in segmentation tasks such as U-Net, U-Net++, Feature Pyramid Network (FPN), Path Aggregation Network (PAN), DeepLabV3, DeepLabV3+, and the SegFormer for satellite image segmentation.

A. U-Net and U-Net++

U-Net is a widely recognized architecture for semantic segmentation tasks [1]. Its distinctive U-shaped structure consists of a contracting path (encoder) and an expanding path (decoder). This design enables U-Net to capture both low-level and high-level features, making it well-suited for fine-grained satellite image segmentation. The skip connections between the encoder and decoder facilitate the recovery of spatial details, which are crucial for precise segmentation. However, U-Net may suffer from vanishing gradient issues in

very deep networks and may not fully leverage contextual information.

U-Net++ is an extension of U-Net that addresses some of its limitations [2]. It incorporates skip pathways and nested skip pathways to enhance feature representation in both the encoder and decoder. U-Net++ improves upon U-Net's performance by capturing more contextual information, which is essential for accurate satellite image segmentation. It often achieves higher IoU scores and smoother segmentations. Like U-Net, U-Net++ may still struggle with very large-scale satellite images.

B. FPN

FPN is a feature pyramid architecture designed to handle multi-scale object detection and segmentation tasks [3]. It connects the feature maps from different stages of a backbone network to create a pyramid of features. FPN effectively captures multi-scale features, making it suitable for satellite image segmentation where objects can vary in size. It's especially useful for detecting small objects within large scenes. FPN might be computationally intensive and may require substantial computational resources for training.

C. PAN

PAN builds upon FPN by introducing a mechanism called the spatial attention module. This module helps the network focus on relevant spatial regions, improving segmentation accuracy. PAN enhances the discriminative power of the FPN by incorporating attention mechanisms [4]. This is particularly useful when dealing with complex satellite imagery with intricate structures. PAN's increased complexity may require longer training times and more computational resources.

D. DeepLabV3 and DeepLabV3+

DeepLabV3 and its successor, DeepLabV3+, are renowned for their effectiveness in capturing fine details and semantic context in images [5-6]. These models employ atrous spatial pyramid pooling (ASPP) and dilated convolutions to capture multiscale information. DeepLabV3+ further enhances performance with a feature pyramid network (FPN) backbone. These architectures excel in preserving spatial information, making them well-suited for high-resolution satellite image segmentation tasks, especially when fine details are crucial.

E. SegFormer

SegFormer represents a departure from traditional convolution-based architectures [7]. It introduces the concept of Transformers, originally developed for natural language processing, into the realm of computer vision. SegFormer leverages self-attention mechanisms to capture long-range dependencies and context in satellite images. This architecture offers the potential to learn global features effectively, making it suitable for tasks that require understanding complex spatial relationships in satellite data.

IV. METRICS

To evaluate the performance of semantic segmentation models, various metrics are used to assess their accuracy, robustness, and generalisation capabilities. For this task following metrics were chosen:

1) Pixel accuracy is the simplest metric, and it measures the percentage of correctly classified pixels in the entire image (1).

$$Pixel\ accuracy = \frac{Number\ of\ correctly\ classified\ pixels}{Total\ number\ of\ pixels} \quad (1)$$

2) Intersection over Union measures the overlap between the predicted and ground truth masks for each class. It's calculated as the intersection area divided by the union area. mIoU is the average IoU across all classes and provides a better measure of segmentation quality than pixel accuracy (2).

$$IoU = \frac{TP}{(TP+FP+FN)}, \quad (2)$$

where TP – True Positive: the area of intersection between Ground Truth and segmentation mask, FP – False Positive: The predicted area outside the Ground Truth. This is the logical OR of GT and segmentation minus GT, TP – False Negative: Number of pixels in the Ground Truth area that the model failed to predict.

3) Cross entropy loss, that measures the dissimilarity between predicted pixel-wise class probabilities and ground truth labels.

