UDC 004.83 DOI: 10.30837/0135-1710.2025.185.043

A.R. KOVTUNENKO, S.V. MASHTALIR

A REVIEW OF MODERN NEURAL NETWORK ARCHITECTURES FOR IMAGE SEGMENTATION

The paper presents a chronological review and analysis of the evolution of image segmentation methods. It examines the development of architectures from early fully convolutional networks (FCN) to modern transformer zero-shot segmentation models such as Segment Anything Model (SAM). The key architectures, their innovations, technical features, advantages, and limitations were discussed in detail. The review allows understanding the main trends in the development of segmentation methods, evaluating the effectiveness of different approaches, and making an informed choice of architectures for solving practical problems depending on the requirements for accuracy, computational resources, and application area.

1. Introduction

Computer vision has been developing rapidly and continues to do so. Every year, existing algorithms are improved, radically new methods are created, and hybrid approaches are introduced that combine the strengths of different methods to solve previous or still unsolved problems. For example, segmentation has seen a significant transformation of paradigms and approaches in the last decade, from traditional heuristic methods to fully-connected convolutional methods and deeper transformer-based models that can handle multimodal data. Transformers were originally developed for natural language processing. They were later adapted for computer vision tasks due to a key advantage: they can identify correlations between disparate regions in images regardless of how far apart these regions are. This capability was difficult to achieve with convolutional networks, as their receptive field is limited to local neighborhoods that only gradually expand with increased network depth. Notably, this development has not been linear, and early methods have not been superseded; rather, many current approaches successfully combine classical methods with neural network methods, creating hybrid solutions that outperform their predecessors. However, despite significant progress, aspects such as finding object boundaries in complex scenes, finding objects of small size, domain adaptation, utilizing prior knowledge, and computational limitations still require further research and improvement.

2. Relevance and purpose of the article

The paper's relevance is due to the rapid increase of methods and approaches for different segmentation tasks and conditions, as well as the need to understand the strengths and weaknesses of existing methods. The systematization of existing approaches in chronological order will allow not only tracing the evolution of architectural solutions, but also understanding the main ideas and research directions, as well as limitations and unsolved problems that remain relevant in the present time. This will avoid the repetition of already known limitations and mistakes from past approaches, allowing for a more informed development of new models.

The purpose of this paper is to chronologically review and analyze segmentation models to identify the main innovations, limitations, and advantages, thus allowing an informed choice of methods and approaches for solving applied problems or designing new, more effective methods.

3. Review of the models

The evolution of neural network architectures for segmentation tasks is reasonable to start with the consideration of a full convolutional network (FCN) (Fig. 1) [1]. This architecture enables the processing of images at arbitrary resolution and obtaining a segmentation result that does not require additional post-processing. This method is an adaptation of the existing SOTA (State-of-the-Art) models (VGG, AlexNet, etc.) by replacing the fully connected layers with convolutional ones and adding layers for dimensionality recovery. However, despite these advantages, this architecture showed unsatisfactory results for the segmentation of complex scenes, small objects, and boundary detection.

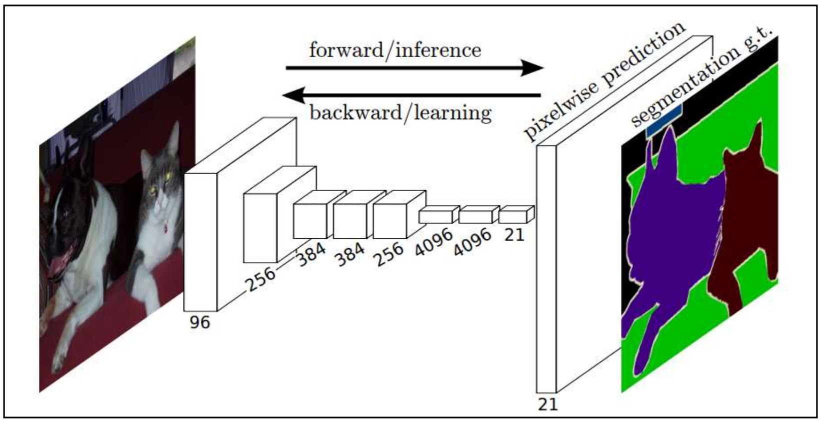

Fig. 1. FCN architecture [1]

