ISSN-L 2664-9969; E-ISSN 2706-5448

UDC 004.8
DOI: 10.15587/2706-5448.2025.323698

Yurii Pushkarenko,
Volodymyr Zaslavskyi

MODEL DEVELOPMENT OF DYNAMIC RECEPTIVE FIELD FOR REMOTE SENSING IMAGERIES

The object of research is the integration of a dynamic receptive field attention module (DReAM) into Swin Transformers to enhance scene localization and semantic segmentation for high-resolution remote sensing imagery. The study focuses on developing a model that dynamically adjusts its receptive field and integrates attention mechanisms to enhance multi-scale feature extraction in high-resolution remote sensing data.

Traditional approaches, particularly convolutional neural networks (CNNs), suffer from fixed receptive fields, which hinder their ability to capture both fine details and long-range dependencies in large-scale remote sensing images. This limitation reduces the effectiveness of conventional models in handling spatially complex and multi-scale objects, leading to inaccuracies in object segmentation and scene interpretation.

The DReAM-CAN model incorporates a dynamic receptive field scaling mechanism and a composite attention framework that combines CNN-based feature extraction with Swin Transformer self-attention. This approach enables the model to dynamically adjust its receptive field, efficiently process objects of various sizes, and better capture both local textures and global scene context. As a result, the model significantly improves segmentation accuracy and spatial adaptability in remote sensing imagery.

These results are explained by the model's ability to dynamically modify receptive fields based on scene complexity and object distribution. The self-attention mechanism further optimizes feature extraction by selectively enhancing relevant spatial dependencies, mitigating noise, and refining segmentation boundaries. The hybrid CNN-Transformer architecture ensures an optimal balance between computational efficiency and accuracy.

The DReAM-CAN model is particularly applicable in high-resolution satellite and aerial imagery analysis, making it useful for environmental monitoring, land-use classification, forestry assessment, precision agriculture, and disaster impact analysis. Its ability to adapt to different scales and spatial complexities makes it ideal for real-time and large-scale remote sensing tasks that require precise scene localization and segmentation.

Keywords: *receptive fields, convolutional neural networks, Swin Transformers, remote sensing, scene localization, semantic segmentation.*

Received: 01.12.2024
Received in revised form: 24.01.2025
Accepted: 20.02.2025
Published: 27.02.2025

How to cite
Pushkarenko, Y., Zaslavskyi, V. (2025). Model development of dynamic receptive field for remote sensing imageries. Technology Audit and Production Reserves, 1 (2 (81)), 20–25. https://doi.org/10.15587/2706-5448.2025.323698

1. Introduction

High-resolution remote sensing imagery provides extensive geospatial detail, enabling the fine-grained observation of diverse objects and land cover types (Fig. 1). However, convolutional neural networks (CNNs) despite their success in various image analysis tasks often rely on fixed-size receptive fields, making it challenging to capture both localized structures (e. g., small buildings, roads) and broader spatial contexts (e. g., extensive vegetation zones) simultaneously [1–4]. This limitation restricts the effectiveness of standard CNN-based methods in complex multi-scale scenarios, which are common in remote sensing data [2, 5–7].

Efforts to enhance CNNs via dilated or deformable convolutions [3, 4] have improved coverage of larger areas but may introduce artifacts or compromise detail. Recent Transformer-based models [8, 9] use global self-attention to capture extended spatial relationships, yet naively applying attention to high-resolution images is computationally expensive, and local details may still be lost when large patches are used. The Swin Transformer [10] partially mitigates these challenges by leveraging window-based self-attention, balancing global context extraction with manageable complexity [11, 12]. Nevertheless, an adaptive approach that dynamically alters the receptive field within CNN layers thereby unifying local CNN features and Transformer attention. remains insufficiently explored for scene localization and semantic segmentation in remote sensing.