TABLE I. EXPERIMENT RESULTS FOR ALL MODELS

Criterion	Models						
	U-Net	***U-Net ++***	***FPN***	***PAN***	***DeepLabV3***	***DeepLabV3+***	***SegFormer***
Train Loss	0.782	0.736	0.493	0.673	0.542	0.54	0.271
Val Loss	0.615	0.582	0.389	0.52	0.44	0.405	0.299
Train IoU	0.252	0.246	0.448	0.356	0.413	0.417	0.582
Val IoU	0.267	0.282	0.457	0.374	0.406	0.428	0.546
Train Acc	0.777	0.782	0.847	0.79	0.834	0.834	0.917
Val Acc	0.821	0.825	0.877	0.84	0.863	0.875	0.904
Test IoU	0.278	0.301	0.421	0.348	0.39	0.348	0.499
Test Acc	0.808	0.818	0.866	0.835	0.864	0.867	0.924

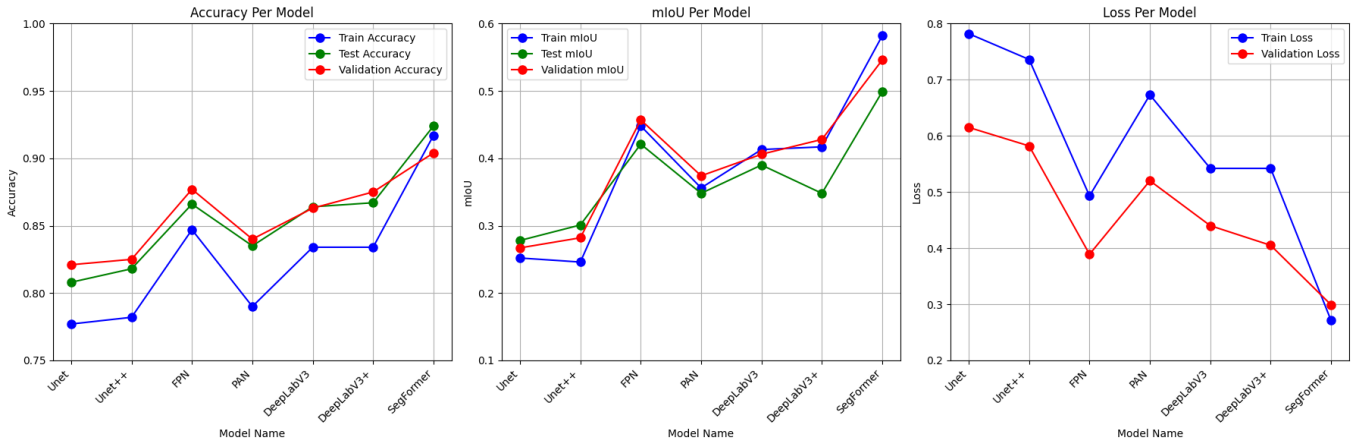