The advent of the U-Net (Fig. 2) [2] architecture has solved some of the problems of FCN, especially in the context of handling fine details in images and boundary extraction. U-Net is also a fully convolutional neural network, which already has skip connections, allowing the model to retrieve information about objects from different levels of representations and recover their spatial information. The introduction of a weighted loss function, with increased coefficients for boundary pixels, allowed the accurate extraction of object boundaries. The architecture was originally developed for medical images and small amounts of raw data. Applying augmentation to the data allowed to achieve high recognition accuracy, but it should be noted that with insufficient variability of the data set, there is a risk of overfitting the model. The disadvantages of U-Net include computational cost, especially as the input image size increases, and the limited receptive field, which can make it difficult to analyze structures that require a larger context. To solve the problem of high-resolution image processing, the authors proposed an overlap-tile strategy.

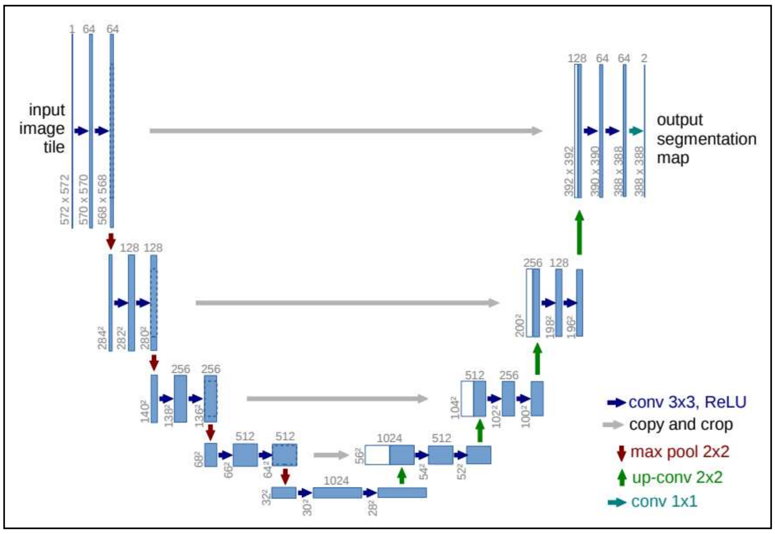

Fig. 2. U-Net architecture [2]

The SegNet (Fig. 3) [3] architecture, compared to U-Net, allows processing high-resolution images with lower resource costs, while still finding object boundaries accurately enough. It has been specifically designed for systems requiring high-speed data processing, even in real time. SegNet is also a two-part encoder-decoder, but uses modified upsampling in the decoder module. The spatial resolution is restored using max pooling indices stored at the corresponding encoder levels during the max pooling operation. The feature maps themselves are not transferred, which significantly reduces the number of model parameters. Since skip connections are not applied, this can lead to partial loss of information about the global context of the image, which potentially reduces the effectiveness of segmentation in scenarios with complex spatial structures.

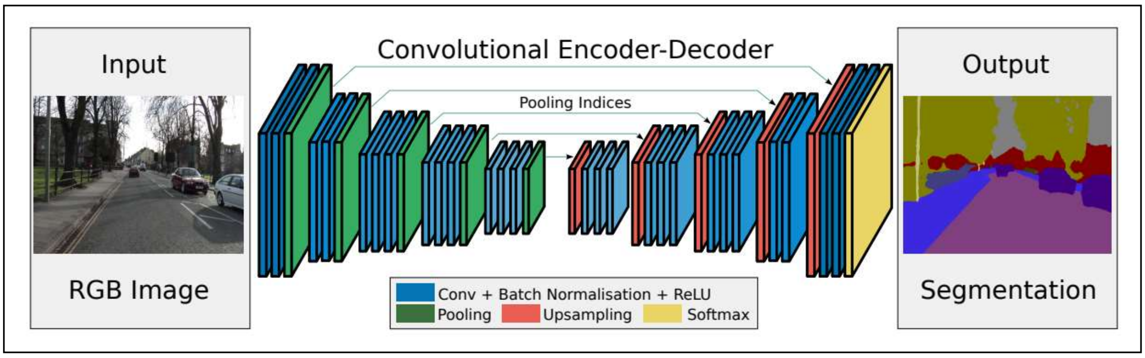

Fig. 3. SegNet architecture [3]

