UDC 528.8:004.93:355.422(477) **doi: 10.32620/reks.2024.2.01**

Yurii PUSHKARENKO, Volodymyr ZASLAVSKYI

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

RESEARCH ON THE STATE OF AREAS IN UKRAINE AFFECTED BY MILITARY ACTIONS BASED ON REMOTE SENSING DATA AND DEEP LEARNING ARCHITECTURES

The invasion of Ukraine by the Russian Federation and the escalation of military actions in the regions have led to significant damage to residential buildings, civilian infrastructure, various critical infrastructure objects, dams, and extensive pollution of the territories. In this context, the tasks of remote sensing using satellite imagery and aerial observation arise to analyze the impact and conduct an economic assessment of damage in these areas. This work investigates and employs deep neural network (DNNs) models in computer vision (CV) tasks (classification, segmentation) and combines their derivatives, such as convolutional networks (CNNs) and vision transformer models (ViTs), to enhance the accuracy of damage assessment. ViTs have demonstrated significant success, often surpassing traditional CNNs, and have potential applications in remote sensing for damage assessment and the protection of critical infrastructure. The research conducted in this work confirms the importance of applying such technologies in environments where labeled data are rare or non-existent, particularly evaluating the use of DNNs, including CNNs and ViTs, in analyzing regions affected by military actions using synthetic aperture radar (SAR) and multispectral images. The aim and subject of this research also include reviewing the possibilities of combining CNNs and ViTs to improve the speed of image feature extraction, landscape detection, and the detection of complex structural contours of objects, where data are usually insufficient. The results of this study provide a critical review of the application of CNNs and ViTs in remote sensing, identifying significant gaps and challenges, especially in the context of the economic consequences of destruction due to military actions. The technical aspects of using CNNs and transformer-based models for complex CV tasks and transfer learning under data-scarce conditions, as well as the challenges in analyzing large volumes of geophysical data, are considered. The conclusions emphasize the transformational potential of DNNs, especially transformers, in remote sensing under conflict and disaster conditions. Their adaptability and accuracy in various environments underscore their utility in both strategic military and humanitarian contexts, establishing a practical standard for their application in key real, real-world scenario-based territory condition assessment.

Keywords: *vision transformers; computer vision; data fusion; remote sensing; CNN; damage assessment.*

1. Introduction

1.1. Motivation

The 2022 Russian aggression in Ukraine demonstrates the profound impact that extensive military confrontations can have on individuals, environments, and economic structures. Within just a month, this aggression led to the uprooting of nearly 10 million people, resulted in damage exceeding US$100 billion to infrastructure, and raised alarms about potential global wheat supply disruptions [1]. Satellites, many commercially managed, are increasingly documenting the aftermath of the most significant European conflict since the Second World War. Very-high-resolution (VHR <5 m) satellite imagery has spotlighted the 64 km Russian vehicle procession near Kyiv, the extensive damage in Mariupol, the destruction of the Kakhovka Dam, and water resource issues (the assessment method was proposed by V. Zaslavskyi et.al. in [2]), and extensive civilian vehicle queues at border checkpoints (Fig. 1). High-resolution imagery is being produced by private and public satellite operators in near real time, and this information is being used to track troop movements, verify attacks in inaccessible areas, assess infrastructure damage, and document possible war crimes. However, access to such data often comes at the cost of conditions or steep prices, limiting its availability to researchers and humanitarian organizations. For instance, acquiring imagery from Maxar's WorldView-4, which offers sub-meter resolution, would cost approximately $22.50 per square kilometer, translating to an exorbitant US$13.6 million for a nation as vast as Ukraine. Conversely, a plethora of lower-resolution satellite data, stemming from publicly supported initiatives, is freely available. Freely accessible data have played a crucial role in identifying and tracking significant landscape transformations, including those precipitated by conflict, such as urban development, deforestation, and shifts in agriculture.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

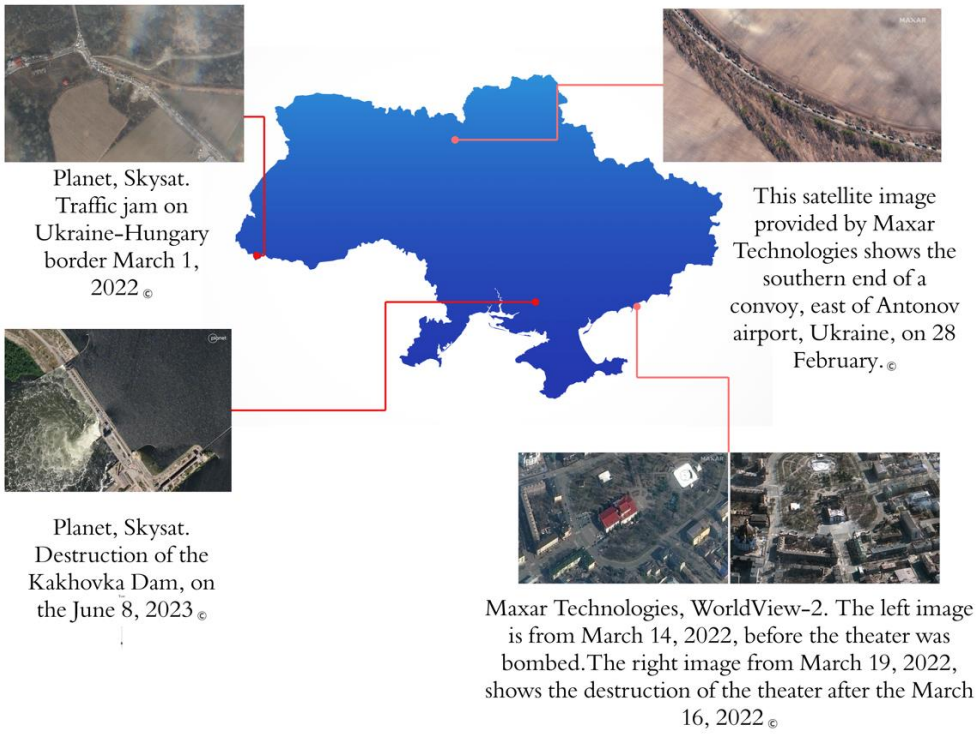

Fig. 1. VHR imageries provide a detailed view of the situation in Ukraine during the first days of full-scale invasion. Sources: Planet, Skysat, and Maxar Tech., WorldView-2. Inspired by [1]

These observations, whether captured through SAR (synthetic-aperture radar), VHR, or Multi-Spectral imagery, even if not of the highest resolution, highlight the potential to make war imagery more universally accessible for scientific analysis, intelligence gathering, and humanitarian efforts.

The overview in this work aims to cover several main topics that are dependent on each other:

1. Assessment of war consequences tasks can expedite decision-making at tactical and strategic levels. For instance:

1.1. Multimodal analysis: combining the strengths of different imaging modalities (SAR, VHR multispectral, hyperspectral) with captioning can provide a more comprehensive understanding. For instance, SAR might detect metal objects under foliage, optical imaging can provide color details, and multispectral can provide material insights. A combined caption might read, "Metallic object, possibly a vehicle, camouflaged under trees with a green tarp (Fig. 2).

1.2. Real-time Tactical Decision Making under high risk and uncertainty: This involves classifying military vehicle types, their numbers, and troop movements to provide commanders with real-time data for making informed tactical decisions [3].

1.3. Terrain Analysis & Pattern Recognition: This entails detecting patterns such as military formations or routine patrols and understanding the nature of the terrain to plan troop movements, set up bases, or strategize defenses.