Fig. 3. Model mean pixel accuracy, IoU and loss comparison

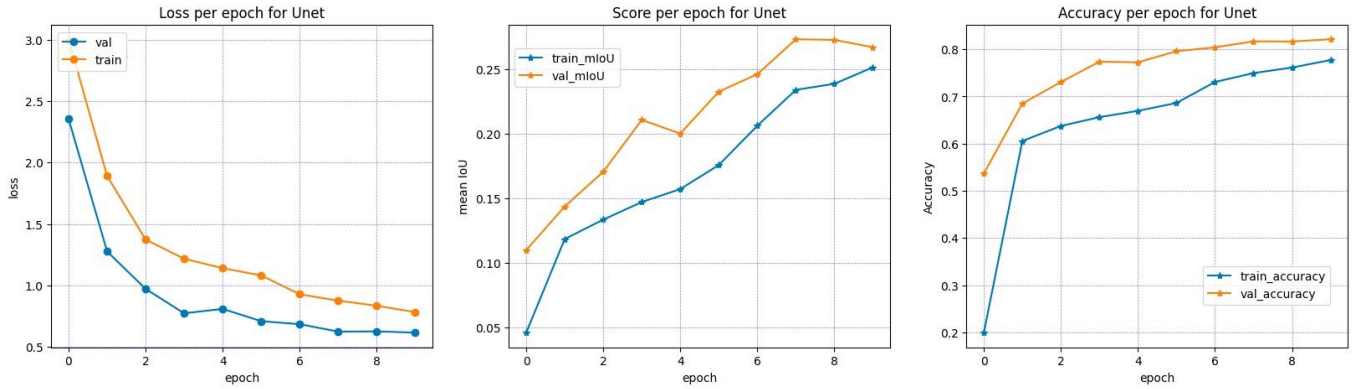

Fig. 4. Loss, IoU and mean pixel accuracy per epoch for Unet

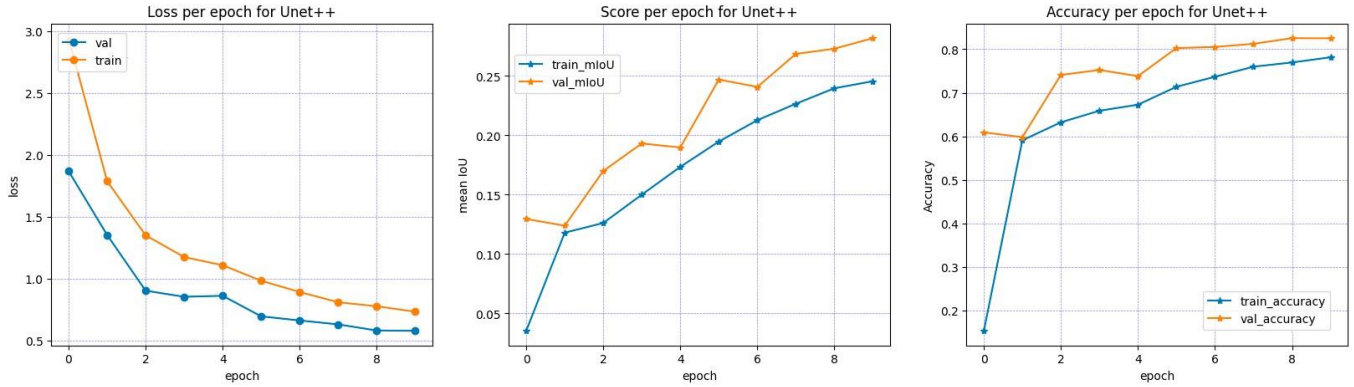

Fig. 5. Loss, IoU and mean pixel accuracy per epoch for Unet++

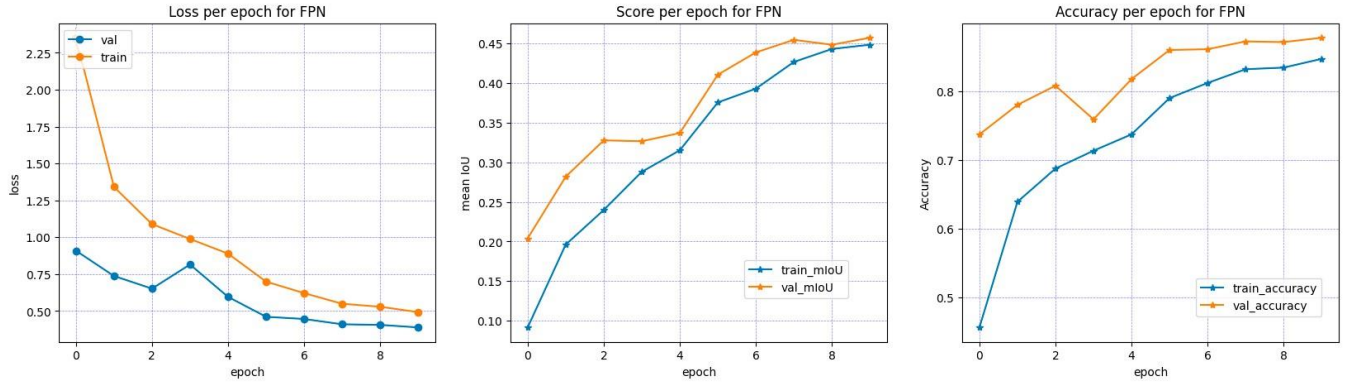

Fig. 6. Loss, IoU and mean pixel accuracy per epoch for FPN

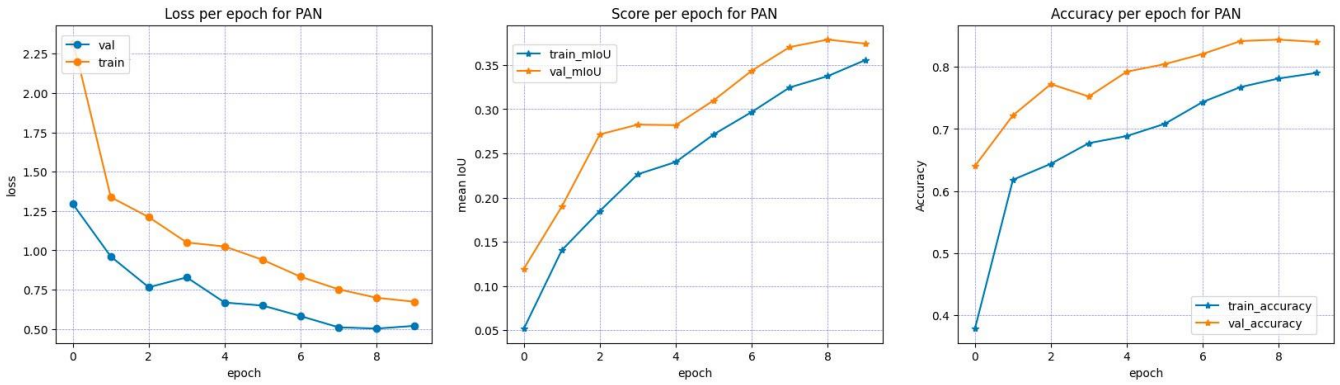

Fig. 7. Loss, IoU and mean pixel accuracy per epoch for PAN

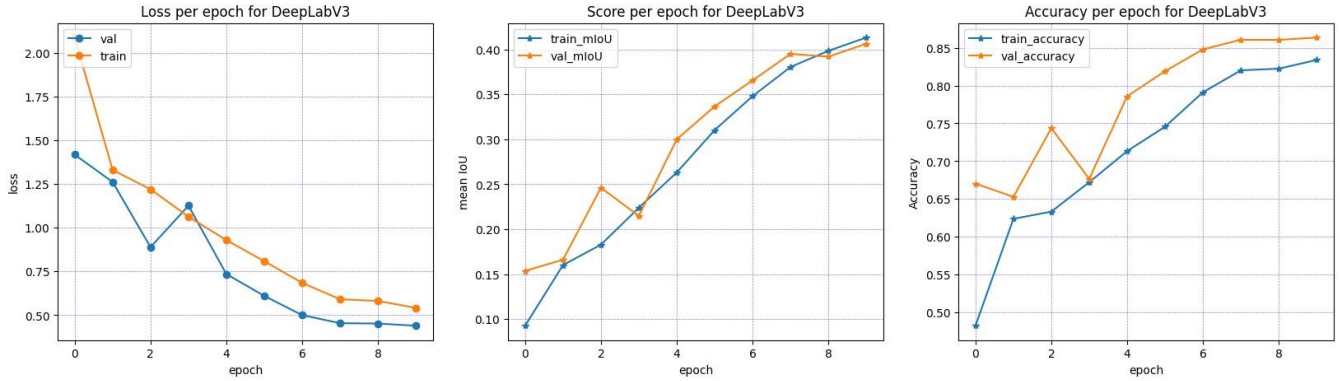

Fig. 8. Loss, IoU and mean pixel accuracy per epoch for DeepLabV3

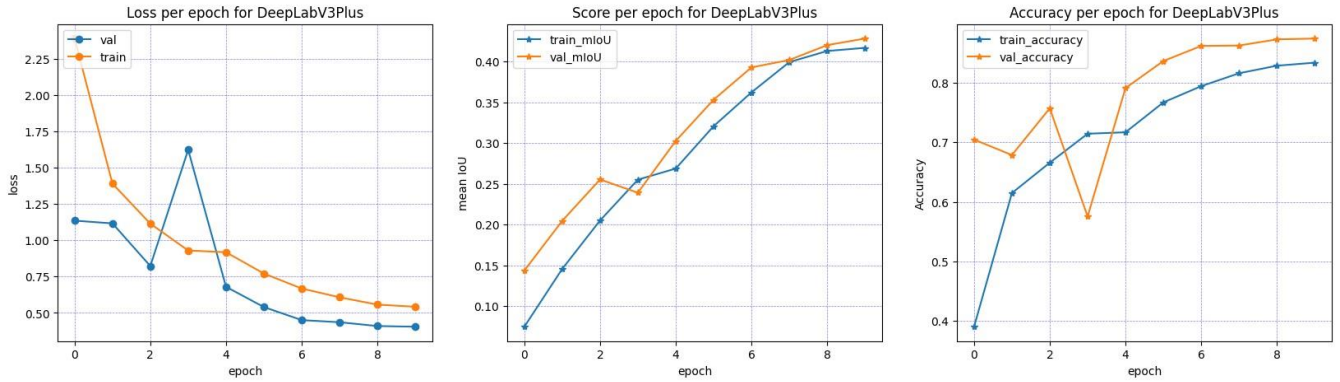