The DeepLab family of models (DeepLabV1 [4], DeepLabV2 [5], DeepLabV3 [6], DeepLabV3+ (Fig. 4) [7]) represents a sequential evolution of architectures aimed at improving the accuracy of semantic segmentation. Each iteration of the model has brought its additions and improvements. DeepLabV1 applied Atrous Convolutions, characterized by the introduction of specific gaps between convolutional kernel elements. This approach allows for increasing the receptive field without changing the feature map resolution. DeepLabV2 develops the concept of dilated convolutions by introducing Atrous Spatial Pyramid Pooling (ASPP) – a set of convolutions with different values of dilation parameters. This solution has significantly improved the quality of segmentation of objects of different scales in images. In DeepLabV3, the ASPP architecture was supplemented with a global pooling component. Its advantages are as follows: it takes the entire feature map and reduces it to a single vector, it captures image-level context that helps understand the overall scene, and it complements the local features captured by convolutional layers. Also, dilated convolutions were added to all parts of the model. DeepLabV3 introduces a «multi-grid» approach where different atrous rates are applied within consecutive blocks. This provides a hierarchical structure of receptive fields and shows significant improvement over using uniform rates. Despite the evolution of the DeepLab family, finding small or heavily occluded objects, or objects with rare views, is still a challenge for this architecture. In DeepLabV3+, the architecture has been changed to encoder-decoder, similar to U-Net. ResNet networks used in previous versions or a modified Xception architecture are used as an encoder-decoder in this architecture. Upsampling was made in two stages to improve granularity. At the first stage, decoder gets low-level features from the encoder and combines them with results from ASPP, which are preliminarily upsampled by a factor of four. The second stage is another 4X increase to get the final result. The authors experimented with a more complex decoder structure, including two levels of skip connections, but found no significant performance improvement over their simpler version. The complexity of the architecture improved the quality of segmentation, especially of object boundaries, but naturally led to an increase in computational cost.

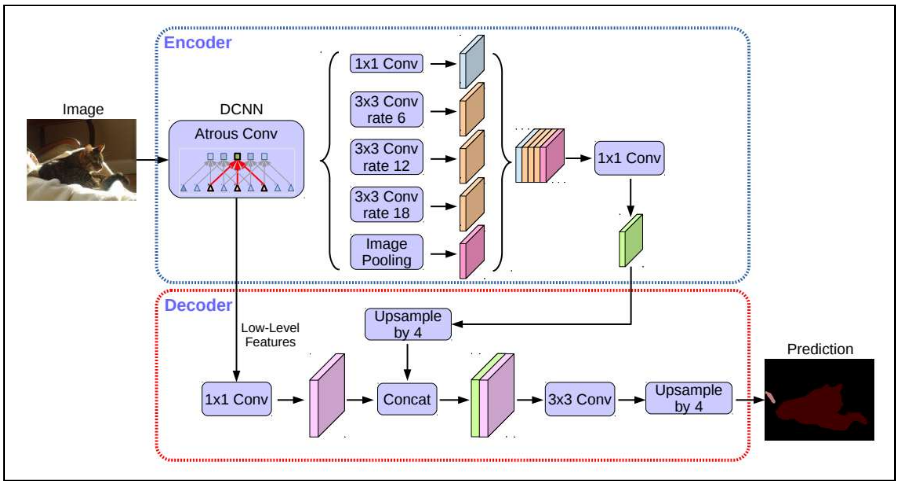

Fig. 4. DeepLab V3+ architecture [7]

Mask R-CNN (Fig. 5) [8] is an architectural evolution of the approaches laid down in Faster R-CNN, but with extended functionality to solve the instance segmentation, where it is required to separate each object within the same class. This architecture implements a two-stage approach: the first stage involves predicting regions of interest (RoI) in the input image, after which RoIAlign accurately extracts the corresponding regions from the feature maps generated by the feature extraction network (backbone). In the second stage, the extracted regions are classified, and in parallel, a separate branch of the model generates segmentation masks for each identified object. This approach became the standard for a long period and was able to predict accurate masks, even for objects with significant overlap. However, again, this approach is very expensive in computational resources and still shows suboptimal results for small-sized objects. An important advantage of Mask R-CNN is its versatility – it can be applied to classification, detection, and segmentation tasks. The disadvantages include the difficulty of training the two-stage approach on new data due to sensitivity to hyperparameters and dependence of mask prediction on the results obtained from the region proposal network (RPN).