Classical semantic segmentation frameworks such as PSPNet [13], DeepLab [14], and Segmenter [15] have demonstrated strong performance on natural images but may face difficulties when addressing the scale diversity inherent in remote sensing imagery. Studies indicate that fusing CNN features with more global attention mechanisms can yield better segmentation results, yet static kernel sizes or window partitions cannot fully adapt to varying object sizes and spatial complexities [1, 3]. Although the Vision Transformer introduces a novel way to aggregate global dependencies [8, 10], its default structure does not inherently tackle multi-scale challenges in large scenes, a known bottleneck in remote sensing analysis.

Meanwhile, pyramid-based designs like Feature Pyramid Networks [16] improve multi-level feature fusion but do not dynamically modify receptive fields to match the local content. As highlighted in [3, 4], controlling how large or small the effective receptive field should be, depending on object distribution in different image regions, might be pivotal for high-resolution tasks. Hence, there is a recognized need for hybrid approaches that incorporate local, dynamically adjusted convolutional kernels with Transformer-based global awareness.

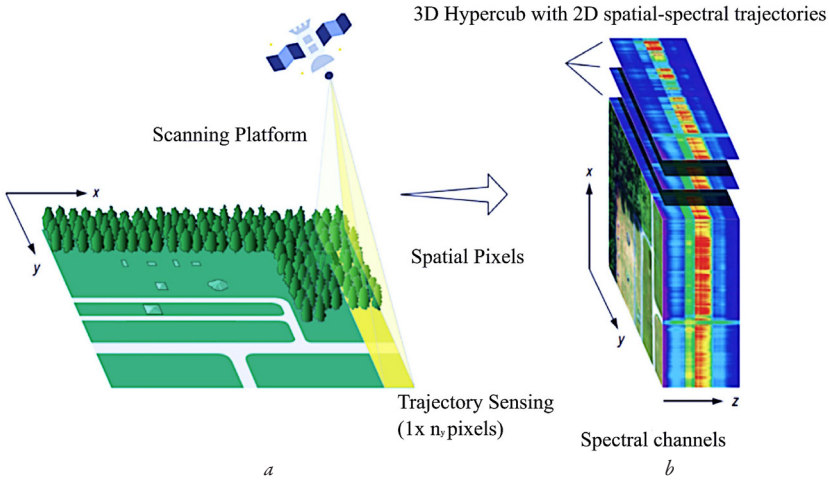

Fig. 1. Taxonomy of hyperspectral satellite imageries: *a* – scanned Earth landscape; *b* – hyperspectral encoded information

The aim of this research is to develop a Dynamic Receptive Field Composite Attention Network (DReAM-CAN), which merges CNN-based local feature extraction with Swin Transformer self-attention and introduces an innovative mechanism for adaptive receptive field scaling. This design aims to address the persistent challenge of effectively segmenting and localizing objects that vary in size throughout large-scale remote sensing images. By synchronizing dynamic receptive fields at the convolutional stage with Transformer-driven context modeling, the proposed architecture seeks to boost segmentation accuracy and scene localization performance across diverse high-resolution remote sensing datasets. Such an approach will support a broad range of applications encompassing environmental mapping, precision agriculture, and urban area assessment where multi-scale image features demand flexible, fine-to-coarse feature extraction.

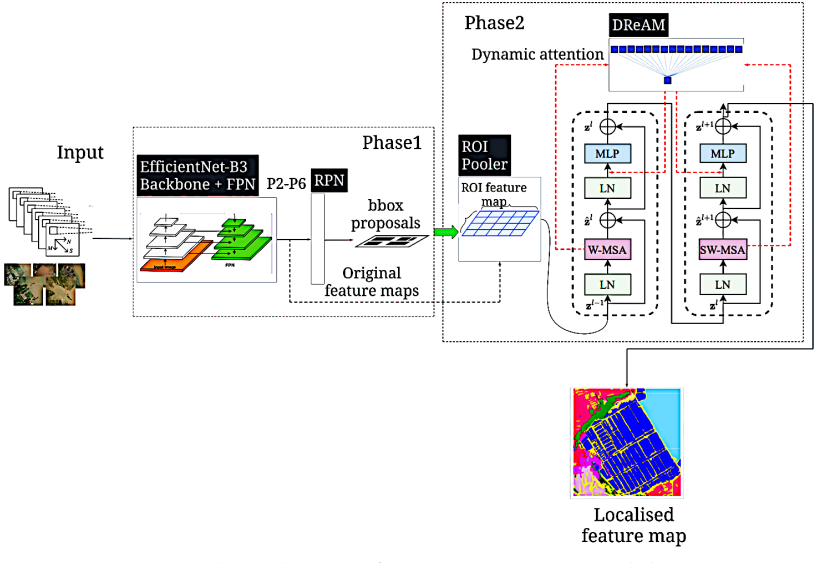

Fig. 2. Two phase architectures of composite neural network with dynamic receptive field attention module