Fig. 9. Loss, IoU and mean pixel accuracy per epoch for DeepLabV3Plus

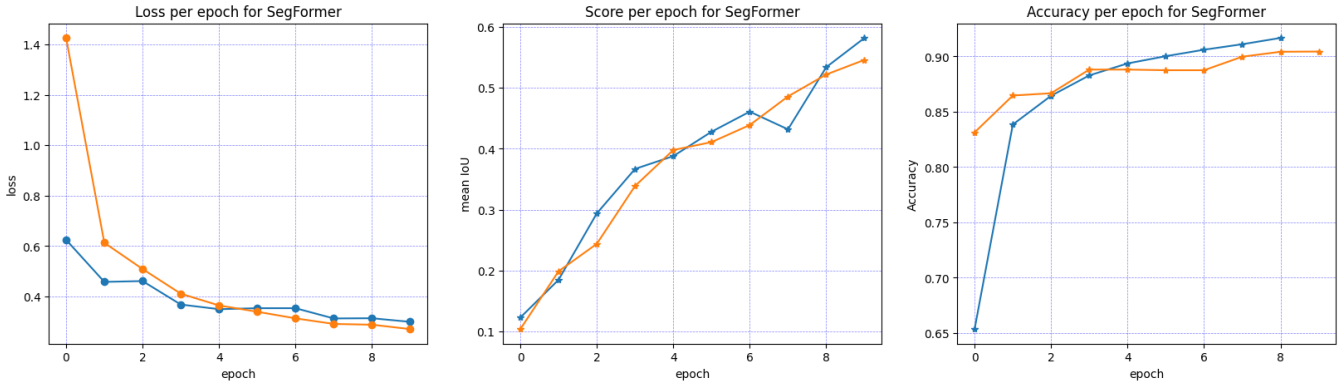

Fig. 10. Loss, IoU and mean pixel accuracy per epoch for SegFormer

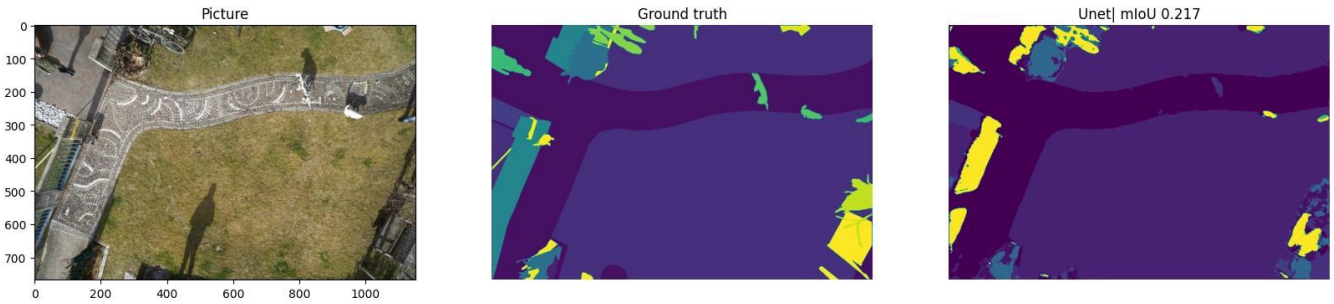

Fig. 11. Original image, ground truth mask and mask predicted by U-net

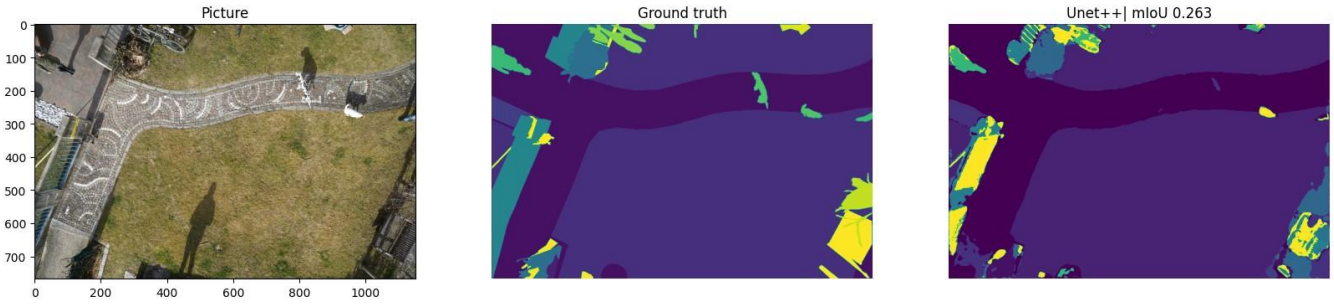

Fig. 12. Original image, ground truth mask and mask predicted by U-net++

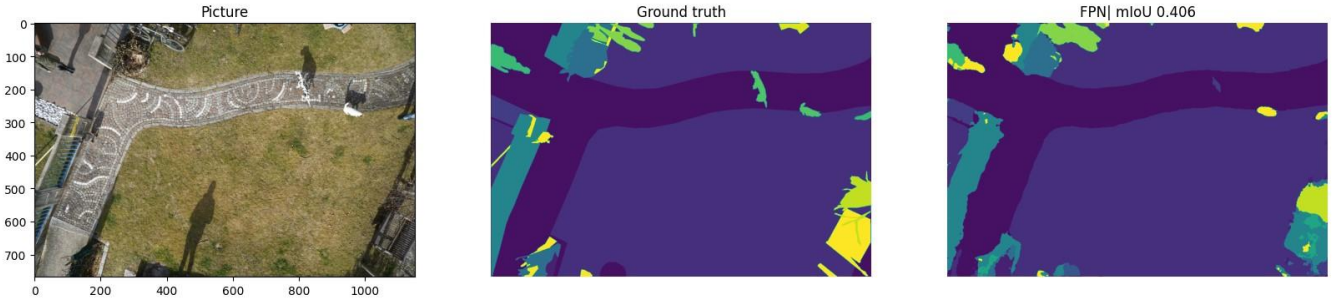

Fig. 13. Original image, ground truth mask and mask predicted by FPN

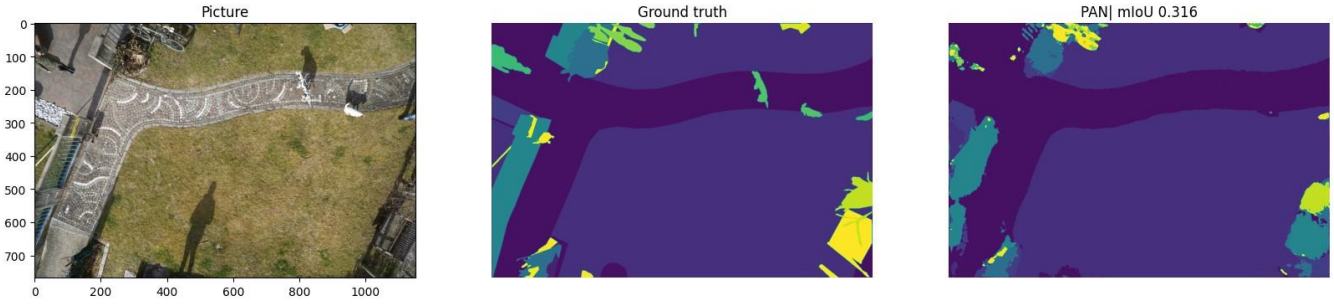

Fig. 14. Original image, ground truth mask and mask predicted by PAN