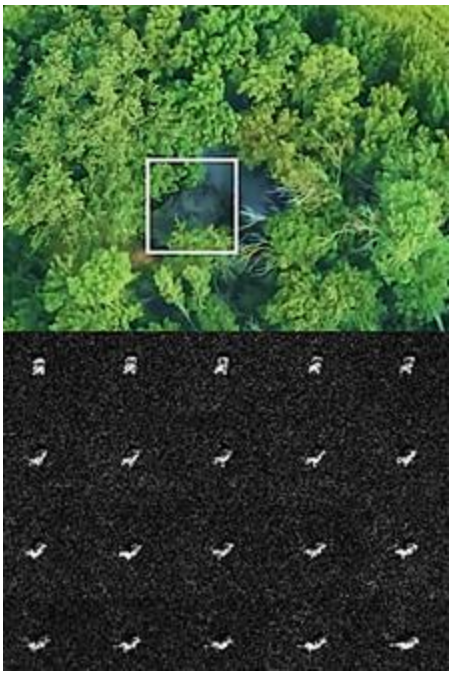

Fig. 2. VHR aerial image of a camouflaged tank in white bounding box captured by the State Border Guard Service of Ukraine and SAR's masked T72 tank objects from MSTAR dataset. Source: https://www.sdms.afrl.af.mil/

1.4. Damage assessment – this aspect can be exceptionally valuable for various purposes, including:

1.4.1. Extracted building footprints: this can be used to compare pre-conflict and post-conflict imagery, thereby aiding in the quantification and visualization of infrastructure damage.

1.4.2. building structures assists ground troops during urban warfare by providing detailed maps indicating potential shelters, ambush points and vantage positions.

1.4.3. Relocation and Evacuation: Identifying undamaged buildings can help create safe zones, relocation centers, or medical hubs.

1.4.4. Critical infrastructure protection: Critical infrastructure, including transportation networks and bridges, is often deliberately targeted during wars and natural disasters. This is because such infrastructure plays a crucial role in maintaining connectivity and facilitating the movement of people and goods, thereby supporting national and international economic development.

2. Attention Mechanisms as potential applications of war consequence analysis, specifically focusing on the following aspects:

2.1. Handle large-scale variations. Can dynamically adjust its focus to different scales, which is ideal for satellite imagery where objects can vary greatly in size (e.g., from individual vehicles to entire buildings).

2.2. Provides end-to-end learning, eliminating the need for manual feature engineering and revealing novel features relevant to the task.

1.2. State of the Art

The intersection of deep learning, primarily ConvNets, and remote sensing has witnessed considerable advancements over the past decade, especially for damage assessment [4] since the latest earthquake in Turkey, as well as military consequences analysis [5] and military vehicles detection, and situational awareness [6]. This section highlights some pivotal works that have laid the foundation in these areas, specifically focusing on accelerating attention towards war consequences and near-real-time tactical/strategic decision-making. For example, Huang et al. (2023) [7] employed classical ConvNets to detect war-induced infrastructural damage in Mariupol's case. However, such approaches have limitations in terms of capturing intricate patterns and long-range dependencies, which are crucial for nuanced assessments like war consequence analysis. Transition to transformer architectures and their application in computer vision tasks were described by Dosovitskiy et al. (2021) [8] due to their capability to capture 16x16 size patches and their contextual information and provide a more holistic understanding of scenes. This makes transformers particularly suitable for complex tasks [9] that extend beyond simple object detection in satellite imageries, such as assessing the aftermath of military activities. While ViTs have found diverse applications, there remains a noticeable vacuum in leveraging them to assess military and war consequences, particularly inregions affected bygeopolitical events. Recent eventsin Ukraine underscore the importance of this area ofstudy. To the best of the authors' knowledge, research explicitly focusing on this intersection is limited.

The literature includes methods for evaluating various catastrophes, such as war aftermath and terrorism, and presents algorithms for proactive disaster protection of critical infrastructure. The most recent study by V. I. Norkin et al. (2018) investigated the stochastic, informational, and behavioral uncertainties in aggressive actions against Ukraine's critical infrastructure. This research applies a bilevel stochastic min-max game problem, which is detailed in [10]. In future, we plan to focus on utilizing these developed methods to create robust DNN models, particularly by framing hyperparameter optimization in DNNs and ViTs as a problem.

The exploration of the concept of multimodality in solving optimization problems, as proposed by H. Yailymova et al. in [11], centers on diversity and type diversity, which is referred to as multimodality in this paper. This review examines the proposed method in the context of integrating various sources of truth, including SAR, VHR, HSR, HS, and MS imageries, and merging DNNs with Vision Transformers (ViT) to achieve an optimized effect. Significant contributions by V. Kharchenko. et al. [12] introduced new mathematical methods and qualitative analysis techniques for imagery and high-volume data processing, which are crucial for tuning hyperparameters in CNNs and ViTs.

1.3. Objectives and approach

This work comprehensively reviews transformer-related advances in remote sensing and their potential applications in the context of war.

The primary objectives are as follows:

1. Assess Transformer-Based Models in Remote Sensing:

– the applicability of transformer-based models to Synthetic Aperture Radar (SAR), Very High Resolution (VHR), and multispectral context-based image analysis (CBIA) in remote sensing;

– investigate the use of Vision Transformers (ViTs) to address challenges related to limited labeled data and enhance image captioning and real-time tactical decision-making;

2. Comparative Analysis:

– a comparative analysis between Convolutional Neural Networks (CNNs) and ViTs to highlight their

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

respective advantages and limitations in the context of imaging and analysis of military tasks.

3. Literature Review:

– examine existing transformer-based studies to identify the latest advancements and potential applications in remote sensing, particularly in war-affected areas.

4. Identify Research Challenges:

– explore various challenges and potential research trajectories related to the application of transformers in remote sensing, with a focus on situational awareness and damage assessment.

The methods used in this research are provided by a set of review approaches, summarized below:

1. Review Transformer Advancements:

– to conduct an extensive review of the recent literature on transformer models, particularly their application in computer vision tasks such as classification and segmentation;

– compare the performance of transformer models with traditional CNNs in remote sensing applications.

2. Conduct Comparative Analysis of CNNs and ViTs:

– provide a detailed comparison of CNN and ViT architectures and highlight their individual advantages and limitations;

– discuss the scalability and flexibility of ViTs relative to capturing global interactions and modeling data nuances compared to content-independent CNN operations.

3. Highlight Key Transformer Architectures:

– discuss important transformer architectures such as Vision Transformers (ViTs), Conditional ViTs, and Detection Transformers (DETR);

– evaluate their potential applications in remote sensing, especially in assessing military and war consequences.

4. The most popular datasets were reviewed, and challenges in potential applications.

2. Methodology

The proposed methodology involves several sequential steps to preprocess the data, extract features, train models, and perform evaluation (Fig.3):

1. Data Preprocessing;

1.1. Normalize the input images to standardize pixel values;

1.2. Noise Reduction:

$$I_{smoothed} = I' * G, \quad (1)$$

where G is a Gaussian kernel;

