УДК: 004.8:004.932

Chekhmestruk Roman Y.
Ilchuk Marina R.

IMAGE SEGMENTATION: COMPARISON OF U-NET, MASK R-CNN AND DEEPLAB ALGORITHMS (PIXEL-WISE CLASSIFICATION)

Vinnytsia National Technical University

Анотація

У роботі досліджено сучасні підходи до сегментації зображень на основі глибокого навчання з акцентом на pixel-wise класифікацію. Детально розглянуто механізм формування карт сегментації, принципи навчання нейронних мереж, функції втрат та метрики оцінювання. Проведено порівняльний аналіз архітектур: U-Net, Mask R-CNN та DeepLab. Визначено переваги, недоліки та області застосування кожної моделі, а також окреслено сучасні тенденції розвитку комп'ютерного зору.

Ключові слова: сегментація зображень, pixel-wise класифікація, глибоке навчання, U-Net, Mask R-CNN, DeepLab, комп'ютерний зір.

Abstract

The paper investigates modern deep learning approaches to image segmentation with a focus on pixel-wise classification. The principles of segmentation map generation, model training, loss functions and evaluation metrics are described. A comparative analysis of U-Net, Mask R-CNN and DeepLab architectures is presented. Their advantages, limitations and application areas are identified, along with current trends in computer vision.

Keywords: image segmentation, pixel-wise classification, deep learning, U-Net, Mask R-CNN, DeepLab, computer vision.

Image segmentation is an important task in computer vision that involves dividing an image into semantic regions with high accuracy. Unlike classification or detection, it operates at the pixel level (pixel-wise classification), allowing a detailed representation of a scene [1]. Such systems are widely used in medicine, autonomous driving, satellite analysis, and quality control, where segmentation accuracy directly affects decision-making [2].

The principle of pixel-wise classification

Pixel-wise classification is implemented using convolutional neural networks (CNNs) that transform an image into a class map. The main stages include feature extraction (edges, textures), formation of semantic features, resolution recovery, and classification of each pixel. Training commonly uses cross-entropy, Dice, and IoU loss functions, while model performance is evaluated using Pixel Accuracy, Dice coefficient, and Intersection over Union metrics [3,4].

Context information also plays an important role, since the value of a pixel depends on neighboring regions. Therefore, modern models use dilated convolutions, enlarged receptive fields, and contextual modules to improve segmentation accuracy. The effectiveness of pixel-wise classification thus depends on the network architecture, contextual mechanisms, and training strategy [3].

U-Net Architecture

U-Net is a symmetric encoder–decoder architecture with skip connections that transfer detailed information between layers and help accurately restore object boundaries. The encoder reduces spatial resolution and extracts high-level features, while the decoder reconstructs the spatial structure and produces a segmentation map [4]. Skip connections combine global context and local details, enabling precise segmentation of small structures and complex contours.

The architecture can be extended with variants such as Attention U-Net, U-Net++, and Residual U-Net, which improve segmentation accuracy, especially in medical imaging and small datasets.

Advantages: high segmentation accuracy, fast training, good performance with limited data, accurate boundary reconstruction, flexible architecture, and relatively low computational requirements.

Disadvantages: limited instance-segmentation capability, difficulties with many objects, reduced accuracy on very large and heterogeneous datasets, and sensitivity to class imbalance [3,4].

Overall, U-Net is effective for tasks requiring precise boundaries and limited training data, though complex scenes may require architectural modifications.

Mask R-CNN Architecture

Mask R-CNN performs instance segmentation by combining object detection with pixel-level segmentation. It extends Faster R-CNN and includes three main components: a feature extraction backbone (e.g., ResNet), a Region Proposal Network (RPN) to detect object regions, and a mask head that generates a segmentation mask for each object. A key feature is ROIAlign, which improves region alignment and produces more accurate object boundaries. After region proposals, the model simultaneously performs classification, bounding box refinement, and mask prediction [5].

Advantages: precise object localization, ability to separate overlapping objects, effective performance in complex scenes, and flexible backbone selection.

Disadvantages: high computational cost, long training time, significant hardware requirements, and slower inference compared to semantic segmentation models.

Overall, Mask R-CNN is suitable for tasks that require accurate separation of individual objects in complex scenes, though it requires considerable computational resources and well-annotated data.

DeepLab Architecture

DeepLab is a modern semantic segmentation architecture that uses dilated (atrous) convolutions to expand the receptive field without reducing spatial resolution. Unlike traditional CNNs with pooling, it preserves spatial information, which helps detect object boundaries more accurately. Recent versions (DeepLabv3, DeepLabv3+) employ deep backbone networks such as ResNet or Xception along with decoder modules to improve segmentation accuracy.

The main components include Atrous Convolution for preserving spatial details, ASPP (Atrous Spatial Pyramid Pooling) for multi-scale context analysis, and a decoder that refines object boundaries by combining high- and low-level features [6].

Advantages: high segmentation accuracy, effective handling of objects at different scales, strong use of global context, and good boundary reconstruction.

Disadvantages: complex architecture, high computational and memory requirements, slower training, and the need for large training datasets.

Overall, DeepLab is effective for complex segmentation tasks, though simpler or hybrid models may be preferable when computational resources are limited.

