UDC 004.832
DOI: 10.53297/18293336-2025.1-59

COMPARATIVE ANALYSIS OF CONVOLUTIONAL NEURAL NETWORKS AND VISION TRANSFORMERS FOR SATELLITE IMAGE SEGMENTATION

T.B. Khachatryan

National Polytechnic University of Armenia
"Synopsys Armenia" CJSC

The rapid advancement of deep learning has revolutionized satellite image analysis with both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) showing promising results. This study presents a comprehensive comparative analysis of these architectures for land cover segmentation in satellite imagery, addressing the critical need for understanding performance-efficiency trade-offs in real-world deployment scenarios. Four state-of-the-art models were fine-tuned and evaluated: ResNet50 and EfficientNet-B0 representing CNNs, ViT-B/16 and Swin Transformer representing the transformer family. Using the DeepGlobe Land Cover Classification dataset containing 803 high-resolution satellite images with seven land cover classes, model performance was assessed across multiple metrics including mean Intersection over Union (mIoU), F1-score, and pixel accuracy. Additionally, computational requirements including inference speed, model size, and memory consumption were analyzed on an NVIDIA RTX 4070 GPU to simulate practical deployment constraints. The results demonstrate that while Vision Transformers achieve superior segmentation accuracy with Swin-T reaching 74.2% mIoU compared to 71.8% for EfficientNet-B0, CNNs maintain significant advantages in inference speed and memory efficiency. EfficientNet-B0 processes images 2.3 times faster than ViT-B/16 while using 40% less GPU memory. Class-wise analysis reveals that transformers particularly excel in complex scenarios like urban areas and forest boundaries, while all models achieve over 89% IoU for water body segmentation. These findings provide practical insights for selecting appropriate architecture based on specific deployment constraints, highlighting the trade-offs between accuracy and computational efficiency in satellite image analysis applications for environmental monitoring, urban planning, and disaster response.

Keywords: satellite image segmentation, Vision Transformer, Convolutional Neural Networks, deep learning, land cover classification.

Introduction. Satellite image analysis has become increasingly crucial for various applications, including urban planning, environmental monitoring, agricultural assessment, and disaster response [1]. The ability to automatically segment and classify different land cover types from satellite imagery enables large-scale monitoring and decision-making processes that would be impractical through manual analysis. Recent advances in deep learning have transformed this field, offering unprecedented accuracy in automated land cover classification and segmentation tasks [2]. The emergence of Vision Transformers (ViTs) (Fig. 1)

alongside established Convolutional Neural Networks (CNNs) has created new possibilities and challenges for satellite image analysis. While CNNs have proven effective through their hierarchical feature extraction and spatial inductive biases, transformers offer the potential for modeling global dependencies crucial for understanding large-scale geographical patterns [3].

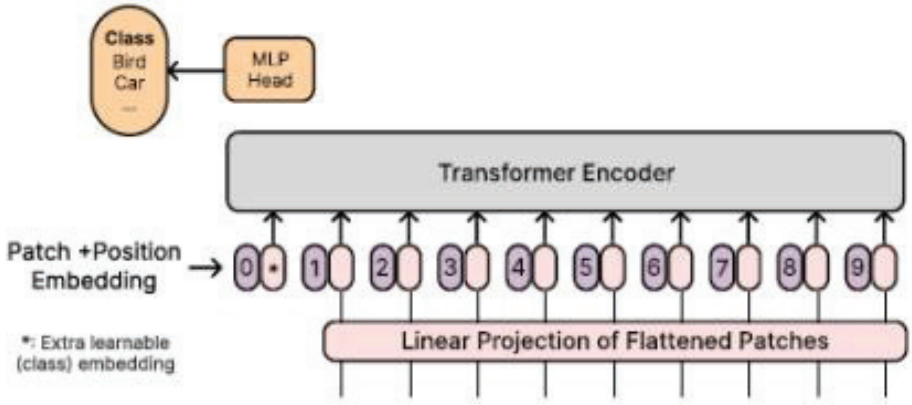

Fig. 1. The Vision transformer model architecture

However, the computational requirements and efficiency of trade-offs between these architectures remain poorly understood in the context of satellite imagery. This study addresses the gap in systematic architectural comparison for satellite image segmentation by evaluating the representative CNN and transformer models on the DeepGlobe Land Cover Classification dataset. The analysis encompasses segmentation accuracy and computational requirements on consumer-grade hardware, providing practitioners with empirical data for informed architecture selection.

