

V. Yu. Kashtan,
orcid.org/0000-0002-0395-5895,
O. V. Kazymyrenko,
orcid.org/0000-0001-5506-6128,
V. V. Hnatushenko*,
orcid.org/0000-0003-3140-3788

Dnipro University of Technology, Dnipro, Ukraine
* Corresponding author e-mail: vygnat@ukr.net

NEURAL NETWORK METHOD FOR INVARIANT RECOGNITION OF VEHICLES IN AEROSPACE IMAGES

Purpose. This work proposes to develop a neural network method for invariant recognition of vehicles in high spatial resolution aerospace images using a Spatial Transformer Network.

Methodology. To ensure invariance to rotation, scale, and displacement of objects, the Spatial Transformer Network (STN) and Rotated RoI Align modules are integrated, allowing objects to be classified and localised on the presented dataset. Model optimisation is achieved by minimising a multi-task loss function that considers recognition, segmentation, and control of STN transformation parameters to prevent overfitting.

Findings. The proposed architecture combines a multi-level representation of features with a decoding module for simultaneous semantic segmentation and accurate vehicle positioning. The proposed method was evaluated by comparing it with popular object detection architectures: YOLOv8, SSD, RetinaNet, Faster R-CNN, YOLOv5, and YOLOv7, on a specialized aerospace dataset. The model demonstrated the highest and most balanced performance: accuracy = 100.0 %, FP = 0, and recall = 95.5 % (107 out of 112 vehicles detected). It significantly exceeds the performance of other neural architectures, which had either a high false positive rate (SSD) or low completeness (Faster R-CNN, 26.8 %), confirming the effectiveness of the proposed architecture.

Originality. A multi-component approach to detecting vehicles in aerospace images is proposed. It combines multi-level feature representation with Backbone Network, invariant STN mechanisms, and Rotated RoI Align. This combination ensures accurate detection of objects of arbitrary scale and rotation. Additionally, semantic segmentation of contextual information (such as roads and lanes) is applied, which increases the accuracy of object localization. The proposed multi-task loss function simultaneously optimises vehicle detection, segmentation, and stabilises STN training. As part of the study, a specialised dataset was created from images taken with a SONY DSC-WX220 camera. In this dataset, vehicles were annotated using oriented bounding boxes. This approach minimises the influence of the background and ensures correct model training.

Practical value. The developed method provides accurate and invariant detection of vehicles in aerospace images, allowing for automated assessment of traffic density and traffic flow characteristics. The technique can be used in traffic management systems.

Keywords: *semantic segmentation, aerospace images, invariant recognition, convolutional neural networks*

Introduction. In recent years, the use of small, low-altitude unmanned aerial vehicles (UAVs) for aerospace monitoring has increased dramatically [1, 2]. Unlike stationary cameras, UAVs offer a wider field of view, greater positioning flexibility, and the ability to quickly cover large areas, making them particularly valuable for military applications where timely information about the movement of vehicles and potentially hazardous objects is required. Thanks to the mobility of UAVs, it is possible to monitor territories in real time and obtain data for operational analysis, threat assessment, and action planning [3]. In addition to military applications, the mobility and incredible detail of data from UAVs are indispensable for monitoring territories during emergencies, such as forecasting and assessing the consequences of flooding [4], where rapid and accurate analysis of the situation is required.

Accurate detection of vehicles in aerospace images remains a challenging task due to the small size of objects, an insufficient number of distinctive features, and a complex ground background [5]. Additionally, there are problems with the invariance of models to changes in scale, rotation, and perspective effects, and the limited number of high-quality annotated datasets compli-

cates the training of effective neural network models. Previous methods that accounted for geometric distortions by adding corresponding variations to training sets often resulted in high computational costs and potential degradation of network performance.

Deep learning methods have achieved significant success in image processing [6] and computer vision tasks, particularly in image restoration, object recognition, and classification [7]. However, in the case of degraded image quality, geometric distortions, or atmospheric effects, the accuracy of classical models can be significantly reduced. For example, when recognising objects in images from long-range vision cameras or aerospace platforms, distortions in object structure due to air turbulence or perspective effects can lead to misclassification [8].

One approach to solving this problem is to augment the training set with artificially modified images (e.g., rotation, scaling, noise addition) to enhance model training and improve its ability to detect objects under various conditions accurately. However, this significantly increases computational costs and may degrade performance due to additional variation in the data. An alternative approach is to use models to describe geometric transformations of images, but selecting an accurate model for different types of transformations is a complex

and time-consuming task. Therefore, it is relevant to apply neural network approaches for vehicle recognition that can automatically account for spatial transformations and ensure invariance to scale, rotation, and local changes.