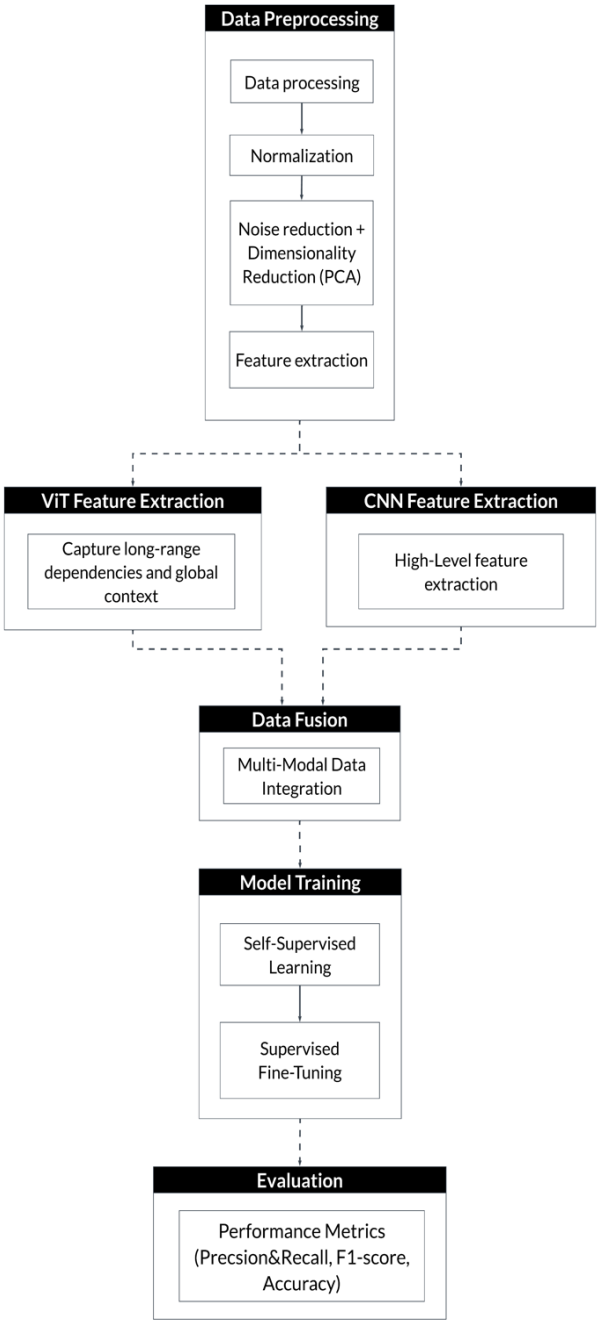

Fig. 3. The main flowchart describes the methodology of a chain of transformations and the fusion of image data by combining optical and SAR data and applying them to attention mechanisms.the attention mechanisms

$I' = \frac{I-\mu}{\sigma}$ is normalization formula;

μ is the mean and σ is the standard deviation of the pixel values in I.

1.3. Dimensionality Reduction (PCA):

1.3.1. Reduce the dimensionality of image data to minimize redundancy and computational load;

1.3.2. Compute covariance matrix C, find eigenvalues and eigenvectors, and project data onto the principal components.

2. Feature Extraction;

2.1. Using Convolutional Neural Networks (CNNs);

2.1.1. The high-level features are extracted from images via convolution operations:

$$(\mathbf{W} * I')(i,j) = \\ = \sum_{m=1}^{f}\sum_{n=1}^{f} \mathbf{W}(m,n)\cdot I'(i+m-1, j+n-1), \quad (2)$$

where $(\mathbf{W} * I')(i,j)$ – the corresponding convolutional layer;

Activation as $\mathrm{ReLU}(x) = \max(0,x)$;

Pooling Layer as,

$P(I,j) = \max_{0\le m<p, 0\le n<p} I'(i+m, j+n)$;

And Fully Connected Layer is described as follows,

$o = W_{fc}\cdot f + b_{fc}$.

2.2. Using vision transformers (ViTs);

2.2.1. Capture long-range dependencies and global contexts using self-attention mechanisms;

2.2.2. Patch Embedding: The image is split into patches, and each patch is projected into a lower-dimensional space;

2.2.3. Transformer Encoder: This encoder applies multi-head self-attention and feed-forward networks to learn contextual relationships;

2.2.4. Classification Head: The final representation is aggregated to produce class predictions.

3. Multi-Modal Data Integration;

3.1. Data Fusion. Combine different types of remote sensing data (e.g., SAR, multi-spectral, hyper-spectral) to enhance robustness:

$$I_f = \alpha\cdot I_1 + (1-\alpha)\cdot I_2, \quad (3)$$

where I_f – the resulting fused image or dataset that combines the information from I_1, I_2;

I_1 – the first input image or dataset, which could be, for example, a SAR image;

I_2 – the first input image or dataset, which could be, for example, a SAR image;

α and $\alpha - 1$ – the complementary weighting factors that determines the contribution of I_1 and I_2 to the fused image.

4. Model Training;

4.1. Self-Supervised Learning:

4.1.1. Pre-training models on unlabeled data to generate task-agnostic latent representations through learning representations:

$$\mathbf{h} = f_\theta(I'), \quad (4)$$

where h denotes the latent representation or feature vector generated by the model;

f_θ The DNN model (e.g., a CNN or ViT) parameterized by θ, where θ represents the learned weights of the model;

I' the preprocessed input image.

4.1.2. Fine-tune pre-trained models using labeled data for specific tasks through some loss function.

5. Evaluate model performance using metrics such as accuracy, precision, recall, and F1-score, etc.

3. Methods of transfer from CNNs to Vision Transformers

CNNs (ConvNets) have become the standard for computer vision tasks, particularly with optical imagery (Fig. 4), including VHRs and High Spectral Resolution (HSR) images. Consider a high-dimensional input space: $I \in \mathbb{R}^{H\times W\times C}$.

The objective is to transform input I into a representation suitable for classification or semantic segmentation tasks. The ConvNet uses layers of convolutions, each represented by a kernel tensor: $K \in \mathbb{R}^{h\times w\times C_{in}\times C_{out}}$, which convolves the input volume to produce feature maps.

The operation for a single layer is defined as follows:

$$F_l(I) = \sigma(K * I + b), \quad (5)$$

where b denotes the bias vector;

σ represents an element-wise non-linear;

activation function;

F_l is the output feature map of the l-th layer.

Subsequent layers, including pooling layers, are applied recursively to produce increasingly abstract representation.

Pooling operations reduce the spatial dimensions (H,W) to (H',W'), often using functions like max or average pooling. From other side vision transformers decompose I into a sequence of flattened 2D patches P (Fig. 5) where each patch is linearly embedded and positional encodings. The resulting sequence $X \in \mathbb{R}^{N\times D}$ is processed through the self-attention mechanism in each transformer block B as follows:

$$B(X) = \mathrm{MHA}\big(\mathrm{LN}(X)\big) + \mathrm{MLP}\Big(\mathrm{LN}\big(B(X)\big)\Big), \quad (6)$$

where MHA is Multi-Head Attention;

LN is the Layer Normalization;

MLP is the Multilayer Perceptron.

The MHA, in other cases MHSA (Multi-Head-Self-Attention), the multi-head self-attention is defined as follows:

$$\mathrm{MHSA}(X) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O, \quad (7)$$

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

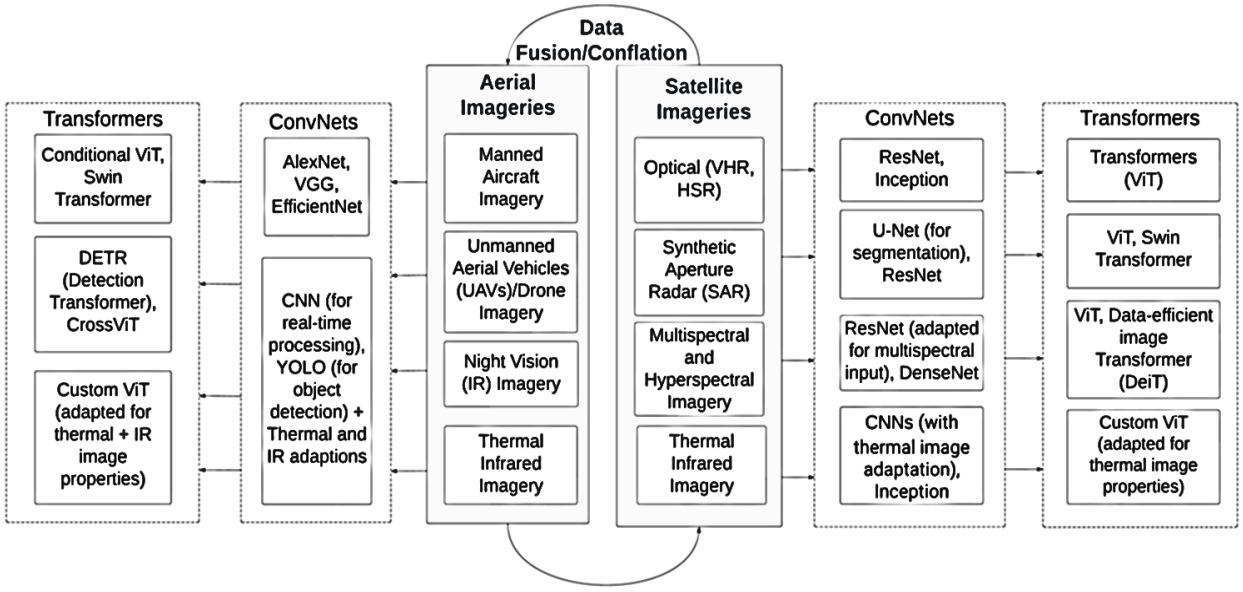

Fig. 4. The platform-based imagery sources taxonomy and corresponding possible DNNs architectures