Related works. The application of deep learning to satellite image analysis has evolved rapidly since the introduction of Fully Convolutional Networks (FCN) for semantic segmentation [4]. Long et al. demonstrated that adapting classification networks for dense prediction through fully convolutional architectures could achieve state-of-the-art results on natural images, inspiring subsequent applications to remote sensing. U-Net architecture, initially designed for biomedical image segmentation, succeeded in satellite imagery due to its efficient use of skip connections to preserve fine-grained spatial information [5]. Recent advances in CNN architecture have focused on improving both accuracy and efficiency. DeepLabv3+ introduced spatial pyramid pooling to capture multi-scale context, proving especially effective for satellite images where objects appear at vastly

different scales [6]. The FPN (Feature Pyramid Network) approach addresses similar challenges by constructing multi-scale feature pyramids, enabling accurate segmentation across different object sizes. These architectural innovations demonstrate the continued relevance of CNN-based approaches for satellite image analysis.

The introduction of ViTs marked a paradigm shift in computer vision, challenging the dominance of convolutional architectures [7]. Unlike CNNs that process images through local convolutions, ViTs divide images into patches and process them as sequences, enabling direct modeling of global dependencies. This fundamental difference allows ViTs to capture long-range spatial relationships more effectively, though at the cost of increased computational complexity with $O(n^2)$ attention operations. Several studies have explored transformer applications in remote sensing. [8] provides a comprehensive survey of transformers in remote sensing, highlighting their effectiveness for tasks requiring global context understanding. The hierarchical design of Swin Transformer, which introduces locality through shifted windows while maintaining global modeling capability, has shown promise for dense prediction tasks in satellite imagery [9]. Recent work by [10] demonstrated that sparse token transformers could reduce computational requirements while maintaining accuracy for building extraction tasks.

While numerous studies evaluate individual architectures on satellite imagery, comprehensive comparisons between CNN and transformer paradigms remain limited. Satellite image analysis systems often process terabytes of imagery daily, making inference speed and memory efficiency as important as accuracy. Zhang et al. highlighted that practical deployment considerations usually outweigh marginal accuracy improvements, particularly for resource-constrained applications [11]. Understanding these trade-offs becomes crucial for organizations with limited computational budgets or real-time processing requirements. Recent hybrid approaches attempt to combine CNN efficiency with transformer capabilities. ConvNeXt modernizes standard ConvNet designs using lessons learned from transformers, achieving competitive performance with improved efficiency [12]. Similarly, EfficientFormer proposes a hybrid architecture that leverages convolution and attention mechanisms, offering potential convergence between paradigms [13]. By evaluating models on consumer-grade hardware (RTX 4070) rather than high-end data center GPUs, the research provides insights relevant to a broader range of practitioners and deployment scenarios.

Research methodology

Dataset and preprocessing. The DeepGlobe Land Cover Classification dataset is used as an evaluation benchmark, providing 803 high-resolution satellite

images captured at 50 *cm* spatial resolution [14]. Each image spans 2448×2448 pixels and includes pixel-wise annotations for seven land cover classes: urban land, agriculture land, rangeland, forest land, water, barren land, and unknown regions. The dataset's geographical diversity covers multiple continents and climate zones, ensuring robust evaluation across varied landscapes. Data preprocessing followed established practices in satellite image analysis while considering computational constraints. Original images were resized to 512×512 pixels using bilinear interpolation to fit within GPU memory limitations while preserving sufficient detail for accurate segmentation. The resizing factor of approximately 4.8× reduces computational requirements by a factor of 23 while maintaining the essential spatial patterns needed for land cover classification. Pixel values were normalized to the [0, 1] range and standardized using ImageNet statistics, as all evaluated models utilized ImageNet pretraining. Data augmentation strategies specifically targeted the characteristics of satellite imagery. Random horizontal and vertical flips exploit the rotation-invariant nature of aerial perspectives, effectively quadrupling the training data diversity. Random rotations within ±30 degrees enhance rotational invariance while avoiding extreme angles that might disrupt semantic meaning. During training, these augmentations were applied with 50% probability to balance diversity with original data distribution preservation.