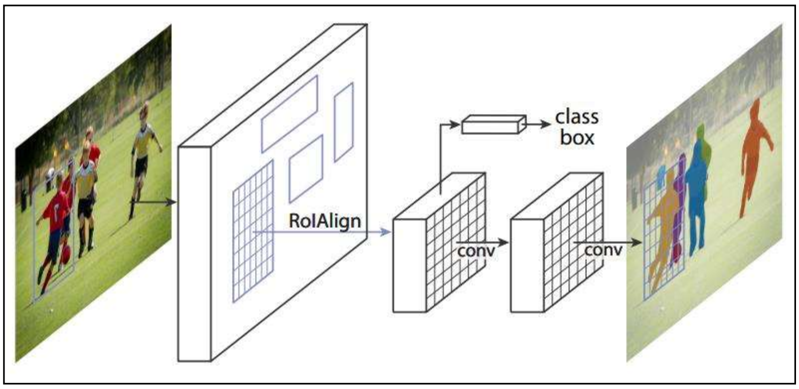

Fig. 5. Mask R-CNN architecture [8]

The problem of incorrect detection of object boundaries, especially in cases of their mutual overlap, remains one of the fundamental limitations of semantic segmentation algorithms. In the Gated-SCNN (Fig. 6) [9] method, the authors aimed to solve this problem. For this purpose, they used two branches in the architecture. The main backbone branch deals with extracting features from the input image, which are then fed via Gated Convolution Later into the second branch, which works with the original image pre-processed with the Canny filter – i.e., with gradients.

The results of these two branches are then combined via ASPP to form the final segmentation result. It should be noted that the Gated Convolution Layer module proposed within this architecture is conceptually equivalent to modern attention mechanisms. However, the main contribution of this paper is a new loss function – dual task loss, which provides the ability to compare the predicted object boundaries with the true ones. The main drawbacks of this model are its pronounced dependence on the quality of the training data annotations, because inaccuracies in the boundary labeling inevitably lead to a decrease in the accuracy of model predictions, and sensitivity to hyperparameter tuning, which is a typical characteristic of composite architectures and loss functions.

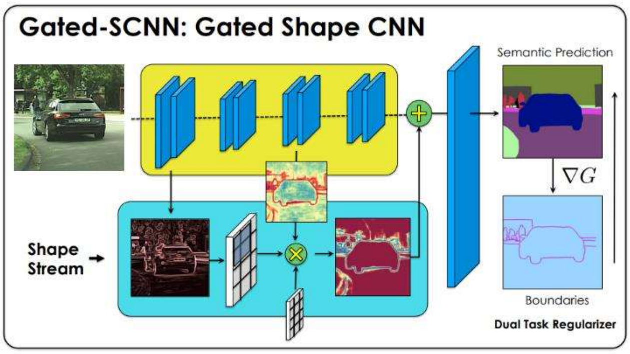

Fig. 6. Gated-SCNN architecture [9]

FastFCN (Fig. 7) [10] is another approach to semantic segmentation in which a new Joint Pyramid Upsampling (JPU) block has been proposed for upsampling. In this block, the input feature maps are first processed with standard convolution, followed by multiple separable convolutions with varying dilation rates. The authors of the research use the last three convolutional layers from the backbone network, although the number of convolutional layers can vary. This approach improves the reconstruction of fine details and significantly reduces computational cost. After JPU, there is a multi-scale/global context module to generate the final segmentation map, such as ASPP. The main advantages of this approach are the increased computational speed and the possibility to apply JPU to other already existing architectures or their modules, e.g., DeepLabV3 or other individual backbones. The main drawback, however, is that during training, the input image size had to be restricted due to memory limitations, and even a resolution difference of 96 pixels affected the performance of the model.

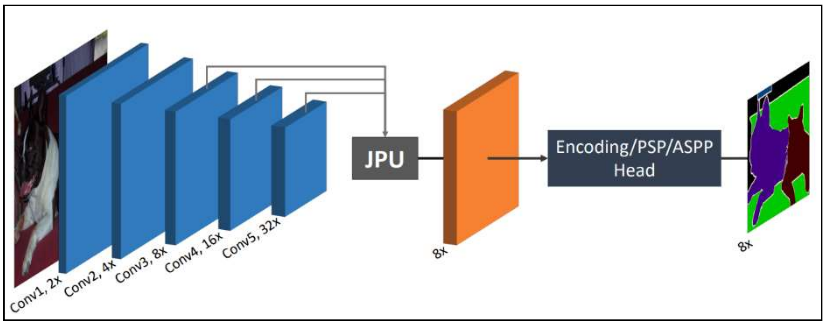

Fig. 7. FastFCN architecture [10]