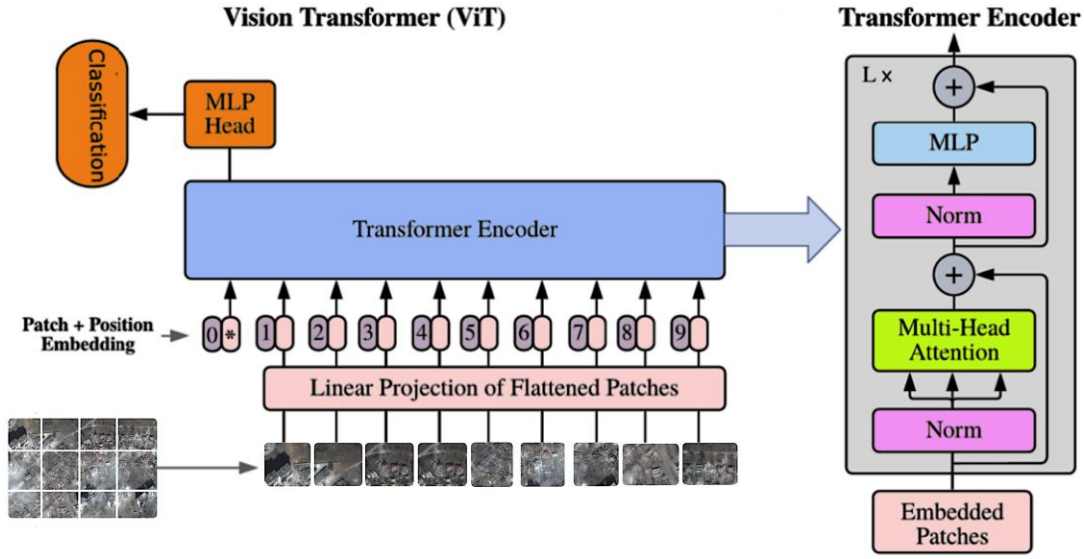

Fig. 5. The ViT Architecture [8]

each head_i in the MHSA is computed independently using the scaled dot-product attention function [8]:

$$\text{head}_i = \text{Attention}\left(XW_i^Q, XW_i^K, XW_i^V\right) \quad (8)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d^k}}\right)V, \quad (9)$$

where Q, K, V are the query, key, and value matrices which are projections of the input X;

$W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{D \times d_k}$ are the parameter matrices for the query, key, and value for the i-th;

$W^O \in \mathbb{R}^{hd_k \times D}$ is the output projection matrix, where i is the number of heads and d_k is the dimensionality of each head's output.

The attention output for each head is concatenated and projected to match the original embedding dimensionality D.

Layer normalization (LN) and position-wise feed-forward networks (FFN) are also critical in the transformer's encoder block [8]:

$$\text{LN}(x) = \gamma \odot x - \frac{\mu}{\sigma} + \beta, \quad (10)$$

where μ, σ are the mean and standard deviation computed across the feature dimension of x;

γ, β are trainable scale and shift parameters.

Following the LN and MHSA, the output Z undergoes another layer normalization and then a position-wise FFN:

$$\mathrm{FFN}(Z) = \max(0, ZW_1 + b_1)\, W_2 + b_2, \quad (11)$$

where W_1, W_2 are weight matrices and b_1, B_2 are bias vectors of the FFN.

$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{1000^{\frac{2i}{D}}}\right), \quad (12)$$

Positional encoding (PE) is added to the input embeddings to retain the positional information of the image patches as follows:

$$\mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{1000^{\frac{2i}{D}}}\right), \quad (13)$$

where pos is the position and i is the dimension.

Training ViTs with Masked Self-Attention. During training with masked self-attention, certain positions in the input sequence are masked (denoted by M) to prevent the model from attending to them:

$$A_{masked} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d^k}} + M\right) V, \quad (14)$$

The mask M applies a large negative value to masked positions prior to SoftMax operation, effectively zeroing out their contributions.

Attention remains focused on the main transformer architectures (backbones) in the context of assessing war-affected regions, with an exploration of the latest transformer-based backbones. ViT is an architecture that directly applies a pure transformer to sequences of image patches for image classification tasks [8] and serves as a fundamental baseline in the industry. This ViT architecture does not incorporate typical image-specific assumptions, such as translation equivariance and locality, as well as more specific applications, such as control frames in SAR or VHR imagery. It undergoes pre-training on extensive datasets like ImageNet21k or JFT-300M, which are not applicable in real-world cases where time and performance are crucial.

Conditional ViTs. A notable study in the realm of Conditional ViTs is the introduction of Conditional Positional Encodings (CPE) for ViTs, which was explored by Xiangxiang Chu et al. [13]. The proposed method differs from traditional fixed or learnable positional encodings by dynamically generating CPEs conditioned on the local neighborhood of the input tokens. This method allows for better generalization to longer input sequences than observed during training and maintains translation invariance in image classification tasks, which leads to improved accuracy. The CPEs are implemented using a Position Encoding Generator (PEG), seamlessly integrating into the existing Transformer framework, resulting in a CPVT model that achieves state-of-the-art results on the ImageNet classification task. For remote sensing, especially for aerial imagery, improved performance in [13] can potentially be applied to UAV which have demonstrated technology success in context the Russian-Ukrainian War.

Since the topic of UAV was touched in terms of forces operation, it is worth mentioning the *Detection*

The Transformer (DETR) model, as presented in a publication by Nicolas Carion et al [14], is a novel approach to object detection. DETR views object detection as a direct set prediction problem, streamlining the detection pipeline and eliminating the need for components like non-maximum suppression and anchor generation. The model, based on a transformer encoder-decoder architecture, achieves comparable accuracy and run-time performance with established baselines like Faster R-CNN, on challenging datasets like COCO. DETR can be extended to tasks like panoptic segmentation. The remainder of this paper provides detailed insights into the architecture, training settings, and performance of DETR. Generally, DETR can be divided into several simplified subtasks:

Bipartite Matching Loss: Let us suppose that, given ground truth objects $\mathbf{y}$ and predicted objects $\widehat{(\mathbf{y})}$, the matching cost between a ground truth object y_i and prediction $\hat{y}_j$ is defined as follows:

$$C(y_i, \hat{y}_j) = -1_{\{c_i \neq \emptyset\}}\, \hat{p}_\theta(c_i \mid \hat{y}_j) 1_{c\{c_i \neq \emptyset\}} \mathcal{L}_{box}(b_i, \hat{b}_j), \quad (15)$$

where c_i and b_i are the class label and bounding box, respectively, for the ground truth;

$\hat{p}_\theta$ is the predicted probability for class c_i;

$\mathcal{L}_{box}$ is a box loss (like generalized IoU (the Intersection over Union)).