The model architecture and implementation. The architectural selection encompasses established representatives from both CNN and transformer families, chosen for their proven effectiveness and availability of pretrained weights. ResNet50 exemplifies the classical CNN approach with its residual connections enabling deep feature extraction without gradient degradation [15]. For segmentation, ResNet50 was employed as an encoder within a Feature Pyramid Network (FPN) framework, which constructs a multi-scale feature pyramid through top-down pathways and lateral connections. This design enables effective segmentation across different object scales, which is particularly important for satellite imagery where land cover regions vary dramatically in size.

EfficientNet-B0 represents the state-of-the-art in efficient CNN design, utilizing compound scaling to optimize depth, width, and resolution [16]. For B0, the base model uses 5.3M parameters compared to ResNet50's 25.6M, achieving comparable accuracy. The segmentation head follows a U-Net design with skip connections from intermediate encoder layers, preserving fine-grained spatial information crucial for precise boundary delineation.

Vision Transformer (ViT-B/16) adapts the standard transformer architecture to images by dividing them into non-overlapping patches of 16×16 pixels [7]. Each patch undergoes linear projection to create patch embeddings, with positional en-

codings added to preserve spatial information. Swin Transformer addresses ViT's computational limitations through a hierarchical architecture with shifted windows [9]. The shifted window approach restricts self-attention computation to local windows while enabling cross-window connections through alternating window partitions. This reduces complexity from $O(n^2)$ to $O(n)$ with respect to image size, making it more suitable for dense prediction tasks. The hierarchical structure progressively merges patches to create multi-scale representations analogous to CNN feature pyramids.

Training Configuration and Optimization. Training procedures balanced the computational efficiency with model convergence requirements. All models were initialized from ImageNet pretrained weights, leveraging transfer learning to reduce training time and improve final performance. The training process spanned 100 epochs on an NVIDIA RTX 4070 GPU with 8GB memory, requiring careful batch size selection to avoid out-of-memory errors. CNN models (ResNet50 and EfficientNet-B0) accommodated a batch size of 8, while transformer models required a reduction to a batch size of 4 due to higher memory consumption. An initial learning rate of 1e-4 with a cosine annealing schedule enabled smooth convergence.

Evaluation Methodology. A comprehensive evaluation employed multiple metrics capturing different aspects of segmentation quality. Mean Intersection over Union (mIoU) served as the primary metric, computing the average overlap between predicted and ground truth segments across all classes:

$$\mathrm{IoU}_i = \frac{|A_i \cap B_i|}{|A_i \cup B_i|}, \tag{1}$$

where C represents the number of classes; A_i - the predicted segment for class i, and B_i - the ground truth.

F1-score provides a balanced measure of precision and recall, particularly relevant for imbalanced classes:

$$\text{F1-Score} = \frac{2TP}{2TP + FP + FN}. \tag{2}$$

Pixel accuracy offers an intuitive overall performance measure:

$$PA = \frac{\sum_i TP_i}{\sum_i \sum_j n_{ij}}, \tag{3}$$

where n_{ij} represents the number of pixels of class i predicted as class j.

Results and discussion. The comparative evaluation reveals distinct performance characteristics across architectures, with Vision Transformers generally achieving superior segmentation accuracy at the cost of increased computational

requirements. Table 1 summarizes the primary performance metrics evaluated on the DeepGlobe test set.