MaskFormer (Fig. 8) [11] is a segmentation model that incorporates an attention mechanism into its architecture. Its fundamental conceptual difference is not in per-pixel classification or selecting regions of interest with objects, but in first generating a set of object masks, which are then classified. To implement this approach, the backbone extracts embeddings, which are processed by a pixel decoder and a transformer decoder to produce binary object masks. The architecture of the decoders is a box-free DETR [12]. This proposed architecture demonstrates high efficiency and compatibility with various backbones. For example, compared to DeepLabV3+, MaskFormer requires fewer parameters while achieving higher accuracy when using the same backbone. The disadvantages include the need for large training datasets and careful parameter selection, as transformers tend to achieve better results when trained on large amounts of data.

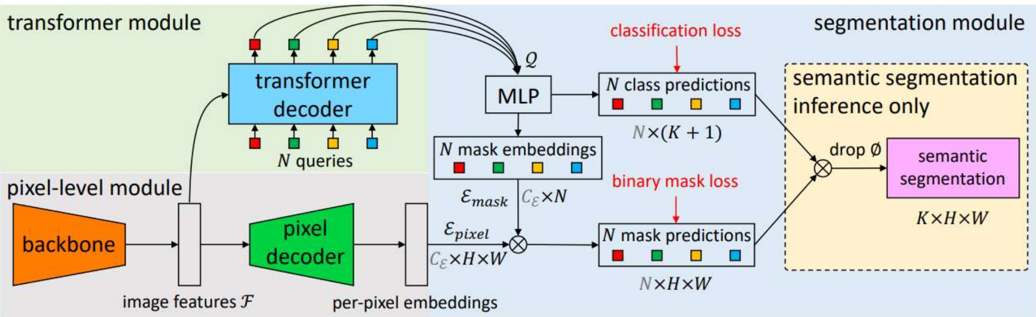

Fig. 8. MaskFormer architecture [11]

The SegFormer (Fig. 9) [13] architecture is an encoder-decoder semantic segmentation model. Its encoder is implemented as a hierarchical transformer that processes images in 4×4 patches, progressively reducing the spatial resolution and generating multi-level features. This design provides the ability to process input images of varying sizes without the need for positional encoding mechanisms. The main focus of SegFormer architecture is its minimalist decoder, which consists solely of a multilayer perceptron (MLP) and upsample layers, significantly reducing computational complexity. The decoder aggregates the feature maps produced by the encoder, aligning them to a unified resolution to produce the final result. The presented solutions and innovations in the SegFormer model provide an optimal balance between computational efficiency and segmentation accuracy, as demonstrated by SOTA results among the models of the corresponding period.

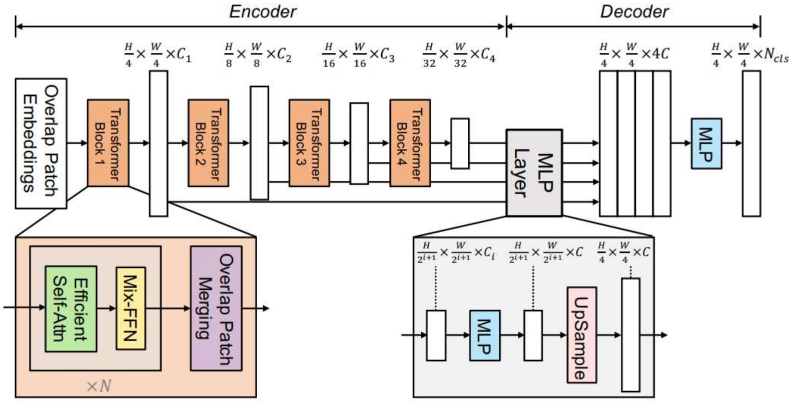
Encoder
Decoder
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{4}\times\frac{W}{4}\times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Efficient Self-Attn
Mix-FFN
Overlap Patch Merging
$\times N$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C$
$\frac{H}{4}\times\frac{W}{4}\times C$
MLP
UpSample

Fig. 9. SegFormer architecture [13]

Mask2Former (Fig. 10) [14] is a development of the concepts introduced in MaskFormer. The main architectural innovation is the replacement of cross-attention with masked-attention and adding multi-resolution feature maps to the pixels decoder, similar to the Feature Pyramid

Network (FPN) design. Masked attention allows the model to focus more effectively on objects or regions of interest, which accelerates the convergence process. The use of multi-scale high-resolution features in the pixel decoder allows the model to segment small objects or regions. To optimize memory usage during training, the authors proposed an approach in which only a randomly selected subset of ground-truth mask points is matched, rather than all pixels. This technique is inspired by the PointRend [15] and the Implicit PointRend [16] techniques and reduces memory consumption by up to three times. The key advantages of the model are its versatility for different types of segmentation tasks and its high accuracy. However, as with many high-performing models, these benefits come at the cost of increased computational demands and architectural complexity.