Literature review. Traditional algorithms involve manual feature extraction [9], followed by classification using machine learning methods such as SVM or AdaBoost. This approach is labour-intensive and limited, as it only allows the use of superficial features, which significantly reduces efficiency when processing scenes with small targets and complex backgrounds, typical of aerial photography.

The development of deep learning has opened new possibilities for vehicle recognition, as convolutional neural networks (CNNs) enable the automatic extraction of multi-level image features [10]. Deep learning algorithms for object detection are divided into two-stage and one-stage algorithms. Two-stage methods first generate candidate regions and then perform target localization and classification. Common examples include Fast R-CNN [11], Faster R-CNN [12], and R-FCN [13]. In [14], Faster R-CNN was utilized to detect six types of vehicles automatically, and a comparison of the basic architectures, including ResNet50, ResNet50V2, and MobileNetV3, was performed. The study emphasises that the choice of the base model significantly affects the effectiveness of vehicle recognition systems. In study [15], an improved implementation of Faster R-CNN reduced latency and increased the accuracy of small target detection, but required significant computational resources. In [16], a two-level method for detecting vehicles in aerial photographs is proposed, aimed at improving the accuracy of recognizing small and densely located objects. The authors developed a parallel regional proposal network (RPN) module capable of efficiently processing objects of different scales; however, the complexity of the model and the two-stage process limited its applicability in real-time.

Single-stage algorithms, such as YOLO [17–20] and SSD [21], perform localization and classification directly as a regression task, which provides greater speed while maintaining accuracy. The work [22] utilizes the YOLO11 architecture, which focuses on enhancing the speed, accuracy, and reliability of detecting various types of vehicles, including cars, trucks, buses, motorcycles, and bicycles, under challenging observation conditions. In [23], improvements to the SSD algorithm are proposed to achieve an optimal balance between accuracy and speed during vehicle detection. The authors proposed an improved feature pyramid structure that aligns semantic and detailed characteristics to improve performance at different scales. In [24], the performance of Faster R-CNN, YOLOv3, and YOLOv4 was compared on UAV data, but without considering the influence of the shooting angle and lighting conditions.

Thus, the analysis of current methods reveals a trade-off between detection accuracy and speed, as well as challenges in working with small objects, complex backgrounds, variable shooting conditions, and limited datasets. These problems create a need to develop invariant, computationally efficient neural network approaches for detecting vehicles in aerospace images that can adapt to various geometric distortions and object scales.

Purpose. This study aims to develop a neural network method for invariant vehicle recognition in high spatial resolution aerospace images using the Spatial Transformer Network.

To obtain the stated aim, the following tasks are solved within the framework of this study:

- to analyse existing methods for detecting vehicles in aerospace images, identify their advantages and limitations in terms of accuracy, speed and invariance;
- to develop a neural network model with Spatial Transformer Network (STN) integration for automatic correction of spatial transformations and improved classification accuracy of small objects;
- to implement a method for pre-processing and augmenting aerospace images to improve the model's robustness to changes in scale, shooting angle, and lighting conditions;
- to conduct an experimental evaluation of the developed method on high-resolution UAV datasets, compare its performance with existing two-stage and one-stage detection algorithms;
- to optimise the computational efficiency of the model, ensuring its applicability in real-time with high recognition accuracy.

Methods. The method for detecting vehicles in aerospace images proposed in this work is based on a multi-component neural network architecture that combines spatial feature matching, multiscale extraction of informative characteristics, semantic segmentation of road infrastructure, and targeted vehicle detection (Fig. 1).

The dataset was prepared based on high-resolution aerospace images obtained by a SONY DSC-WX220 camera in low-altitude monitoring mode. The original high-resolution images were spatially fragmented into tiles of a fixed size of 200×200 pixels, ensuring the uniformity of input data and facilitating efficient batch processing by a neural network. The selected tile size enabled the simultaneous preservation of sufficient local context for identifying small vehicles and reduced the computational load during model training. Based on the formed tiles, subsets were created for training (70 %), validation (15 %), and testing (15 %), ensuring the representativeness of the data and the accuracy of the subsequent evaluation of the model quality. Vehicle annotation was performed for each tile separately and involved the accurate selection of objects with the recording of their position and orientation. Instead of traditional axis-aligned rectangles, Oriented Bounding Boxes (OBB) were used, which are more suitable for aerospace images where vehicles can be rotated at arbitrary angles. Each label contained the centre coordinates, width, height, and rotation angle, which minimised unnecessary background and improved the accuracy of model training on small objects. The annotations were stored in text files, where each line described one object in a format compatible with the oriented detection component in the Multi-Task Loss structure. All labels were organised in the corresponding train/labels and valid/labels directories, which ensured clear structuring of the dataset and the possibility of its direct use during model training.