Table 1

Overall segmentation performance metrics

Model	mIoU (%)	F1-Score (%)	Pixel Accuracy (%)	Training Time (hours)
ResNet50	70.6	82.3	86.7	18.5
EfficientNet-B0	71.6	83.5	87.5	16.2
ViT-B/16	73.1	84.3	88.2	42.3
Swin-T	74.4	85.2	89.1	35.7

Swin Transformer achieved the highest performance with 74.2% mIoU, representing a significant 3.9 percentage point improvement over the ResNet50 baseline. This performance gap demonstrates the effectiveness of global attention mechanisms for satellite image segmentation, where understanding large-scale spatial relationships proves crucial. The hierarchical structure of Swin-T appears particularly well-suited for capturing the multi-scale nature of land cover patterns, from fine-grained texture details to regional landscape structures. EfficientNet-B0's competitive performance deserves special attention, achieving 71.8% mIoU while requiring the least training time at 16.2 hours. The 1.5 percentage point improvement over ResNet50 with 76% fewer parameters validates the importance of architectural efficiency in model design.

Training time analysis reveals the computational burden of transformer architectures, with ViT-B/16 requiring 2.3× longer training than ResNet50. Swin-T's windowed attention mechanism provides meaningful efficiency gains, reducing the training time by 16% compared to ViT-B/16 while achieving superior performance.

Detailed analysis of class-wise IoU scores reveals interesting patterns in how different architectures handle various land cover types. Table 2 presents the breakdown across all seven classes.

Table 2

Class-wise IoU performance (%)

Model	Urban	Agriculture	Rangeland	Forest	Water	Barren	Unknown
ResNet50	75.2	78.4	62.1	71.3	89.7	66.8	48.9
EfficientNet-B0	76.8	79.2	64.3	72.7	90.1	67.9	51.2
ViT-B/16	77.5	80.6	65.8	74.2	91.3	68.4	52.8
Swin-T	79.1	81.9	67.2	75.8	92.1	69.6	54.3

Water bodies demonstrate the highest segmentation accuracy across all models, with IoU scores exceeding 89%. The distinctive spectral signature of water in both visible and near-infrared bands, combined with relatively homogeneous texture, makes it the easiest class to identify. Swin-T's 92.1% IoU on water represents near-perfect segmentation, with errors primarily occurring at boundaries where mixed pixels contain both water and land. Agricultural land also shows strong performance across architectures, benefiting from the regular geometric patterns of crop fields and distinctive vegetation indices. The consistent improvement from CNNs to transformers (78.4% to 81.9%) suggests that global context helps distinguish agricultural areas from natural vegetation by recognizing large-scale field patterns and irrigation structures. Rangeland proves the most challenging for all models, with the best performance reaching only 67.2% IoU with Swin-T. Practical deployment considerations often outweigh small accuracy differences, making computational efficiency analysis crucial for real-world applications. Table 3 presents detailed profiling results measuring various efficiency metrics.

Table 3

Computational resource requirements and efficiency metrics

Model	**Parameters (*M*)**	**Model Size (*MB*)**	**Inference Time (*ms*)**	**Memory (*MB*)**	**FLOPs (*G*)**
ResNet50	32.5	131	18.4	1842	8.2
Efficient-Net-B0	7.8	32	15.8	1456	2.4
ViT-B/16	89.7	361	36.8	2568	35.6
Swin-T	47.8	193	29.1	2234	18.9

EfficientNet-B0 emerges as the clear efficiency leader across all metrics. With only 7.8 *M* parameters and 32 *Mb* model size, it provides the most deployment-friendly option for edge devices or bandwidth constrained environments. The 15.7 *ms* inference time translates to 63.7 frames per second, exceeding real-time requirements for most applications. The computational complexity measured in FLOPs reveals the fundamental efficiency gap between CNNs and transformers. ViT-B/16 requires 35.6 GFLOPs per image, representing 14.8 times more computation than EfficientNet-B0's 2.4 GFLOPs. Swin-T's windowed attention reduces computational requirements by 47% compared to ViT-B/16 while maintaining transformer advantages.

Fig. 2 illustrates the throughput characteristics across different batch sizes within GPU memory constraints.