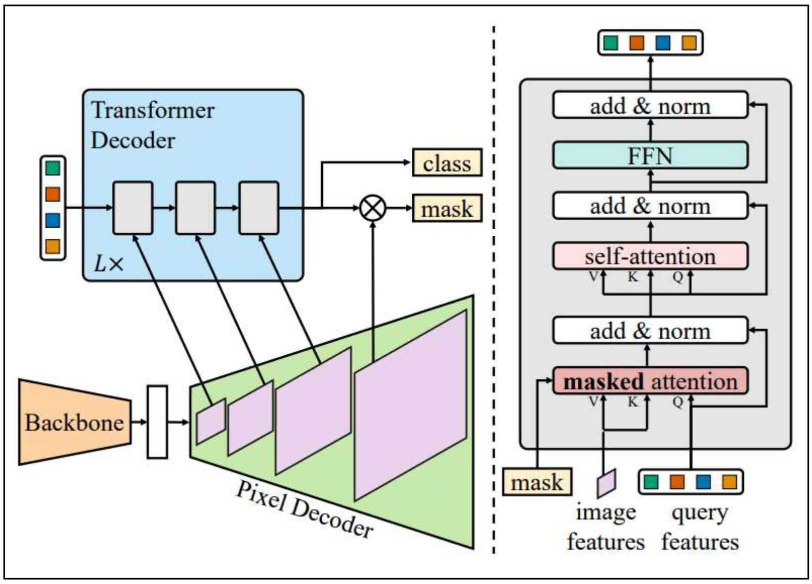

Fig. 10. Mask2Former architecture [14]

OneFormer (Fig. 11) [17] is a logical continuation of the MaskFormer and Mask2Former architectures, offering a unified approach to various segmentation tasks, including semantic, instance, and panoptic segmentation within a single model. The key difference of OneFormer is the use of a task-conditional token that defines the type of task being performed. The model architecture consists of three main components: an image encoder based on the Swin Transformer, a task token, and a decoder with multi-level attention mechanisms. The segmentation process in OneFormer involves extracting visual features from the input image, incorporating a task token, generating object queries in the decoder depending on the type of task, and predicting binary masks along with their corresponding classes. Thanks to its unified architecture and task-conditioned design, OneFormer achieves high performance across all three segmentation tasks without requiring task-specific reconfiguration. However, the model has several limitations, including high training complexity, the need for large amounts of annotated data, and significant computational cost.

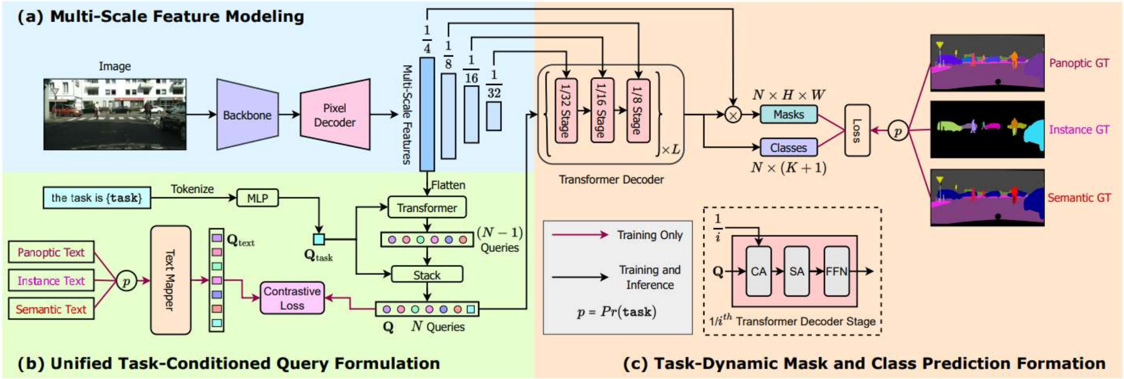

Fig. 11. OneFormer architecture [17]