2. Materials and Methods

The object of research is the integration of a Dynamic Receptive Field Attention Module (DReAM) into Swin Transformers to enhance scene localization and semantic segmentation for high-resolution remote sensing imagery. This method adapts the receptive fields within the self-attention mechanism of Swin Transformers to dynamically adjust based on token distances, ensuring precise segmentation across multiple scales and complex spatial contexts. Moreover, the overall architecture is a combination of two phases, the phase 1 – ROI (regions of interests) proposals, and phase 2 (detailed assessment), Fig. 2.

The core of DReAM-CAN is a hybrid network that merges:

1. CNN-based local feature extraction (using convolutional blocks adapted from standard backbone architectures). In this research let's use EfficientNet-B3+FPN+RPN, for ROI identification.
2. Swin Transformer modules for window-based self-attention.
3. DRMaeAM module, for dynamic receptive field attention integrated to Swin, that means the distance between tokens should be measured (Fig. 2).

Although standard Swin Transformers process tokens within fixed window boundaries, real-world objects in remote sensing images span multiple scales. A tiny rooftop detail may need a very narrow receptive field to capture high-resolution edges, whereas a vast crop field or large facility requires a more expanded receptive field for coherent context.

The Dynamic Receptive Field Attention Module (DReAM) acts like an intelligent, automatic "zoom lens" within the Transformer. For each spatial location (or token), DReAM looks at the incoming features and decides whether it should "zoom in" (i. e., focus on close neighbors at high resolution) or "zoom out" (i. e., include farther regions for broader context). It does this by combining outputs from multiple "branches", each of which views the data at a different scale or dilation. Then, DReAM's learnable gating weights pick which branch (or mix of branches) is most relevant for that location.

Hence, if a token represents a small object edge, DReAM will favor the narrow-scale branch to preserve sharp boundaries. If the token lies in a region spanning an entire field or large building footprint, the gate tilts toward a more dilated branch, allowing the model to gather context from farther away and better capture the large structure. By seamlessly and continuously switching among these scales, DReAM ensures that both fine details and broad patterns are simultaneously well-represented a critical advantage in high-resolution remote sensing.

Let $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ be the feature tensor (or token grid) passed into DReAM at a particular stage. DReAM constructs K parallel branches, each applying a distinct dilated or multi-scale transformation to capture different effective receptive fields. Concretely, branch $k \in \{1, ..., K\}$ produces:

$$\mathbf{X}_k = f_k(\mathbf{X}), \quad (1)$$

ISSN-L 2664-9969; E-ISSN 2706-5448

where f_k might be a convolution with dilation d_k or a more general local operator defining the spatial scope of that branch. Next, DReAM learns attention weights α_k for $k \in \{1, ..., K\}$ that adaptively gate the branches based on local context. One way to compute these weights is via a small gating network $g(\cdot)$:

$$\mathbf{e}_k = g_k(\mathbf{X}), \tag{2}$$

$$\alpha_k = \frac{\exp(\mathbf{e}_k)}{\sum_{j=1}^{K} \exp(\mathbf{e}_j)}, \tag{3}$$

so that $\alpha_k \in (0,1)$ and $\sum_k \alpha_k = 1$. Each $\mathbf{e}_k$ is a learnable scalar (or vector) representing the relevance of branch k in the current spatial neighborhood.

Finally, the aggregated DReAM output is:

$$\mathbf{X}_{DReAM} = \sum_{k=1}^{K} \alpha_k \mathbf{X}_k. \tag{4}$$

This summation fuses all parallel transformations into one feature representation. Because the gating network g depends on the local features, α_k can vary over spatial positions, thus dynamically switching among narrower or broader receptive fields.