Transformer Encoding/Decoding: For an input feature map $\mathbf{z}$, the transformer encoder output $\mathbf{z}'$ is computed as follows:

$$\mathbf{z}' = \mathrm{TransformerEncoder}(\mathbf{z}). \quad (16)$$

The transformer decoder then processes a fixed number of objects queries $\mathbf{q}$, and the encoder output is expressed as follows:

$$\hat{\mathbf{y}} = \mathrm{TransformerEncoder}(\mathbf{q}, \mathbf{z}'). \quad (17)$$

Prediction Head: Each output of the decoder is fed into a feed-forward network (FFN) to predict the class and bounding box:

$$\mathrm{class}(\hat{y}_j), \mathrm{box}(\hat{y}_j) = \mathrm{FFN}(\hat{y}_j). \quad (18)$$

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

DETR simplifies the object detection pipeline by directly predicting a set of objects and employing transformers to model the global relationships and dependencies between these objects. This is crucial when using drones for object detection tasks.

The Swin Transformers, introduced by Liu et al. [15], represent a novel computer vision architecture that addresses the challenges of adapting Transformers from language to vision. This architecture features a hierarchical structure calculated using shifted windows, thereby enhancing efficiency by limiting self-attention to local, non-overlapping windows and allowing cross-window connections. The Swin Transformer exhibits linear computational complexity relative to image size, making it suitable for a broad range of vision tasks. The performance of the proposed model surpasses previous models on key benchmarks like COCO and ADE20K. The main advantage of Swin Transformers is their linear computational complexity. Unlike global self-attention, which has a quadratic complexity with respect to the number of tokens, the complexity of self-attention within local windows is linear with respect to image size. The computational complexity of a window-based MSA on an image with N patches is approximately $O(N)$, given a fixed number of patches in each window. This is applicable to SAR imageries where analysis can also include attention to speckle noise.

4. Results of the analysis

The central idea behind this review is to understand perspectives DNNs and their impact on studying situational awareness and critical infrastructure defense in war-affected regions of Ukraine using different remote sensing sources that have been described in the taxonomy section.

Remote data from remote sensing have been crucial for conveying critical information on damage following natural disasters, focusing on surface impacts and infrastructural damage. In contrast to natural disasters, the worsening regional security in armed conflicts has complicated field research and damage assessment by international teams, with low-altitude flights being hindered by airspace restrictions. Consequently, the challenge of rapidly acquiring precise, up-to-date information on humanitarian crises, evaluating losses, and guiding global efforts in conflict resolution, humanitarian assistance, and rebuilding efforts has become a critical issue that the international community urgently needs to address amid these conflicts.

TThe review focused on multiple cases in which DNNs, primarily ViTs, were combined with different remote sensing sources. In addition, potential methods to identify damage in images include object detection and segmentation techniques. For example, ConvNets have been employed to identify global views and classify various types of damage but not a subject of damage. However, current techniques frequently regard damage as a broad concept rather than a distinctly defined object, leading to conceptual discrepancies. To tackle this challenge, the current work describes existing damage detection strategies that utilize transformers, particularly ViT, to leverage the attention mechanism at the pre-training level. To delve deeper into the subject, summarizing the available data for addressing this issue, we outline the sources actively covering Ukraine.

Medium resolution optical images (e.g., Landsat, ASTER, and Sentinel-2) have provided data for a more general interpretation of damaged areas in disaster zones. Although these images, typically ranging from 10 to 30 m, may not be suitable for detailed damage assessment, they are effective for gaining a comprehensive understanding of the overall damage situation. However, this research did not address medium-spatial resolution due to contextual reasons.

High and Very High Spectral Resolution optical imageries (HSR, VHR) also provided by public and commercial organizations (e.g., IKONOS, QuickBird, Geo-Eye, and the WorldView series) have enabled building-scale damage detection after a disaster using pixel or object-based change detection techniques using CNNs. The upcoming work is also focused on the self-attention mechanisms and fusion of CNN and ViT. These models excel in applications like damage assessment, due to their capacity to autonomously identify hierarchical feature levels in images, from basic to complex. Numerous studies have exploited deep learning to evaluate structural damage, marking significant advancements in this field. For example, Abdi et al. [16] classified building damage into four categories using CNN-based UAV imagery post-hurricane. Similarly, Zhang et al. [17] utilized pre- and post-disaster imagery to create a comprehensive method for building damage assessment via semantic segmentation. Gaining situational awareness can be challenging using only optical sensors because they offer clear imagery primarily in daylight and under cloud-free conditions. This study, specifically focusing on , encounters a significant challenge: from October to March, cloud cover is common across much of the country, which complicates the analysis of optical remote sensing sources.

4.1. Experiments for HSR and VHR datasets

xView2 xBD Dataset [18] is the largest optical satellite imagery benchmark dataset available for building segmentation and damage assessment in remote sensing communities. This dataset, part of Maxar's Open Data program, provides high-resolution satellite imagery

that is crucial for emergency planning and damage assessment. The model comprises annotated images featuring polygons and damage scores for buildings affected by natural disasters. Encompassing 18,336 images from 15 countries, the dataset spans six disaster types and includes over 850,000 polygons across 45,000 square kilometers. Notably, the dataset captures scenes both before and after disasters, intending to facilitate risk assessment and response research. However, note that the dataset is unbalanced towards the "no-damage" classification (Fig. 6).

Fig. 6. Example of xView Dataset.
Source: https://xview2.org/

Custom\manual datasets, in most cases mined from Google Earth service, sometimes such datasets were collected for augmentation needs or to provide a source of truth (Fig. 7).

Microwave SAR offers an alternative that does not have these constraints. Thus, various SAR sensors (L, C, and X-band SAR images provided by ALOS-2 PALSAR-2, RADARSAT-2, Sentinel-1, TerraSAR-X, and the COSMO-SkyMed constellation) (Fig. 8) are being incorporated into remote sensing disaster response.

In SAR imagery, segmentation can be challenging due to the appearance of speckles, which is a type of multiplicative noise that increases with increasing back-scattering radar magnitude. Many new backbones have been proposed recently GCBANet [19].

Meanwhile, Xia et al. [20] introduced the CRTransSar model, which integrates CNNs and transformers, effectively capturing both detailed and broad perspectives for object detection in SAR images. The proposed model employs a structure that combines attention mechanisms with convolutional layers to enhance object detection performance.

Sentinel-1, a satellite mission for Earth observation equipped with Synthetic Aperture Radar (SAR) provides medium resolution (~10m) C-band (λ = 5.6 cm) SAR measurements with dual polarization, allowing data acquisition during night-time or through cloud cover (Fig. 10). The two Sentinel-1 satellites operate in sun-synchronous orbits with a 12-day repeat cycle. In this work, VV and VH polarized data from the main Interferometric Wide-swath mode of Sentinel-1 are utilized.

Fig. 7. Lugansk International Airport pre-event (left) and post-event (right) images captured by Google Earth.
Source: https://earth.google.com/

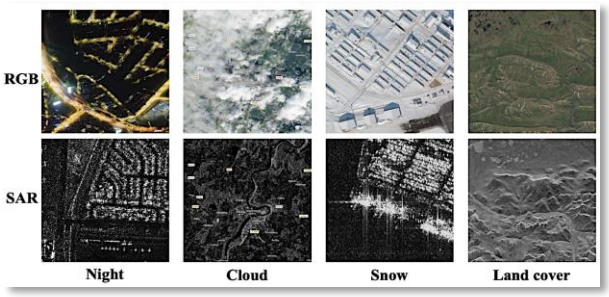

Fig. 8. SAR vs. Optical Imageries exemples.
Source: https://sentinels.copernicus.eu/

While Sentinel-2 is a constellation of two sun-synchronous satellites that enables optical Earth observation at medium resolution. The on-board instrument provides multi-spectral observations in the visible, near-, and short-wave infrared in 13 bands with up to 10 m pixel resolution. The two Sentinel-2 satellites achieve a revisit rate of 5 days at the equator.