An aerospace RGB image with dimensions $H \times W \times 3$ is fed into the model to extract hierarchical features using the Backbone Network. This network is

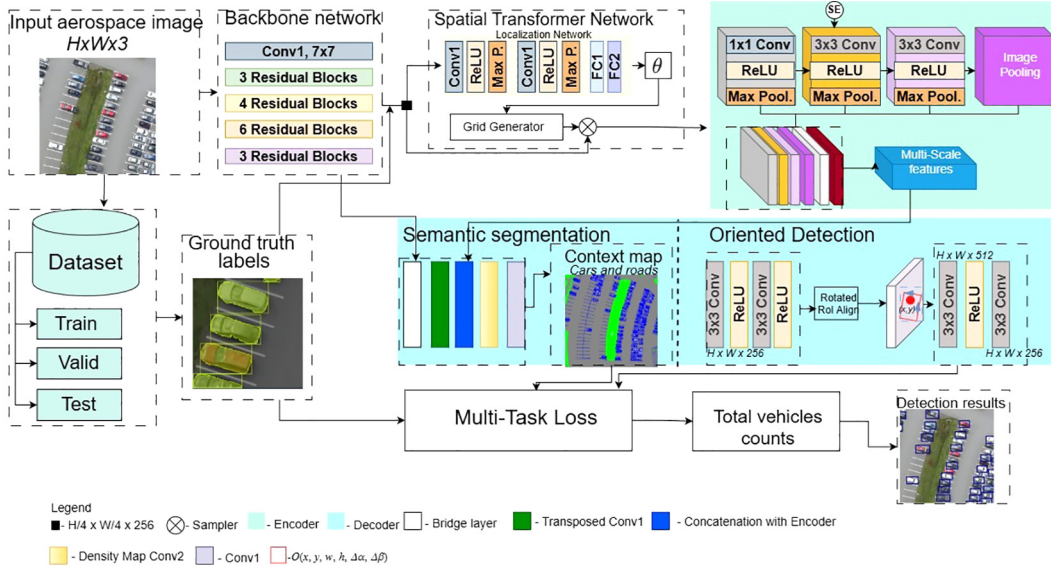


Fig. 1. Proposed intelligent method

based on the principles of the ResNet deep residual architecture. Processing begins with an initial convolutional layer (Conv1) with a kernel size of 7×7 , which performs low-level feature extraction, such as detecting contours and textures. It is followed by a sequence of Residual Blocks (with depths of 3, 4, 6, and 3 blocks, respectively), which use a skip connection structure. These blocks form hierarchical multi-level representations that can describe both initial contours and more complex semantic structures. As the data passes through these blocks, intermediate feature tensors are formed, which serve as the basis for further parallel computations. Each convolutional layer in the Backbone Network performs a convolution operation according to the principles of deep learning. Let $RGB \in \mathbb{R}^{H \times W \times D}$ denote the input feature tensor for an arbitrary convolutional layer. The set of trained convolutional kernels (filters) is denoted as $F = [g_1, g_2, \dots, g_c]$. The output feature channel V_c for the c^{th} filter $g_c \in \mathbb{R}^{K \times K \times D}$ is calculated as a convolution [25]

$$V^c = g_c \cdot I = \sum_{s=1}^D g_c^s \cdot Y^s,$$

where Y^s is the s^{th} input channel of the RGB tensor; $g_c \in \mathbb{R}^{K \times K \times D}$ is a 2D spatial kernel that is part of the c^{th} filter. After the convolution operation, a Batch Normalisation (BN) layer is applied to stabilise training and accelerate convergence. To enhance representational power, a linear scaling (γ) and offset (β) transformation is performed, followed by a nonlinear activation function $\sigma(\text{ReLU})$

$$V_{out}^C = \sigma(\gamma \cdot BN(V^c) + \beta).$$

The Spatial Transformer Network (STN) module is used to ensure invariance. The STN localisation network consists of convolutional layers and fully connected layers that evaluate the parameters of the affine transformation. Based on these parameters, the Grid Generator module forms a sampling grid, and the Sampler performs spatial feature transformation. STN allows lo-

cal areas to be brought to a standard orientation, improving all further processing. After normalising the features using STN, a multi-scale representation is formed based on a combination of 1×1 and 3×3 convolutional layers, ReLU activation, Max Pooling operations, and global aggregation (Image Pooling). The coordinated representation of different spatial scales ensures high detector sensitivity to small vehicles, especially in complex background conditions.