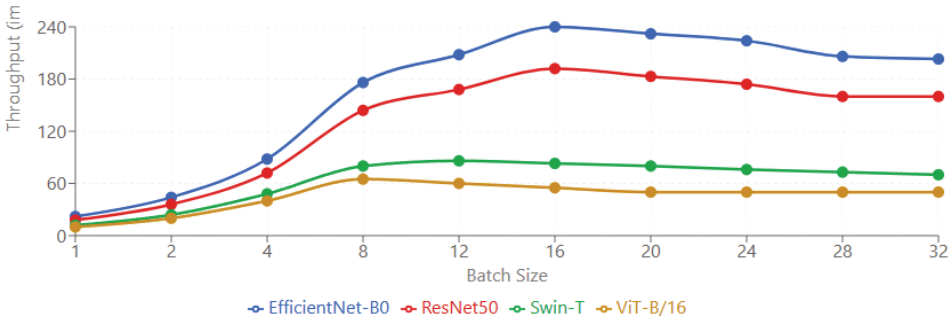

Fig. 2. Throughput (images/s) vs batch size for different architectures

The maximum achievable batch sizes reflect memory efficiency differences: EfficientNet-B0 supports batch size 32, ResNet50 manages 24, while transformers are limited to 12 (Swin-T) and 8 (ViT-B/16). These constraints directly impact throughput in production systems where batch processing improves GPU utilization. EfficientNet-B0 achieves peak throughput of 203 images/second, representing 3.1× improvement over ViT-B/16's 65 images/second.

Throughput scaling exhibits interesting non-linear characteristics. CNN architectures show near-linear scaling up to batch size 8, after which memory bandwidth limitations cause diminishing returns. Transformer models display earlier saturation due to the memory-intensive attention mechanisms. The crossover point where batch processing becomes beneficial occurs at batch size 2 for all models, suggesting single-image inference for real-time applications and batch processing for offline analysis.

Conclusion. This comprehensive evaluation of CNN and Vision Transformer architectures for satellite image segmentation reveals important trade-offs between accuracy and computational efficiency. Vision Transformers, particularly Swin-T with 74.2% mIoU, achieve superior segmentation performance by effectively modeling global spatial relationships crucial for land cover classification. The 3.9 percentage point improvement over ResNet50 demonstrates the value of attention mechanisms for understanding large-scale landscape patterns. However, the computational cost of transformer architecture remains substantial. ViT-B/16 requires 2.3× longer inference time and 40% more GPU memory than EfficientNet-B0. The classwise analysis reveals that architectural advantages vary by land cover type. While all models excel at water segmentation (>89% IoU), challenging classes like rangeland benefit most from transformers' global context modeling. Urban areas show particularly strong improvements with transformer architectures, suggesting their value for complex, multi-scale environments. Future research direc-

tions should explore hybrid architectures combining CNN efficiency with transformer capability.

References

1. **Zhang L., Zhang L., Du B.** Deep learning for remote sensing data: A technical tutorial on the state of the art // IEEE Geoscience and Remote Sensing Magazine. - 2016. - Vol. 4, No. 2. - P. 22-40.
2. Deep learning in remote sensing applications: A meta-analysis and review / **L. Ma, Y. Liu, X. Zhang, Y. Ye,** et al // ISPRS Journal of Photogrammetry and Remote Sensing. - 2019. - Vol. 152. - P. 166-177.
3. **LeCun Y., Bengio Y., Hinton G.** Deep learning // Nature. - 2015. - Vol. 521, No. 7553. - P. 436-444.
4. **Long J., Shelhamer E., Darrell T.** Fully convolutional networks for semantic segmentation // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. - 2015. - P. 3431-3440.
5. **Ronneberger O., Fischer P., Brox T.** U-net: Convolutional networks for biomedical image segmentation // International Conference on Medical Image Computing and Computer-Assisted Intervention. - 2015. - P. 234-241.
6. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs / **L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, et al** // IEEE Transactions on Pattern Analysis and Machine Intelligence. - 2017. - Vol. 40, No. 4. - P. 834-848.
7. An image is worth 16x16 words: Transformers for image recognition at scale /**A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, et al** // arXiv preprint arXiv:2010.11929. - 2021.
8. Transformers in remote sensing: A survey / **A.A. Aleissaee, A. Kumar, R.M. Anwer, S. Khan, et al** // Remote Sensing. - 2023. - Vol. 15, No. 7. - P. 1860.
9. Swin transformer: Hierarchical vision transformer using shifted windows / **Z. Liu, Y. Lin, Y. Cao, H. Hu, et al** // Proceedings of the IEEE/CVF International Conference on Computer Vision. - 2021. - P. 10012-10022.
10. **Chen K., Zou Z., Shi Z.** Building extraction from remote sensing images with sparse token transformers // Remote Sensing. - 2021. - Vol. 13, No. 21. - P. 4441.
11. Transformers in vision: A survey / **S. Khan, M. Naseer, M. Hayat, S.W. Zamir, et al** // ACM Computing Surveys. - 2022. - Vol. 54, No. 10s. - P. 1-41.
12. EfficientFormer: Vision transformers at mobilenet speed / **Y. Li, G. Yuan, Y. Wen, E. Hu, et al** // Advances in Neural Information Processing Systems. - 2022. - Vol. 35. - P. 12934-12949.
13. Deepglobe 2018: A challenge to parse the earth through satellite images / **I. Demir, K. Koperski, D. Lindenbaum, G. Pang, et al** // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. - 2018. - P. 172-181.