Comparative analysis of algorithms

U-Net and DeepLab are semantic segmentation methods that perform pixel-wise classification, assigning a class to each pixel without separating individual objects [2]. In contrast, Mask R-CNN performs instance segmentation, identifying and separating each object as an individual instance, which is useful for tasks such as object counting and handling overlaps [5].

The architectures differ in how they use contextual information: U-Net combines global and local features through skip connections, DeepLab uses dilated convolutions and the ASPP module for multi-scale context analysis, and Mask R-CNN first detects regions of interest and then segments objects within them [4,6].

They also differ in data and computational requirements. U-Net and DeepLab require pixel-level masks for the whole image, while Mask R-CNN needs separate masks for each object. In practice, U-Net is the fastest and suitable for small datasets, DeepLab performs well in complex multi-class scenes, and Mask R-CNN is most effective for separating individual objects.

The main differences between the considered architectures are summarized in Table 1.

Table 1 – Comparison of U-Net, Mask R-CNN and DeepLab architectures

Architecture	Segmentation Type	Parameters	Pixel Accuracy	Training Speed	GPU Memory Usage	Typical Application
U-Net	Semantic segmentation	~7–30 M	85–92%	Fast	Low–Medium	Medical image segmentation, small datasets
Mask R-CNN	Instance segmentation	~44–63 M	75–85% (mAP)	Slow	High	Object detection and instance segmentation
DeepLab	Semantic segmentation	~35–45 M	86–91% (mIoU)	Medium	Medium–High	Complex scenes and multi-class segmentation

As shown in Table 1, each architecture is suitable for different segmentation tasks.

Current trends

In 2023–2026, image segmentation research has focused on developing architectures that combine high accuracy with efficient processing. Key directions include transformer-based models, hybrid CNN–Transformer architectures, and real-time segmentation methods. Models such as SegFormer and Swin Transformer use self-attention, which helps capture global context and improves segmentation in complex scenes.

Hybrid approaches combine CNNs for local feature extraction with transformers for global context modeling, achieving a balance between detail and contextual understanding. At the same time, real-time models like YOLOv8-seg are being developed for fast processing of video and streaming data. Additionally, self-supervised and semi-supervised learning methods are explored to reduce the need for large labeled datasets.

Conclusion

Pixel-wise classification is the foundation of modern image segmentation, as it assigns a class to each pixel and provides a detailed representation of a scene. This work examines three architectures—U-Net, Mask R-CNN, and DeepLab—designed for different segmentation tasks.

The analysis shows that U-Net performs well on small datasets and tasks requiring precise boundaries (e.g., medical imaging), Mask R-CNN is best for instance segmentation where individual objects must be separated, and DeepLab achieves high accuracy in complex multi-class scenes using dilated convolutions and multi-scale context.

Since no model is universal, the choice of architecture depends on the task, data, and computational resources. Current research focuses on hybrid CNN–Transformer models and self-supervised learning, which reduce the need for large labeled datasets and expand practical applications of image segmentation.

LIST OF REFERENCES

1. DOU Community. Image Segmentation and Deep Learning: Discussion and Practical Aspects. 2022. Тема: основні принципи сегментації зображень та використання методів глибокого навчання у комп’ютерному зорі. С. 1–3.
2. Szeliski R. Computer Vision: Algorithms and Applications. 2nd edition. 2022. Тема: методи комп’ютерного зору, алгоритми обробки зображень та підходи до сегментації. С. 415–430.
3. Zhang Y., Chen H., Li X. Medical Image Segmentation: A Comprehensive Review of Deep Learning Approaches. 2025. Тема: огляд сучасних методів глибокого навчання для сегментації медичних зображень. С. 1–18.
4. Kumar A., Singh P., Verma R. Comparative Study of Deep Transfer Learning Models for Image Segmentation. 2025. Тема: порівняльний аналіз моделей глибокого навчання з використанням перенесення навчання для сегментації зображень. С. 1–12.
5. Li J., Wang H., Zhao Y. Application Mask R-CNN for Object Detection and Segmentation. 2025. Тема: використання нейронної мережі для виявлення та сегментації об’єктів на зображеннях. С. 1–10.
6. Chen L., Zhou T., Zhang Y. Improved DeepLabv3+ Model for Image Segmentation Tasks in Complex Scenes. 2024. Тема: удосконалена модель нейронної мережі для сегментації зображень у складних сценах. С. 1–12.

Чехместрук Роман Юрійович – доцент кафедри програмного забезпечення, Вінницький національний технічний університет, м. Вінниця, e-mail chekhroma@gmail.com

Ільчук Маріна Русланівна – студент факультету ІТКІ, групи 2ПІ-24б, Вінницький національний технічний університет, м. Вінниця, e-mail: marinailchuk05@gmail.com

Chekhmestruk Roman Y. – Associate Professor of the Department of Software Engineering, Vinnytsia National Technical University, Vinnytsia, Ukraine, e-mail: chekhroma@gmail.com

Ilchuk Marina R. – student of the Faculty of ITKI, Vinnytsia National Technical University, Vinnytsia, e-mail: marinailchuk05@gmail.com