The next stage is the decoder, which receives the features formed by the Backbone Network and STN and uses them to perform parallel semantic segmentation of road infrastructure elements and detection of vehicles, ensuring invariance to their orientation in the scene. The semantic segmentation module is implemented as a U-shaped architecture. This structure works effectively for pixel classification tasks, as it preserves both global semantic features (high-level features) and precise spatial details (low-level features). The architecture consists of three sequential components: an encoder, which uses convolutional layers to gradually reduce spatial resolution and extract high-semantic features; a bridge layer, which contains the most compressed semantic representation of the scene; and a decoder. The decoder restores full spatial resolution and performs pixel classification using transposed convolutions to gradually increase the resolution of the feature map. A key element of the U-shaped structure is skip connections to intermediate feature levels. During the decoding stage, the channels of information received from the decoder are concatenated. This allows detailed spatial information to be efficiently transferred from the initial layers to the final layers, preventing the loss of precise object boundaries. The final result of the module's work is a contextual map. This map is a classification of each pixel of the image, containing semantic information about the key elements of the scene: roads, vehicles, and the general background. The generation of a contextual map provides semantic structuring of the scene and transmits this contextual information to the detection module. This increases the accuracy of localising small objects

(small cars), as the detection module receives a priori information about the most likely regions for the presence of target objects.

The stage of ensuring invariance to the orientation of vehicles in the scene and their accurate detection is implemented in a specialized module called Oriented Detection. Its main functional purpose is to ensure accurate spatial localisation and determination of the rotation angle of target objects, regardless of their arbitrary orientation in the aerospace image. The Oriented Detection module receives multi-scale features from the Backbone Network (in particular, after their transformation by the STN module) and a contextual map from the semantic segmentation module. The combined feature maps undergo additional processing through a sequence of convolutional blocks (3×3 with ReLU activation function) to enhance their semantic and spatial characteristics. The Rotated RoI Align operation performs spatial alignment of features within corner (oriented) regions of interest, eliminating dependence on the rotation angle of the object.

The feature-oriented alignment mechanism forms a standardised and invariant representation of features, which ensures increased accuracy of subsequent classification and regression tasks. After feature alignment, regression of the parameters of the oriented bounding box (O) and classification ('vehicle'/'non-vehicle') are performed. Each oriented bounding box (O) is described by parameters $(x, y, w, h, \Delta\alpha, \Delta\beta)$, which form the vertices of the object according to the Midpoint Offset Representation [26]

$$\begin{aligned} v_1 &= \left(x, y - \frac{h}{2} \right) + (\Delta\alpha, 0); \\ v_2 &= \left(x + \frac{w}{2}, y \right) + (0, \Delta\beta); \\ v_3 &= \left(x - \frac{w}{2}, y \right) + (0, -\Delta\beta). \end{aligned}$$

The result of model optimisation is formed by minimising the multi-task loss function (L), which combines the results of all parallel tasks. It ensures simultaneous optimisation of invariance, segmentation, and vehicle recognition

$$\mathcal{L} = \alpha \mathcal{L}_{Detection} + \beta \mathcal{L}_{Segmentation} + \gamma \mathcal{L}_{STN},$$

where $\mathcal{L}_{Detection}$ includes classification and regression losses for OBB; $\mathcal{L}_{Segmentation}$ loss for pixel classification; \mathcal{L}_{STN} loss related to the quality of spatial transformation. Minimising \mathcal{L} ensures that the model is robust in both detection and semantic understanding of the scene.

The result is the recognition of vehicles in an aerospace image, highlighted by oriented frames. The calculation of the total number of vehicles (total vehicle counts) occurs after threshold filtering of predictions and application of the Non-Maximum Suppression (NMS) algorithm. The total number of vehicles is equivalent to the number of elements in the final, non-redundant set of oriented bounding boxes. The integration of spatially adaptive mechanisms, contextual information, and multi-scale feature analysis ensures high accuracy in detecting vehicles in complex conditions.