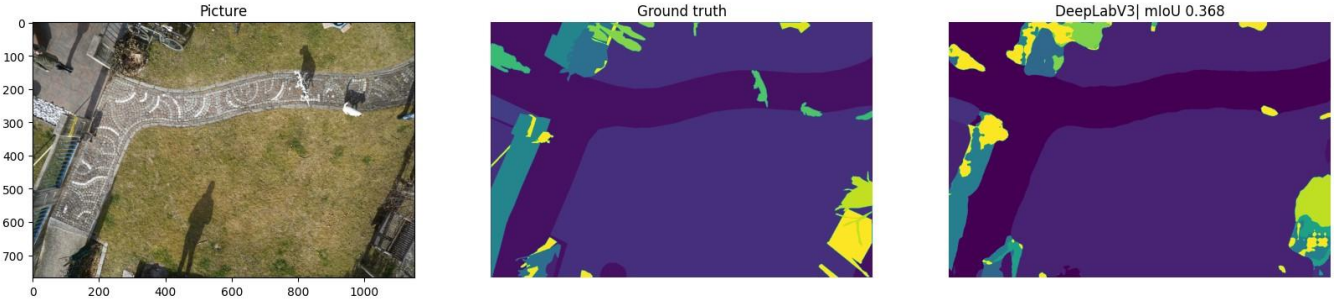

Fig. 15. Original image, ground truth mask and mask predicted by DeepLabV3

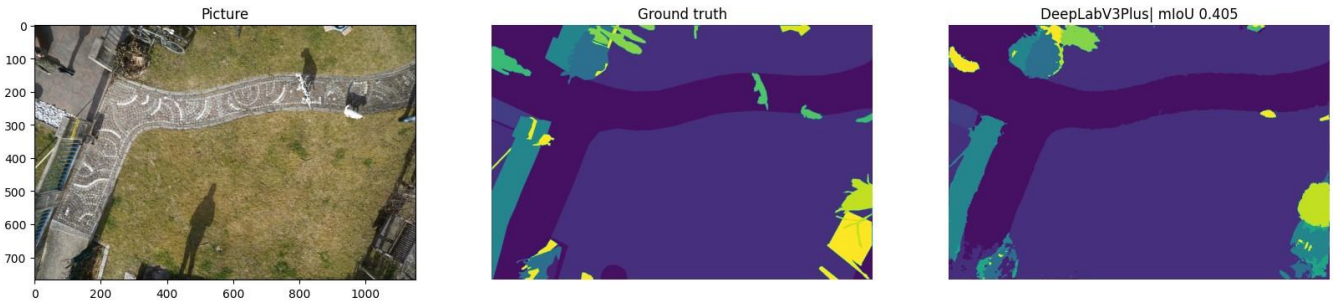

Fig. 16. Original image, ground truth mask and mask predicted by DeepLabV3Plus

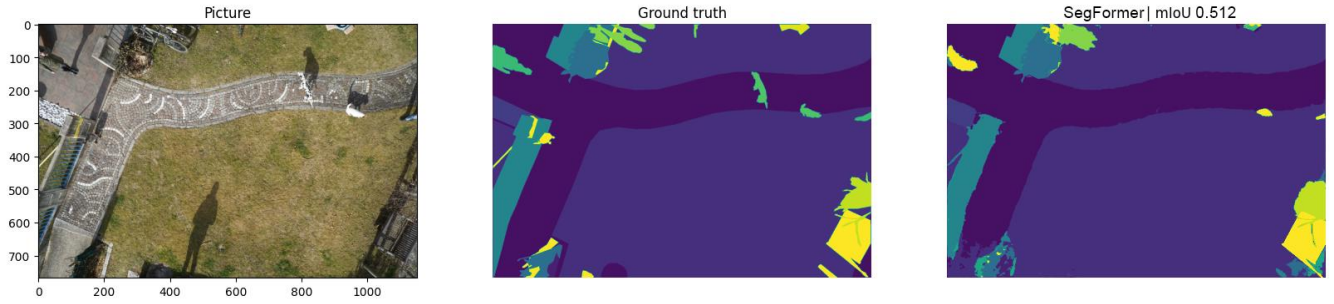

Fig. 17. Original image, ground truth mask and mask predicted by SegFormer

V. Results and Conclusion

Pixel accuracy, mean IoU, presented in Table I, Fig. 3 - Fig.10, and visual analysis of the generated segmentation images Fig. 11 - Fig. 17, allows us to make the following conclusions: FPN, DeepLabV3, DeepLabV3+and SegFormer