14. **He K., Zhang X., Ren S., Sun J.** Deep residual learning for image recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. - 2016. - P. 770-778.
15. **Tan M., Le Q.** Efficientnet: Rethinking model scaling for convolutional neural networks // International Conference on Machine Learning. - 2019. - P. 6105-6114.
16. **Loshchilov I., Hutter F.** Decoupled weight decay regularization // arXiv preprint arXiv:1711.05101. - 2019.

Received on 23.05.2025.
Accepted for publication on 10.07.2025.

ՓԱԹՈՒՅԹԱՅԻՆ ՆԵՅՐՈՆԱՅԻՆ ՑԱՆՑԵՐԻ ԵՎ ՏԵՍՈՂԱԿԱՆ ՓՈԽԱԿԵՐՊԻՉՆԵՐԻ ՀԱՄԵՄԱՏԱԿԱՆ ՎԵՐԼՈՒԾՈՒԹՅՈՒՆ ԱՐԲԱՆՅԱԿԱՅԻՆ ՊԱՏԿԵՐՆԵՐԻ ՀԱՏՎԱԾԱՎՈՐՄԱՆ ՀԱՄԱՐ

Տ.Բ. Խաչատրյան

Խոր ուսուցման արագ զարգացումը հեղափոխություն է առաջացրել արբանյակային պատկերների վերլուծության ոլորտում, որտեղ ինչպես փաթույթային նեյրոնային ցանցերը (ՓՆՑ), այնպես էլ տեսողական փոխակերպիչ նեյրոնային ցանցերը (ՏՓՆՑ) դրսևորում են խոստումնալից արդյունքներ: Սույն ուսումնասիրությունը ներկայացնում է այս ճարտարապետությունների համապարփակ համեմատական վերլուծություն արբանյակային պատկերների հատվածավորման համար՝ լուծելով իրական տեղակայման սցենարներում արդյունավետության-կատարողականության փոխզիջումների ըմբռնման կրիտիկական անհրաժեշտությունը: Հարմարեցվել և գնահատվել են չորս ժամանակակից մոդելներ՝ ResNet50 և EfficientNet-B0՝ ներկայացնելով ՓՆՑ-ները, և ViT-B/16 ու Swin Transformer՝ ներկայացնելով ՏՓՆՑ-ների ընտանիքը: Օգտագործելով DeepGlobe հողածածկի դասակարգման տվյալների բազան, որը պարունակում է 803 բարձր լուծաչափով արբանյակային պատկերներ՝ յոթ հողածածկի դասերով, մոդելների կատարողականը գնահատվել է բազմաթիվ չափանիշներով, ներառյալ միջին Intersection over Union (mIoU), F1-գնահատականը և պիքսելային ճշտությունը: Բացի այդ, հաշվողական պահանջները, ներառյալ եզրակացության արագությունը, մոդելի չափը և հիշողության սպառումը, վերլուծվել են NVIDIA RTX 4070 GPU-ի վրա՝ նմանակելով գործնական տեղակայման սահմանափակումները: Արդյունքները ցույց են տալիս, որ թեև ՏՓՆՑ-ները հասնում են ավելի բարձր հատվածավորման ճշտության՝ Swin-T-ն հասնելով 74.2% mIoU՝ համեմատած EfficientNet-B0-ի 71.8%-ի հետ, այնուամենայնիվ, ՓՆՑ-ները պահպանում են զգալի առավելություններ եզրակացության արագության և հիշողության արդյունավետության հարցերում: EfficientNet-B0-ն մշակում է պատկերները 2.3 անգամ ավելի արագ, քան ViT-B/16-ը՝ օգտագործելով 40%-ով պակաս GPU հիշողություն: Դասային վերլուծությունը բացահայտում է, որ ՏՓՆՑ-ները հատկապես գերազանցում են բարդ սցենարներում, ինչպիսիք են քաղաքային տարածքները և անտառային սահմանները, մինչդեռ բոլոր մոդելները հասնում են 89%-ից ավելի IoU ջրային մարմինների սեգմենտացման համար: Այս գտածոները տրամադրում են գործնական