Segment Anything Model (SAM) (Fig. 12) [18] is a fundamentally new paradigm in segmentation focusing on universality, interactivity, and zero-shot generalization to unseen domains. The architecture consists of three main components: an image encoder based on a modified Vision Transformer (ViT), a prompt encoder capable of handling different types of input data (e.g., text, points, boxes, masks), and a lightweight mask decoder that produces segmentation masks. What sets SAM aside from other models is its training on an unprecedentedly large dataset (more than one billion masks) and its interactive behavior that supports iterative refinement by allowing users to provide new prompts interactively. The model demonstrates strong adaptability to new domains and object types. Nevertheless, SAM has several limitations, including high computational costs, reduced accuracy for segmentation of small objects, complex or fine-grained boundaries, and difficulties with contextual understanding of complex scenes.

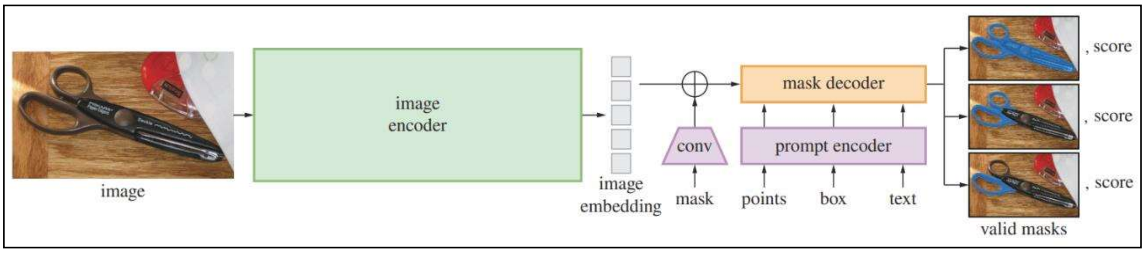

Fig. 12. SAM architecture [18]

HQ-SAM (Fig. 13) [19] is an enhanced version of the original SAM model designed to improve the quality of predicted masks, especially in the context of accurate object boundary detection. Although SAM was trained on 1B masks, its segmentation results may be suboptimal when dealing with objects that have complex structures. The HQ-SAM model addresses two main problems of the original SAM: inaccurate mask boundaries for small or thin objects, and incorrect mask predictions or fragmented predictions in complex scenes. These types of errors significantly reduce the applicability and effectiveness of SAM. To address this, the HQ-SAM architecture introduces two main components while retaining the zero-shot capabilities of the original model. The High-Quality Output Token (HQ-Output Token) is a trainable token added to the SAM mask decoder along with the original prompt and output tokens. The second innovation, Global-local Feature Fusion, combines features from the SAM mask decoder with both early-layer local features and final-layer global features from the ViT encoder. This fusion allows for taking into account boundary information, semantic context, and mask shape representation from both the ViT encoder and mask decoder to improve mask refinement. A TinyViT-based Light HQ-SAM [19] model variant was also introduced to support resource-constrained devices, since the full

HQ-SAM model has even more parameters than the original SAM.