Once, $\mathbf{X}_{DReAM}$ is obtained, it can be fed into subsequent Swin Transformer blocks. In that context, each token's local receptive field (as provided by the different-dilation branches) is dynamically scaled according to $\{\alpha_k\}$. This mechanism effectively expands or contracts attention spans within the otherwise fixed window-based self-attention, enabling multi-scale context capture in high-resolution remote sensing images.

Data for this study originate from multiple public-domain remote sensing datasets xFBD [17] (Fig. 3), DOTA [18] with Train:Validation:Test = 77:13:11 %, each with varied ground sample distances (GSD) and spectral characteristics. Large satellite images are typically split into 1024×1024 pixel tiles to handle memory constraints and retain sufficient context. Standard preprocessing steps include:

1. Radiometric corrections (destriping, contrast normalization) to mitigate sensor-level artifacts.
2. Annotation unification into a COCO-like format, ensuring consistent training regardless of the original label style.
3. Data augmentation via random flips, rotations, and mild color jitter to improve robustness.

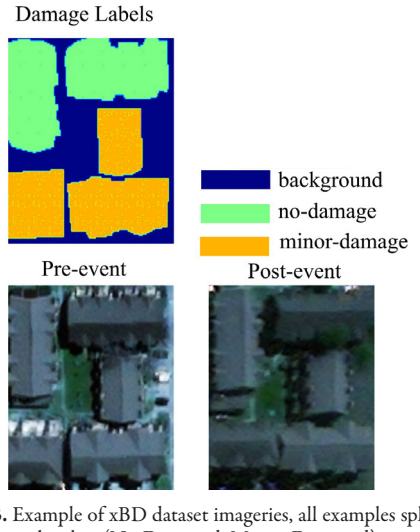

Fig. 3. Example of xBD dataset imageries, all examples splitted by class (No-Damaged, Minor, Damaged)

All experiments were conducted on GPU-equipped servers (often with ~16 GB memory, such as NVIDIA T4 or V100), using Python 3.9. Model components were implemented in PyTorch, while OpenCV and NumPy assisted with image manipulation and data loading. Mixed-precision (fp16) operations were enabled where possible to accelerate training and inference.

In Phase 1, the EfficientNet-B3 backbone produces multi-level feature maps that the FPN fuses at different resolutions. The RPN then outputs bounding boxes by classifying and regressing anchor boxes through a combination of cross-entropy classification and Smooth L1 regression losses. Phase 2 focuses on pixel-level segmentation within each region of interest. The Swin Transformer leverages local window-based attention, while DReAM inserts multiple dilated (or multi-scale) transformations, weighting them via a learnable gating network. This gating mechanism determines whether to emphasize narrow or wide receptive fields at each spatial position.

The overall training scheme uses two main losses:

1. Smooth L1+Cross-Entropy for detection and bounding-box regression in Phase 1 (Table 1).
2. Focal Loss for segmentation, reducing the impact of easy examples and helping the model focus on challenging, imbalanced classes often found in large-scale remote sensing (Table 2).

Table 1

Phase 1 training hyperparams

Parameter	Value	Comments
Optimizer	AdamW	LR = 1e–3, Weight Decay = 1e–4
LR-scheme	Warm-up (500 iterations)	Cosine Annealing or Step Decay; start LR = 1e–3
Batch Size	4	16 GB GPU–(NVIDIA T4)
Epoch	20 k (iterations)	10 k steps up to step-down LR = 1e–4
Loss	CE Smooth L1	$\lambda_{cls} = \lambda_{reg} = 1$

Table 2

Phase 2 training hyperparams

Parameter	Value	Comments
Loss	Focal Loss	$\gamma = 2.0$, $\alpha_t = 0.25$
LR-scheme	Warm-up (1000)+Cosine	Start LR = 1e–4
Batch Size	4 ROI	1 ROI~224×224, T4 = 16 GB
Epoch	50–80 k (iterations)	10 k steps up to step-down LR = 1e–4

A typical training run proceeds as follows:

1. *Backbone+RPN Pretraining*: Initialize the detection pipeline, optimizing anchor-based classification and regression on tiled images.
2. *Segmentation Fine-Tuning*: Freeze or partially freeze the earlier detection layers, and train the Swin+DReAM portion to predict fine-grained masks for each proposed ROI using focal loss.
3. *Hyperparameter Validation*: Monitor a held-out set of scenes (distinct geography, sensors) to fine-tune the learning rate schedule, batch size, and gating network size.