4.2. Experiments on the SAR datasets

Extending the view on potential applications of SAR imageries, the proposal encompasses not only disaster datasets but also offers opportunities to use SAR

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

for analyzing war consequences and pre-event scenarios, which is crucial for understanding the scope of the current work.

1. The most important dataset was recently released, and it deserves attention SARDet-100K [21]. SAR has found extensive applications in critical domains that potentially could be applicable to Ukraine case, including national defense [22] and camouflage detection. A significant obstacle in high-resolution SAR image object detection is the sensitivity of SAR images, coupled with the high costs associated with annotating these images. This severely restricts the availability of public datasets. Current datasets often contain only a single type of object set against a basic background. These datasets are also typically small, which may lead to biases when evaluating the proposed methods. Furthermore, a significant obstacle in the progress of SAR object detection research is the unavailability of source codes to the public. This scarcity hampers the ability to replicate past studies accurately, compare methodologies effectively, or enhance prior work. This brand-new dataset comprises approximately images and 246k instances of objects across six distinct categories according to official publications (Fig. 9).

2. Another popular ship detection dataset was SSDD provided by Zhang et.al [23].

3. This section describes freely available datasets specifically conceived for flood mapping applications using SAR. The availability of such datasets is relevant for the validation and testing of flood detection algorithms and is even more stringent for DL detection methods, which require a large amount of data to accomplish the training, testing, and validation phases.

Sen1Floods11 was the Sen1Floods11 dataset developed by Bonafilia et al. [24]. To support the training and validation of DL algorithms for flood detection and mapping based on SAR imagery. It consists of 4831 labeled patches, each sized 512 × 512, encompassing the entire globe and covering 11 flood events at a spatial resolution of 10 m. In addition to capturing flooding events, the dataset also includes permanent water bodies. Most of the patches (namely, 4370) were labeled automatically by means of simple classification algorithms and can be used as weakly supervised training data, whereas the remaining patches were hand-labeled and can be utilized for a refined training, as well as testing and validation purposes.

4. S1S2-Water/Flood. The S1S2-Water dataset is a global dataset designed for the semantic classification of water bodies in S1 and S2 imagery. The model comprises over 100,000 non-overlapping patches, each sized 256 × 256, for each sensor. Each patch is complemented by a corresponding DEM tile derived from the Copernicus DEM, a quality-controlled binary water mask, and other metadata.

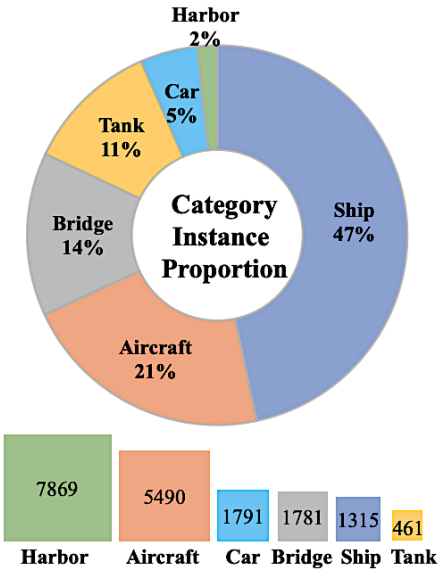

Fig. 9. Percentage of instances in each category and average instance area (in pixels) in SARDet-100K

5. TerraSAR-X. In particular, for damage assessment, Yamazaki et al. [25] collected a dataset from 2011 Tohoku, Japan earthquake, which can be considered as a baseline.

6. The MSTAR dataset is widely used for classification and testing of algorithms. This study classifies, recognize and detect military vehicles with the help of DNNs. (see Fig. 2)

Previously mentioned, ***hyperspectral images and multispectral (HSI, MSI)***, which are comprised of numerous spectral bands, are pivotal for solving diverse issues [26]. The complexity and high dimensionality of hyperspectral data, along with their spectral correlation, pose challenges for machine learning [27]. Various techniques, including dimensionality reduction, data fusion, and classification, have also been proposed.

Deep learning methods, such as fully connected, convolutional (CNN), and recurrent neural networks, have shown success in this domain. Recently, hybrid approaches that combine CNNs with transformers have emerged, using CNNs for spatial feature extraction within a transformer framework. In addition, the pure transformer models specifically designed for hyperspectral imagery are advancing the field [28].

Datasets for HIS and MSI:

7. SpaceNet 8: Focusing on flood-disaster scenarios, SpaceNet 8 is designed for building, road network extraction, and flood detection. This includes imagery covering 850km², encompassing over 32,000 buildings and 1,300km of roads.

Although public access to hyperspectral (HSI) and multispectral (MSI) datasets is limited, employing augmentation and fusion techniques can generate various

scenarios to extend the available data for research and application purposes.

5. Discussions and recommendations

The application of deep learning models, particularly Vision Transformers (ViTs), to the analysis of war-affected areas represents a significant advancement in remote sensing. This study highlighted several critical aspects and challenges that must be addressed to fully leverage these technologies.

Effectiveness of Vision Transformers.

Vision Transformers can capture complex patterns and long-range dependencies in imagery. Unlike traditional CNNs, which often struggle with intricate detail and context, ViTs provide a more comprehensive understanding of scenes by analyzing segmented patches of images. This capability is particularly beneficial in conflict zones where detailed and accurate damage assessment is crucial for both military and humanitarian efforts.

Multimodal Data Integration.

One of the key strengths of transformer-based models is their ability to integrate various imaging modalities. By combining SAR, VHR, and multispectral imagery, these models can provide a more holistic view of the affected areas. This multimodal approach not only enhances the accuracy of damage assessments and offers deeper insights into the nature and extent of destruction. For example, SAR can detect metallic objects under foliage, multispectral imagery can identify material properties, and VHR provides detailed spatial resolution.

Challenges in Data Availability and Quality.

Despite their potential, the implementation of ViTs in war zones faces significant challenges. A major issue is the scarcity of labeled data, which is crucial for training deep learning models. The high cost and limited availability of high-resolution imagery, such as those from Maxar's WorldView-4, further complicate this issue. Publicly available data from satellites like Sentinel-1 and Sentinel-2 offer some respite, but their lower resolution can limit the effectiveness of detailed analyses.

Advances and Future Directions.

Recent advancements in transformer models, such as Conditional Vision Transformers (CPEs) and Detection Transformers (DETR), provide promising avenues to overcome some of these challenges. CPEs enhance the adaptability of transformers to various input sequences, improving their accuracy in image classification tasks. DETR simplifies object detection pipelines, making them more efficient and effective for real-time applications.

Practical Implications.

The practical implications of applying ViTs in remote sensing extend to both strategic military and humanitarian contexts. For military purposes, these models can provide real-time data for tactical decision-making, such as identifying troop movements and assessing battlefield damage. For humanitarian efforts, accurate damage assessments can aid in disaster response, resource allocation, and rebuilding efforts.

Economic Considerations.

The economic implications of using advanced deep learning models in remote sensing are also significant. The initial costs of acquiring high-resolution imagery and developing sophisticated models can be high; however, the long-term benefits, in terms of accurate and timely information, can outweigh these costs. Moreover, advancements in self-supervised learning and transfer learning can reduce the dependency on large –labeled datasets, making these technologies more accessible.

6. Conclusion

The transformative capabilities of DNNs, particularly ViTs, represent a significant leap forward in the assessment of war-affected regions using satellite and aerial imagery. This paper has illustrated how these technologies can surpass the limitations of traditional CNNs (ConvNets) by effectively analyzing SARs, VHRs, HSRs, and multispectral images to evaluate the extensive damage caused by armed conflicts, notably evident in Ukraine during the 2022 Russian invasion.