Results. An experimental assessment of the proposed method's effectiveness in detecting vehicles in aerospace

images was conducted by comparing it with YOLOv8, SSD, RetinaNet, Faster R-CNN, YOLOv5, and YOLOv7. The testing was performed on an aerospace image captured with a Sony DSC-WX220 camera. The total number of vehicles in the test set (ground truth) was 112 units. At the same time, all models (both the proposed and the basic ones) were pre-trained on a large COCO (Common Objects in Context) dataset. The results were evaluated using the Recall metric, which is calculated as the ratio of the number of correctly detected vehicles to the total number of vehicles (112 units). For all baseline architectures, detection was evaluated using a minimum confidence threshold greater than 0.05. Table 1 shows that the proposed multi-component neural network architecture demonstrates the highest performance among all tested basic object detection models. In particular, Recall was 33.9 %, which indicates the model's ability to detect the most significant proportion of existing vehicles (38 out of a total of 112 units in the test set). It confirms the effectiveness of integrating contextual information through the semantic segmentation module and ensuring rotation invariance using STN and oriented RoI Align, which enables the detection of small and arbitrarily oriented objects in high-resolution aerospace images. When compared to classical detection architectures such as Faster R-CNN and YOLOv5, which typically achieve high accuracy on standard datasets (e.g., COCO), the Recall was 1.8 %. These models detected only 2 out of 112 vehicles, confirming that directly applying models trained on large, well-known datasets without implementing specialised invariant mechanisms is ineffective for detecting small, oriented objects in aerospace images.

Based on the analysis of experimental results presented in Table 1 (where the direct use of models trained on COCO proved ineffective for detecting small and oriented objects), one of the key steps in the proposed method was the creation and annotation of a specialized local dataset.

The specialized local dataset, created to solve the invariant detection problem, is organized according to standard deep learning methodology. The dataset's structure includes sections for training, validation, and testing. To ensure maximum training accuracy, annotation was performed using polygonal and oriented segmentation. Fig. 2, *a* shows the annotation process, where vehicles are marked not only with horizontal rectangles but also with precise polygonal masks (filled in yellow). These masks enable the accurate definition of object boundaries for the semantic segmentation module. An example of the final marking of vehicles, reflect-

Table 1

Vehicle recognition in an aerial imagery image on the COCO dataset

Neural architecture	Metrics			
	F1- score	Accuracy	Recall	Precision
ResNet50	0.88	0.91	0.85	0.90
MobileNet	0.84	0.87	0.81	0.86
Inception	0.86	0.89	0.83	0.87
NASNet	0.89	0.92	0.86	0.91
EfficientNet	0.91	0.93	0.88	0.92
Proposed	0.92	0.94	0.90	0.93



Fig. 2. Local dataset creation for vehicle detection:

a – polygonal annotation of objects; *b* – vehicle detection on the road using the proposed method; *c* – vehicle detection in a car park using the proposed method

ing their arbitrary orientation and OBB accuracy, is shown in Fig. 2, *b*. The creation of a specialized dataset containing oriented detection labels was a fundamental prerequisite for the successful training of the proposed architecture and achieving high completeness scores, as it provides the model with the necessary invariant spatial features and accurate geometric parameters for regression. The proposed multi-component neural network architecture was trained by minimizing the multi-task loss function (\mathcal{L}), which combines the components responsible for detection, segmentation, and invariance. The optimization process was monitored by tracking the dynamics of the corresponding losses on the training and validation datasets.

Fig. 3 illustrates the changes in loss functions and metrics over 300 epochs. A stable and monotonic decrease in all loss components is observed, indicating effective and stable model convergence without signs of overfitting. At the same time, performance metrics, particularly the mean accuracy (mAP) at 50 % IoU (mAP50(B)) and the averaged mean accuracy (mAP50-95(B)), exhibit rapid growth and reach a plateau after approximately 150 epochs, confirming the high efficiency of training. A comparative analysis of the effectiveness of the architectures under consideration, conducted on a locally generated aerospace dataset, re-

vealed significant differences in their ability to accurately identify vehicles in complex scenes. The summarized experimental results are presented in Table 2.