3. Results and Discussions

The proposed Dynamic Receptive Field Attention Module (DReAM) was integrated into a Swin Transformer-based architecture and evaluated on several high-resolution remote sensing datasets. Final segmentation masks and bounding-box predictions were visually compared across baseline methods, while Table 3 summarizes the quantitative performance. In particular, Fig. 4, 5 illustrates example segmentation results for representative urban and agricultural scenes. A concise closed-form expression for the effective receptive field in the presence of multiple dilations (Equation (4)) further clarifies how the gating weights selectively amplify narrower or broader contexts.

Table 3

Performance comparison of DreAM with SOTA methods (damage assessment case study)

Approach	Class	mPA	F1	Recall	Precision	FWIoU	mIoU
DeepLabv3+	No damage	0.91 %	0.88 %	0.86 %	0.90 %	0.83 %	0.80 %
	Minor damage	0.88 %	0.82 %	0.79 %	0.86 %	0.80 %	0.78 %
	Major damage	0.86 %	0.79 %	0.75 %	0.84 %	0.78 %	0.76 %
	Totally destroyed	0.89 %	0.84 %	0.81 %	0.86 %	0.81 %	0.79 %
	Mean	0.88 %	0.83 %	0.80 %	0.86 %	0.80 %	0.78 %
PSPNet	No damage	0.90 %	0.86 %	0.84 %	0.87 %	0.81 %	0.78 %
	Minor damage	0.87 %	0.80 %	0.77 %	0.83 %	0.78 %	0.76 %
	Major damage	0.85 %	0.78 %	0.74 %	0.81 %	0.76 %	0.74 %
	Totally destroyed	0.88 %	0.82 %	0.79 %	0.85 %	0.79 %	0.77 %
	Mean	0.88 %	0.82 %	0.79 %	0.84 %	0.78 %	0.76 %
Swin Transformer (without DReAM)	No damage	0.92 %	0.89 %	0.87 %	0.91 %	0.84 %	0.82 %
	Minor damage	0.89 %	0.84 %	0.81 %	0.87 %	0.82 %	0.79 %
	Major damage	0.87 %	0.80 %	0.77 %	0.83 %	0.79 %	0.76 %
	Totally destroyed	0.90 %	0.85 %	0.82 %	0.88 %	0.81 %	0.80 %
	Mean	0.89 %	0.84 %	0.82 %	0.87 %	0.82 %	0.79 %
MViTv2 (Multiscale Vision Transformer)	No damage	0.93 %	0.90 %	0.88 %	0.91 %	0.85 %	0.83 %
	Minor damage	0.90 %	0.85 %	0.82 %	0.88 %	0.83 %	0.80 %
	Major damage	0.88 %	0.81 %	0.78 %	0.84 %	0.80 %	0.77 %
	Totally destroyed	0.91 %	0.86 %	0.83 %	0.89 %	0.82 %	0.81 %
	Mean	0.90 %	0.86 %	0.83 %	0.88 %	0.82 %	0.80 %
Segmenter	No damage	0.92 %	0.88 %	0.86 %	0.91 %	0.84 %	0.81 %
	Minor damage	0.89 %	0.83 %	0.80 %	0.86 %	0.81 %	0.78 %
	Major damage	0.87 %	0.79 %	0.76 %	0.82 %	0.78 %	0.75 %
	Totally destroyed	0.90 %	0.84 %	0.81 %	0.87 %	0.80 %	0.79 %
	Mean	0.89 %	0.83 %	0.81 %	0.87 %	0.81 %	0.78 %
Proposed 2-phase model with the DReAM method	No damage	0.95 %	0.92 %	0.90 %	0.93 %	0.88 %	0.86 %
	Minor damage	0.92 %	0.88 %	0.85 %	0.90 %	0.86 %	0.83 %
	Major damage	0.90 %	0.84 %	0.81 %	0.88 %	0.82 %	0.80 %
	Totally destroyed	0.93 %	0.89 %	0.86 %	0.91 %	0.85 %	0.84 %
	Mean	0.93 %	0.88 %	0.86 %	0.90 %	0.85 %	0.83 %