Relying on self-supervised pre-training methods to generate task-agnostic [29] latent representations has demonstrated promising results in land cover classification, damage assessment, and military asset detection, outperforming fully supervised baselines. This shift towards using transformers in remote sensing and UAVs underscores the urgent need to leverage advanced machine learning techniques for military consequence analysis and humanitarian missions.

The paper emphasizes the significance of integrating various imaging modalities — SAR, VHR, multispectral, and hyperspectral — to provide a holistic understanding of the affected areas. This study also addresses the challenges posed by the scarcity of labeled data and the high costs associated with acquiring hyperspectral and VHR images, suggesting the potential utilization of publicly available data from Sentinel-1 and Sentinel-2 satellites for remote sensing tasks.

Furthermore, the exploration of conditional ViTs and their application in improving the accuracy of image classification tasks in remote sensing illustrates ongoing advances in this field. The study of Detection Transformers (DETR) for object detection and Swin Transformers for dealing with scale variations presents

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

new pathways for enhancing situational awareness and tactical decision-making in conflict zones.

In conclusion, the paper advocates for the increased accessibility of war imagery for scientific, intelligence, and humanitarian purposes, emphasizing the role of transformers in advancing remote sensing technologies. As the field progresses, continued exploration of integrating DNN architectures, such as ViTs and CNNs, is crucial to address the multifaceted impacts of geopolitical events on the environment and human settlements.

This comprehensive assessment not only illuminates the current state of transformer-based models in remote sensing for war-affected areas and sets the stage for future research directions, aiming to enhance the effectiveness and accessibility of critical information derived from satellite imagery during crises.

Contributions of authors: General structure of the work, comparison and drafting of manuscript – **Yurii Pushkarenko**; research idea and led with overall supervision, revision – **Yurii Pushkarenko**, **Volodymyr Zaslavskyi**.

Conflict of Interests

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

The research was conducted without financial support.

Data Availability

The manuscript contains no associated data.

Use of Artificial Intelligence

The authors confirm that they did not use artificial intelligence methods in their work. All the authors have read and agreed to the publication of the finale version of this manuscript.

References

1. Bennett, M. M., Van Den Hoek, J., Zhao, B., & Prishchepov, A. V. Improving Satellite Monitoring of Armed Conflicts. *Earth's Feature*, 2022, vol. 10, iss. 9, article no. e2022EF002904. DOI: 10.1029/2022EF002904.

2. Didmanidze, I. Sh., Megrelishvili, Z. N., & Zaslavskyi, V. A. Safety of Water Resources of a River Basin. *CRC Press*, 2023, vol. 348, iss. 358, article no. e9781003260196-15. DOI: 9781003260196-15.

3. Gaivoronski, A. A., Knopov, P. S., & Zaslavskyi, V. A. Modern Optimization Methods for Decision Making Under Risk and Uncertainty. *CRC Press*, 2023, article no. e9781003260196. DOI: 9781003260196.

4. Wang, X., Feng, G., He, L., An, Q., Xiong, Z., Lu, H., Wang, W., Li, N., Zhao, Y., Wang, Y., & Wang, Y. Evaluating Urban Building Damage of 2023 Kahramanmaras, Turkey Earthquake Sequence Using SAR Change Detection. *School of Geosciences and Info-Physics*, 2023, vol. 23, iss. 14, article no. e23146342. DOI: 10.3390/s23146342.

5. Aimaiti, Y., Sanon, C., Koch, M., Baise, L. G., & Moaveni, B. War Related Building Damage Assessment in Kyiv, Ukraine, Using Sentinel-1 Radar and Sentinel-2 Optical Images. *Remote Sens*., 2022, vol. 14, iss. 24, article no. e6239. DOI: 10.3390/rs14246239.

6. Ameta, A., Singh, V., & Devi, V. S. V. A Convolutional Neural Network Based Approach for SAR Image Classification of Vehicles. *International Journal of Engineering Research & Technology*, 2020, vol. 9, iss. 6, article no. e250. DOI: 10.17577/IJERTV9IS060250.

7. Huang, Q., Jin, G., & Xiong, X. Monitoring Urban Change in Conflict from the Perspective of Optical and SAR Satellites: The Case of Mariupol, a City in the Conflict between RUS and UKR. *Remote Sens*., 2023, vol. 15, iss. 12, article no. e3096. DOI: 10.3390/rs15123096.

8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Computer Vision and Pattern Recognition (cs.CV)*, 2021, vol. 10, iss. 11929, article no. e202011929. DOI: 10.48550/arXiv.2010.11929.

9. Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., & Khan, F. S. Transformers in Remote Sensing: A Survey. *Computer Vision and Pattern Recognition (cs.CV)*, 2022, vol. 9, iss. 01206, article no. e20220901206. DOI: 10.48550/arXiv.2209.01206.

10. Norkin, V. I., Gaivoronski, A. A., Zaslavskyi, V. A., & Knopov, P. S. Optimal Resource Allocation for Active Protection of Critical Infrastructure. *Cybernetics and Systems Analysis*, 2018, vol. 54, iss. 5, article no. e0071-7. DOI: 10.1007/s10559-018-0071-7.

11. Yailymova, H., Yang, H., & Zaslavskyi, V. Models and methods in creative computing: diversity and type-variety principle in development of innovation solutions. *Third International Symposium of Creative Computing (ISPAN-FCST-ISCC)*, 2017, vol. 454, iss. 461, article no. e2017081. DOI: 10.1109/ISPAN-FCST-ISCC.2017.81.

12. Kharchenko, V., Yakovlyev, S., Horbachyk, O., Letychevs'kyy, O., Lukin, V., Sydorenko, M., Siora, O., & Uryvs'kyy, L. *Metody ta zasoby intelektual'noyi obrobky velykykh danykh v bezpechnykh systemakh dystantsiynoho zonduvannya, mul'tymedia ta komunikatsii* [Methods and means of intelligent processing of big data in secure remote sensing, multimedia and communication systems]. *Constanta, Kharkiv, Ukraine*, 2019, vol. 68, iss. 99, article no. e2019415-6. ISBN 978-966342-415-6. (In Ukrainian).

13. Chu, X., Tian, Z., Zhang, B., Wang, X., & Shen, C. Conditional Positional Encodings for Vision Transformers. *Computer Vision and Pattern Recognition (cs.CV)*, 2023, vol. 21, iss. 2, article no. e210210882. DOI: 10.48550/arXiv:2102.10882.

14. Carion, N., Massa, F., Synnaeve, G., & Usunier, N. End-to-End Object Detection with Transformers. *Vision and Pattern Recognition (cs.CV)*, 2020, vol. 20, iss. 5, article no. e200512872. DOI: 10.48550/arXiv:2005.12872.

15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Vision and Pattern Recognition (cs.CV)*, 2021, vol. 21, iss. 3, article no. e210314030. DOI: 10.48550/arXiv.2103.14030.

16. Abdi, G., Esfandiari, M., & Jabari, S. A Deep Transfer Learning-Based Damage Assessment on Post-Event Very High-Resolution Orthophotos. *Geomatica*, 2022, vol. 21, iss. 1, article no. e20210014, pp. 237-250. DOI: 10.1139/geomat-2021-0014.

17. Zhang, Y., Yang, G., Gao, A., Lv, W., Xie, R., Huang, M., & Liu, S. An Efficient Change Detection Method for Disaster-Affected Buildings Based on a Lightweight Residual Block in High-Resolution Remote Sensing Images. *International Journal of Remote Sensing*, 2023, vol. 44, iss. 9, pp. 2959-2981. DOI: 10.1080/01431161.2023.2214274.

18. Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., Heim, E., Choset, H., & Gaston, M. Xbd: A Dataset for Assessing Building Damage from Satellite Imagery. *Computer Vision and Pattern Recognition (cs.CV)*, 2019, vol. 19, iss. 11, article no. e191109296. DOI: 10.48550/arXiv.1911.09296.

19. Ke, X., Zhang, X., & Zhang, T. Gcbanet: A global context boundary-aware network for SAR ship instance segmentation. *Remote Sens.*, 2022, vol. 14, iss. 9, article no. e2165. DOI: 10.3390/rs14092165.

20. Xia, R., Chen, J., Huang, Z., Wan, H., Wu, B., Sun, L., Yao, B., Xiang, H., & Xing, M. Crtranssar: A visual transformer based on contextual joint representation learning for SAR ship detection. *Remote Sens.*, 2022, vol. 14, iss. 6, article no. e1488. DOI: 10.3390/rs14061488.

21. Li, Y., Li, X., Li, W., Hou, Q., Liu, L., Cheng, M.-M., & Yang, J. SARDet-100K: Towards Open-Source Benchmark and ToolKit for Large-Scale SAR Object Detection. *Vision and Pattern Recognition (cs.CV)*, 2020, vol. 24, iss. 3, article no. e240306534. DOI: 10.48550/arXiv:2403.06534.

22. Peng, B., Peng, B., Zhou, J., Xie, J., & Liu, L. Scattering Model Guided Adversarial Examples for SAR Target Recognition: Attack and Defense. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, vol. 60, iss. 3, article no. e3213305. DOI: 10.1109/TGRS.2022.3213305.

23. Zhang, T., Zhang, X., Li, J., Xu, X., & Wang, B. SAR Ship Detection Dataset (SSDD): Official Release and Comprehensive Data Analysis. *Remote Sens.*, 2021, vol. 13, iss. 18, article no. e3690. DOI: 10.3390/rs13183690.

24. Bonafilia, D., Tellman, B., Anderson, T., & Issenberg, E. Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 2020, vol. 50498, iss. 1, article no. e113, pp. 835–845. DOI: 10.1109/CVPRW50498.2020.00113.

25. Yamazaki, F., Iwasaki, Y., Liu, W., Nonaka, T., & Sasagawa, T. Detection of damage to building side-walls in the 2011 Tohoku, Japan earthquake using high-resolution TerraSAR-X images. *Proceedings Volume 8892, Image and Signal Processing for Remote Sensing XIX*, 2013, vol. 8892, iss. 1, article no. e2029465. DOI: 10.1117/12.2029465.

26. Deng, S., Deng, L.-J., Wu, X., & Ran, R. Bidirectional Dilation Transformer for Multispectral and Hyperspectral Image Fusion. Proceedings of the Thirty-Second *International Joint Conference on Artificial Intelligence Main Track*, 2023, vol. 32, iss. 1, article no. e404, pp. 3633-3641. DOI: 10.24963/ijcai.2023/404.

27. Scheibenreif, L., Mommert, M., & Borth, D. Masked Vision Transformers for Hyperspectral Image Classification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, vol. 59228, iss. 1, article no. e210. DOI: 10.1109/CVPRW59228.2023.00210.

28. Cong, Y., Khanna, S., Meng, C., & Liu, P. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. *Computer Vision and Pattern Recognition (cs.CV)*, 2023, vol. 22, iss. 7, article no. e08051. DOI:10.48550/arXiv.2207.08051.

29. Hung, C.-C., Lange, L., & Strötgen, J. TADA: Efficient Task-Agnostic Domain Adaptation for Transformers. *ACL (Findings)*, 2023, vol. 23, iss. 5, article no. e12717, pp. 487-503. DOI:10.48550/arXiv.2305.12717.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Received 17.03.2024, Accepted 15.04.2024

ДОСЛІДЖЕННЯ СТАНУ РАЙОНІВ УКРАЇНИ, ПОСТРАЖДАЛИХ ВІД ВІЙСЬКОВИХ ДІЙ, НА ОСНОВІ ДАНИХ ДИСТАНЦІЙНОГО ЗОНДУВАННЯ ТА АРХІТЕКТУР ГЛИБОКОГО НАВЧАННЯ

Ю. В. Пушкаренко, В. А. Заславський

Вторгнення РФ в Україну та ескалація військових дій в регіонах привело до значного пошкодження житлових будинків, доріг та мостів, різноманітних об'єктів критичної інфраструктури, гребель, та значних забруднень територій. У зв'язку з цим, для дослідження стану регіонів та окремих територій постають задачі дистанційного зондування за допомогою космічних знімків, аероспостереження безпілотними літальними апаратами (БПЛА) з метою аналізу впливу та економічної оцінки ушкоджень та руйнувань на територіях. Для дослідження стану територій за допомогою дистанційного зондування використовується різноманітні (різнотипні) окремі типи моделей глибокого навчання або їх комплекси, архітектури. В роботі досліджуються та використовуються моделі глибоких нейронних мереж (DNN) в задачах комп'ютерного зору (класифікація, сегментація) та поєднання їх похідних таких як згорткові мережі та моделі-трансформери з метою підвищення точності оцінки ушкоджень та руйнувань територій. Ці моделі продемонстрували значний успіх, часто перевершуючи традиційні згорткові нейронні мережі (CNN), і мають потенціал застосування в дистанційному зондуванні для оцінки пошкоджень і захисту критичної інфраструктури. Проведені в роботі дослідження цих моделей підтверджують важливість застосування таких технологій у середовищах, де розмічені дані рідкісні або відсутні. Зокрема, оцінка використання глибоких нейронних мереж, включаючи згорткові мережі та трансформери, під час аналізу регіонів, які постраждали від військових дій, за допомогою радарів із синтезованою апертурою (SAR) і мультиспектральних зображень. Метою та предметом дослідження є також огляд можливостей поєднання згорткових мереж та трансформерів для підвищення швидкості отримання ознак, ландшафтів, забудов, та виявлення складних структурних контурів об'єктів, де зазвичай не вистачає даних. Результат цього дослідження забезпечує критичний огляд застосування згорткових мереж та трансформерів у дистанційному зондуванні, визначаючи значні прогалини та виклики в дослідженнях, особливо в контексті економічних наслідків руйнувань через військові дії. Розглядаються технічні аспекти використання як згорткових мереж, так і моделей на основі трансформерів для складних завдань комп'ютерного зору та передавального навчання в умовах дефіциту даних, а також виклики при аналізі великих об'ємів геофізичних даних. Висновки підкреслюють трансформаційний потенціал DNN, особливо трансформерів, у дистанційному зондуванні в умовах конфліктів і зон лиха. Їхня адаптованість і точність у різних середовищах підкреслюють їх користь як у стратегічному військовому, так і гуманітарному контекстах, встановлюючи практичний стандарт для їх застосування в ключових, реальних сценаріях дослідження стану територій.

Ключові слова: зорові трансформери; комп'ютерний зір; злиття даних (різнотипність); дистанційне зондування; оцінка пошкоджень; згорткові нейронні мережі.

Пушкаренко Юрій Валерійович – асп. каф. математичної інформатики, Київський національний університет імені Тараса Шевченка, Київ, Україна.

Заславський Володимир Анатолійович – д-р техн. наук, проф., проф. каф. математичної інформатики, Київський національний університет імені Тараса Шевченка, Київ, Україна.

Yurii Pushkarenko – PhD Student at the Department of Mathematical Informatics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine,
e-mail: yurii.pushkarenko@gmail.com, ORCID: 0009-0007-2560-2971.

Volodymyr Zaslavskyi – Doctor of Technical Sciences, Professor, Professor at the Department of Mathematical Informatics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine,
e-mail: zas@unicyb.kiev.ua, ORCID: 0000-0001-6225-1313.