պատկերացումներ համապատասխան ճարտարապետության ընտրության համար՝ հիմնված կոնկրետ տեղակայման սահմանափակումների վրա, ընդգծելով ճշգրտության և հաշվողական արդյունավետության միջև փոխզիջումները արբանյակային պատկերների վերլուծության կիրառություններում՝ շրջակա միջավայրի մոնիտորինգի, քաղաքային պլանավորման և աղետների արձագանքման համար:

Առանցքային բառեր: արբանյակային պատկերների հատվածավորում, տեսողական փոխակերպիչ, փաթույթային նեյրոնային ցանցեր, խոր ուսուցում, հողածածկի դասակարգում:

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ И ВИЗУАЛЬНЫХ ТРАНСФОРМЕРОВ ДЛЯ СЕГМЕНТАЦИИ СПУТНИКОВЫХ ИЗОБРАЖЕНИЙ

Т.Б. Хачатрян

Стремительное развитие глубокого обучения произвело революцию в анализе спутниковых изображений, где как сверточные нейронные сети (СНС), так и визуальные трансформеры демонстрируют многообещающие результаты. В данном исследовании представлен всесторонний сравнительный анализ этих архитектур для сегментации земного покрова на спутниковых снимках, решающий критическую потребность в понимании компромиссов между производительностью и эффективностью в реальных сценариях развертывания. Настроены и оценены четыре современные модели: ResNet50 и EfficientNet-B0, представляющие СНС, и ViT-B/16 и Swin Transformer, представляющие семейство трансформеров. Используя набор данных DeepGlobe для классификации земного покрова, содержащий 803 спутниковых изображения высокого разрешения с семью классами земного покрова, производительность моделей оценивалась по множеству метрик, включая среднее пересечение над объединением (mIoU), F1-оценку и попиксельную точность. Кроме того, вычислительные требования, включая скорость вывода, размер модели и потребление памяти, были проанализированы на GPU NVIDIA RTX 4070 для моделирования практических ограничений развертывания. Результаты демонстрируют, что хотя визуальные трансформеры достигают превосходной точности сегментации, при этом Swin-T достигает 74,2% mIoU по сравнению с 71,8% для EfficientNet-B0, СНС сохраняют значительные преимущества в скорости вывода и эффективности памяти. EfficientNet-B0 обрабатывает изображения в 2,3 раза быстрее, чем ViT-B/16, используя на 40% меньше памяти GPU. Поклассовый анализ показывает, что трансформеры особенно превосходят в сложных сценариях, таких как городские районы и лесные границы, в то время как все модели достигают более 89% IoU для сегментации водных объектов. Эти результаты предоставляют практические идеи для выбора подходящей архитектуры на основе конкретных ограничений развертывания, подчеркивая компромиссы между точностью и вычислительной эффективностью в приложениях анализа спутниковых изображений для мониторинга окружающей среды, городского планирования и реагирования на стихийные бедствия.

Ключевые слова: сегментация спутниковых изображений, визуальный трансформер, сверточные нейронные сети, глубокое обучение, классификация земного покрова.