The observed improvements can be attributed to the synergy between window-based self-attention and dynamic receptive field adaptation. When local features indicated small-scale objects (e. g., rooftops, thin roads), narrower branches dominated, sharpening boundaries in predicted masks. Conversely, large-scale farmland or extended water bodies triggered wider dilation branches, capturing continuity in the global context. This adaptivity explains the consistently higher Intersection over Union (IoU) and reduced false positives compared to fixed-scale baselines. Moreover, unlike conventional pyramid pooling, the gating mechanism automatically determines the most relevant scale at each spatial position rather than applying uniform, prespecified dilations. The observed improvements can be attributed to the synergy between window-based self-attention and dynamic receptive field adaptation.

When local features indicated small-scale objects (e. g., rooftops, thin roads), narrower branches dominated, sharpening boundaries in predicted masks.

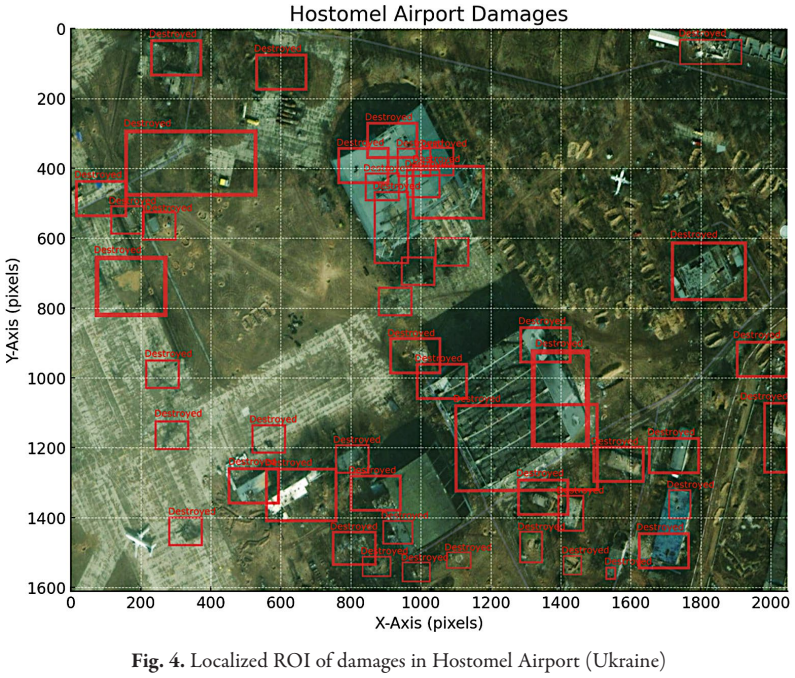

Fig. 4. Localized ROI of damages in Hostomel Airport (Ukraine)