showed good effectiveness on the selected dataset. We have conducted a thorough examination of their capabilities and characteristics. In this comprehensive evaluation, one model, in particular, stands out as a remarkable performer. Notably, it exhibits exceptional learning speed - more than 5 times faster than other models, enabling efficient adaptation to diverse datasets. Furthermore, SegFormer surpasses its peers by delivering superior accuracy and IoU scores, particularly excelling in delineating intricate object boundaries. In the ever-evolving realm of satellite image segmentation, SegFormer's unparalleled combination of efficiency and precision sets a new standard, promising to elevate the field of remote sensing and Earth observation to greater heights.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, Computer Science Department and BIOSS Centre for Biological Signalling Studies, University of Freiburg, Germany, pp. 4–6, 2015.

[2] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, “UNet++: A Nested U-Net Architecture for Medical Image Segmentation”, 2018, pp. 3–6.

[3] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan and Serge Belongie, “Feature Pyramid Networks for Object Detection”, Facebook AI Research (FAIR), Cornell University and Cornell Tech, 2017, pp. 8-9.

[4] Hanchao Li, Pengfei Xiong, Jie An and Lingxue Wang, “Pyramid Attention Network for Semantic Segmentation”, 2018, 3-6.

[5] Liang-Chieh, Chen George, Papandreou, Florian and Rethinking Atrous, “Convolution for Semantic Image Segmentation”, 2017, pp. 3-8.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”, Google Inc., 2018, pp. 4-8.1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez and Ping Luo, “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers”, The University of Hong Kong Nanjing University, NVIDIA, Caltech, 2021, pp. 3-5.