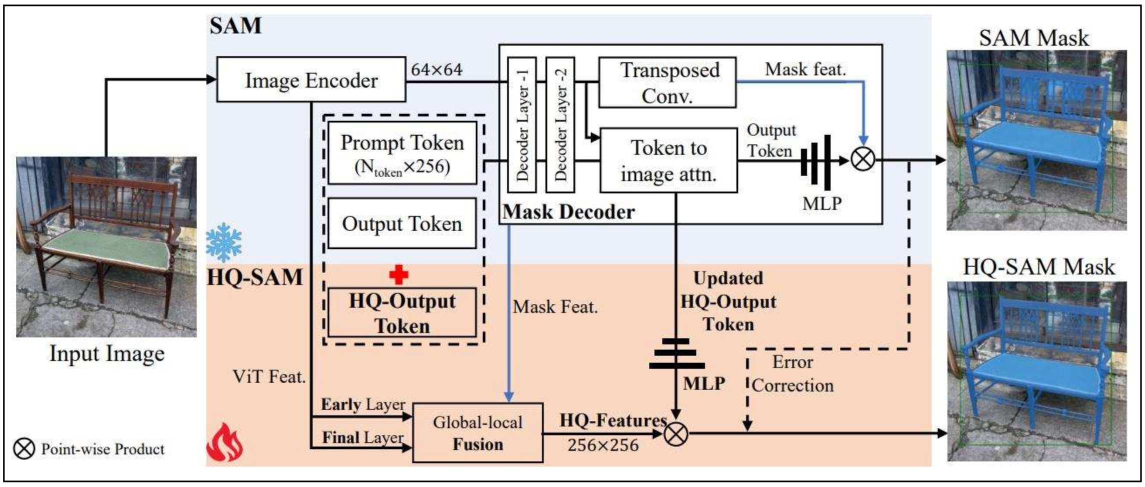

Fig. 13. HQ-SAM architecture [19]

4. Conclusions and prospects of further research

This paper provides a comprehensive review of the evolution of neural network architectures for image segmentation tasks. It examines key milestones in the development of segmentation models, from early fully convolutional networks to modern transformer-based architectures. Technical features, advantages, and limitations of different architectures were analyzed. Reviewed architectures clearly illustrate the progress of models and their approaches to problem solving. Each of the analyzed architectures has its advantages and disadvantages, which determine the areas of their optimal application in different types of segmentation. Modern architectures have successfully overcome many of the limitations of earlier approaches, offering universal solutions for various types of segmentation with an optimal balance between accuracy and efficiency. Nevertheless, several fundamental problems remain unresolved, such as domain adaptation, accurate small objects segmentation, and reducing dependence on large amounts of labeled data. Transformer-based architectures, despite their high accuracy, still require significant computational resources, which limits their application in real-time systems and highlights the need for further research to find the optimal balance between segmentation quality and computational efficiency.

References:

1. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015. doi:10.1109/cvpr.2015.7298965

2. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Lecture Notes in Computer Science, pp. 234–241, 2015. doi:10.1007/978-3-319-24574-4_28

3. V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, Dec. 2017. doi:10.1109/tpami.2016.2644615

4. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs“. arXiv preprint arXiv:1412.7062, Jun. 2016.

5. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834–848, Apr. 2018. doi:10.1109/tpami.2017.2699184

6. L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation,” arXiv preprint arXiv:1706.05587, Jun. 2017.

7. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for Semantic Image segmentation,” Lecture Notes in Computer Science, pp. 833–851, 2018.

doi:10.1007/978-3-030-01234-2_49

8. K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017. doi:10.1109/iccv.2017.322

9. T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape cnns for semantic segmentation," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2019. doi:10.1109/iccv.2019.00533

10. H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation," arXiv preprint arXiv:1903.11816, Mar. 2019.

11. B. Cheng, A. G. Schwing, and A. Kirillov, "Per-Pixel Classification is Not All You Need for Semantic Segmentation," Advances in neural information processing systems, vol. 34, pp. 17864–17875, Jul. 2021.

12. N. Carion et al., "End-to-end object detection with Transformers," Lecture Notes in Computer Science, pp. 213–229, 2020. doi:10.1007/978-3-030-58452-8_13

13. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," Advances in neural information processing systems, vol. 32, pp. 12077–12090, May 2021.

14. B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for Universal Image segmentation," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1280–1289, Jun. 2022. doi:10.1109/cvpr52688.2022.00135

15. A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020. doi:10.1109/cvpr42600.2020.00982

16. B. Cheng, O. Parkhi, and A. Kirillov, "Pointly-supervised instance segmentation," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022. doi:10.1109/cvpr52688.2022.00264

17. J. Jain et al., "OneFormer: One Transformer to rule Universal Image segmentation," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2989–2998, Jun. 2023. doi:10.1109/cvpr52729.2023.00292

18 A. Kirillov et al., "Segment Anything," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026, Apr. 2023.

19. L. Ke et al., "Segment Anything in High Quality," Advances in Neural Information Processing Systems, vol. 36, pp. 29914–29934, Jun. 2023.

Надійшла до редколегії 14.05.2025 р.

Kovtunenko Andrii Romanovych, PhD student, Informatics Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, e-mail: andrii.kovtunenko@nure.ua, ORCID: https://orcid.org/0009-0004-9072-7779

Mashtalir Sergii Volodymyrovych, Doctor of Engineering Science, Professor, Informatics Department. Kharkiv National University of Radio Electronics. Kharkiv, Ukraine, e-mail: sergii.mashtalir@nure.ua, ORCID: https://orcid.org/0000-0002-0917-6622