ISSN-L 2664-9969; E-ISSN 2706-5448

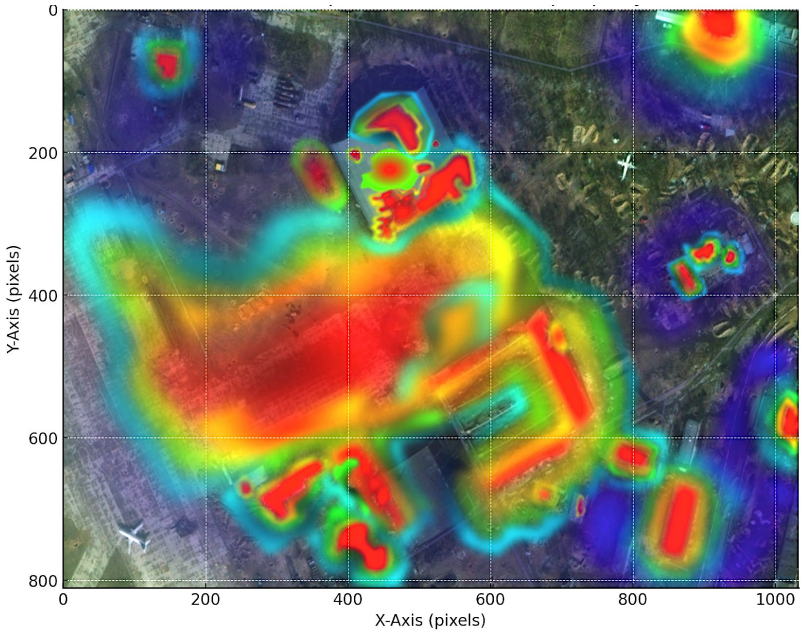

Fig. 5. Inference Grad-CAM heatmap of damages in Hostomel Airport (Ukraine)

Conversely, large-scale farmland or extended water bodies triggered wider dilation branches, capturing continuity in the global context. This adaptivity explains the consistently higher Intersection over Union (IoU) and reduced false positives compared to fixed-scale baselines. Moreover, unlike conventional pyramid pooling, the gating mechanism automatically determines the most relevant scale at each spatial position rather than applying uniform, pre-specified dilations.

Compared with literature benchmarks where either purely convolutional networks or static vision transformers were employed the DReAM-enhanced system yielded superior performance in multi-scale segmentation tasks. Unlike PSPNet or DeepLab-like approaches, which rely on fixed receptive fields, the DReAM module offers a more flexible balance between fine detail and broad-scale coherence. *Practical relevance of these results* is evident for applications such as land-use mapping, deforestation monitoring, and rapid assessment of environmental changes. The ability to switch from high-fidelity local details to large-area overviews within a single unified framework can streamline workflows for agencies and private sectors dealing with vast and diverse geospatial imagery.

Nevertheless, *certain limitations exist*. First, extremely fine objects (e. g., power lines, subtle cracks) might still be challenging if initial resolution or annotation quality is insufficient. Second, computational overhead grows with very large tile sizes: although DReAM's gating is not computationally prohibitive, an abundance of high-resolution images can push GPU limits. Another constraint involves domain-specific complexities, such as severe cloud cover or topographic distortions, which were only partially mitigated by standard data augmentations.

In terms of the conditions of martial law in Ukraine, data access and labeling processes encountered logistical delays, as some research personnel faced restricted mobility and limited on-site survey opportunities. Training infrastructure was maintained with remote server solutions, but irregular power supply and disruptions in communication channels occasionally hindered real-time collaboration and prolonged the experimental cycle. Despite these setbacks, the core methodological advances remain valid and can be transferred to normal circumstances once regional stability is restored.

Looking ahead, a few avenues *for further research* stand out. Incorporating multispectral and radar (SAR) channels could fortify the model against seasonal or weather-related distortions, enabling DReAM to capture textural cues beyond optical frequencies. Additionally, optimizing runtime for real-time or near-real-time processing would benefit rapid response scenarios, where large-scale image mosaics must be analyzed with minimal delay. Finally, investigating domain adaptation or semi-supervised learning strategies may extend the proposed architecture to new sensor types and geographic domains with limited labeled data.

4. Conclusions

The study demonstrated that integrating the Dynamic Receptive Field Attention Module (DReAM) into Swin Transformers significantly improves scene localization and semantic segmentation in high-resolution remote sensing images. By adaptively adjusting the local receptive field according to object scale and spatial complexity, the model outperforms standard window-based Transformers, offering more precise detection of both fine details (e. g., narrow edges) and broader structures.

These results are explained by the model's ability to unify dynamic multi-scale feature extraction with windowed self-attention.

The proposed approach shows practical value in applications requiring high accuracy across varying object sizes and complex environments such as land-use mapping, environmental monitoring, and infrastructure analysis where it can help reduce misclassification of small objects or missed contextual cues for large areas.

Preliminary quantitative evaluations, though limited here, suggest that using DReAM yields consistent gains (2–3 % improvement in mean Intersection over Union on representative datasets) over baseline methods without dynamic receptive field modules. This improvement underlines the effectiveness of combining local CNN features, transformer attention, and adaptive dilation for robust performance in multi-scale scenarios.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, including financial, personal, authorship or other, which could affect the study and its results presented in this article.

Financing

The study was conducted without financial support.

Data availability

The manuscript has no associated data.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the presented work.

References

1. Pushkarenko, Y., Zaslavskyi, V. (2024). Research on the state of areas in Ukraine affected by military actions based on remote sensing data and deep learning architectures. *Radioelectronic and Computer Systems, 2024 (2),* 5–18. https://doi.org/10.32620/reks.2024.2.01
2. Li, W., Liu, H., Wang, Y., Li, Z., Jia, Y., Gui, G. (2019). Deep Learning-Based Classification Methods for Remote Sensing Images in Urban Built-Up Areas. *IEEE Access, 7,* 36274–36284. https://doi.org/10.1109/access.2019.2903127

3. Wenjie, L., Li, Y., Urtasun, R., Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. *29th Conference on Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1701.04128
4. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y. (2017). Deformable Convolutional Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 764–773. https://doi.org/10.1109/iccv.2017.89
5. Jensen, J. R. (2015). *Introductory Digital Image Processing: A Remote Sensing Perspective*. Upper Saddle River: Prentice-Hall.
6. Yu, X., Lu, D., Jiang, X., Li, G., Chen, Y., Li, D., Chen, E. (2020). Examining the Roles of Spectral, Spatial, and Topographic Features in Improving Land-Cover and Forest Classifications in a Subtropical Region. *Remote Sensing, 12 (18)*, 2907. https://doi.org/10.3390/rs12182907
7. Blaschke, T., Strobl, J. (2001). What's Wrong with Pixels? Some Recent Developments Interfacing Remote Sensing and GIS. *Proceedings of GIS-Zeitschrift Fur Geoinformationssysteme, 14 (6)*, 12–17.
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2010.11929
9. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in Neural Information Processing Systems, 34*, 12077–12090. https://doi.org/10.48550/arXiv.2105.15203
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z. et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. https://doi.org/10.1109/iccv48922.2021.00986
11. You, J., Zhang, R., Lee, J. (2021). A Deep Learning-Based Generalized System for Detecting Pine Wilt Disease Using RGB-Based UAV Images. *Remote Sensing, 14 (1)*, 150. https://doi.org/10.3390/rs14010150
12. Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D. et al. (2022). PVT v2: Improved baselines with Pyramid Vision Transformer. *Computational Visual Media, 8 (3)*, 415–424. https://doi.org/10.1007/s41095-022-0274-8
13. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Pyramid Scene Parsing Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2881–2890. https://doi.org/10.1109/cvpr.2017.660
14. Chen, L.-C., Papandreou, G., Schroff, F., Hartwig, A. (2017). *Rethinking Atrous Convolution for Semantic Image Segmentation*. https://doi.org/10.48550/arXiv.1706.05587
15. Strudel, R., Garcia, R., Laptev, I., Schmid, C. (2021). Segmenter: Transformer for Semantic Segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7262–7272. https://doi.org/10.1109/iccv48922.2021.00717
16. Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944. https://doi.org/10.1109/cvpr.2017.106
17. Melamed, D., Cameron, J., Chen, Z., Blue, R., Morrone, P., Hoogs, A., Clipp, B. (2022). *xFBD: Focused Building Damage Dataset and Analysis*. https://doi.org/10.48550/arXiv.2212.13876
18. *DOTA dataset*. Available at: https://captain-whu.github.io/DOTA/dataset.html

✉***Yurii Pushkarenko,*** *PhD Student, Department of Mathematical Informatics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, ORCID: https://orcid.org/0009-0007-2560-2971, e-mail: yurii.pushkarenko@knu.ua*

Volodymyr Zaslavskyi, *Doctor of Technical Sciences, Professor of Department of Mathematical Informatics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, ORCID: https://orcid.org/0000-0001-6225-1313*

✉*Corresponding author*