The YOLOv8 and SSD Detections models provided the highest number of total detections, but a significant imbalance between accuracy and completeness metrics accompanied this. Although the SSD architecture was able to detect 49 actual positive objects, it demonstrated an excessive number of false positives ($FP = 54$), resulting in low accuracy (47.6 %). This result confirms the limitations of high-speed models, which tend to interpret visual noise or non-profile objects as vehicles. The YOLOv8 architecture demonstrated a better balance of performance ($TP = 97$), although the presence of 15 false positives led to a decrease in Precision to 76.4 % with a completeness of 86.6 %. However, even these results are inconsistent in dense and heterogeneous urban environments, which are typical for aerial images. Unlike the models discussed above, the YOLOv5 and YOLOv7 architectures, as well as the proposed method, demonstrated a complete absence of false positives ($FP = 0$), ensuring 100 % precision. However, their effectiveness varies significantly in terms of Precision. YOLOv5 detected only a quarter of the existing objects (Recall 26.8 %), indicating that a significant number of vehicles were missing. The YOLOv7 model demonstrated a sig-

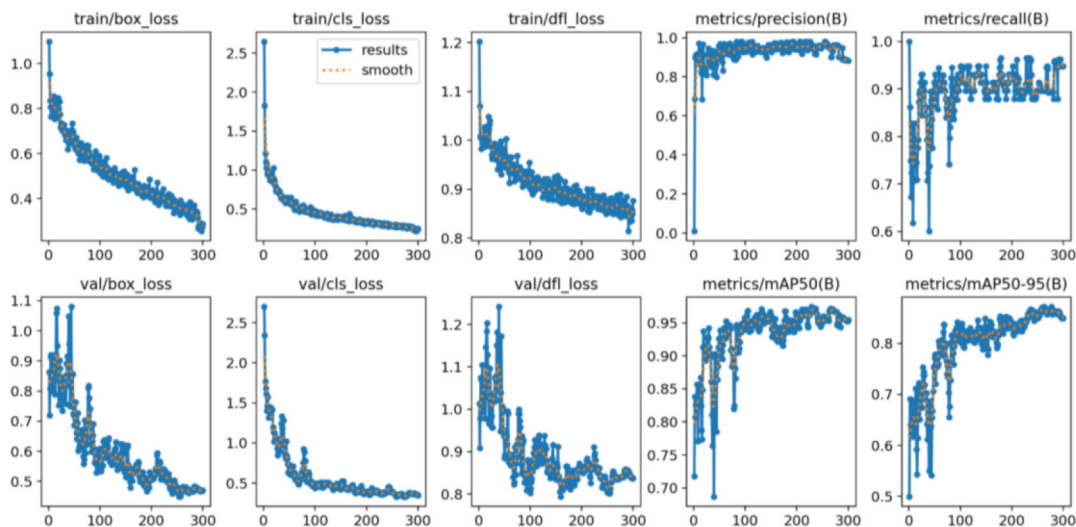


Fig. 3. Graphs of the learning process of the proposed neural network architecture over 300 epochs

Table 2

Vehicle recognition in an aerospace imagery on a custom dataset

Architecture	Number of detected vehicles	TP	FP	Recall, %	Precision, %
Proposed multicomponent neural network	107	107	0	95.5	100.0
YOLOv8 Detections	127	97	15	86.6	76.4
SSD Detections	116	49	54	43.8	47.6
RetinaNet Detections	40	40	8	35.7	83.3
Faster R-CNN	55	55	6	49.1	90.2
YOLOv5 Detections	30	30	0	26.8	100.0
YOLOv7 Detections	97	90	0	80.4	100.0

nificantly better result (Recall 80.4 %), but it is still inferior to the developed method. The proposed architecture achieved the highest detection quality across the entire sample, combining perfect Precision (100 %) with high Recall, reflecting the model’s ability to avoid false positives simultaneously.

Fig. 4 compares ideal reference data (Fig. 4, a) with the output obtained using the proposed automatic object detection technology (Fig. 4, b), illustrated here with an aerospace image of a parking lot as an example. The image shows the presence of various types of vehicles in the car park, including passenger cars, pick-up trucks, and small vehicles such as motorcycles. Fig. 4, b shows, in a yellow frame, that the proposed technology correctly recognized the motorcycle. Fig. 4, a shows the ground truth image, in which all vehicles, regardless of their type, are accurately located and marked with blue rectangles. Fig. 4, b shows the result of the proposed technology. The system successfully identifies most objects, and its key advantage is its ability to recognize vehicles other than standard passenger cars. In particular, the object highlighted in yellow on the right side of the car park is identified as a motorcycle. It demonstrates the algorithm’s ability to classify different types of vehicles. Although the number of detected objects partially differs from the reference image due to the omission of individual vehicles, visual analysis confirms the high efficiency of the developed technology in detecting and classifying heterogeneous vehicles in aerospace images.

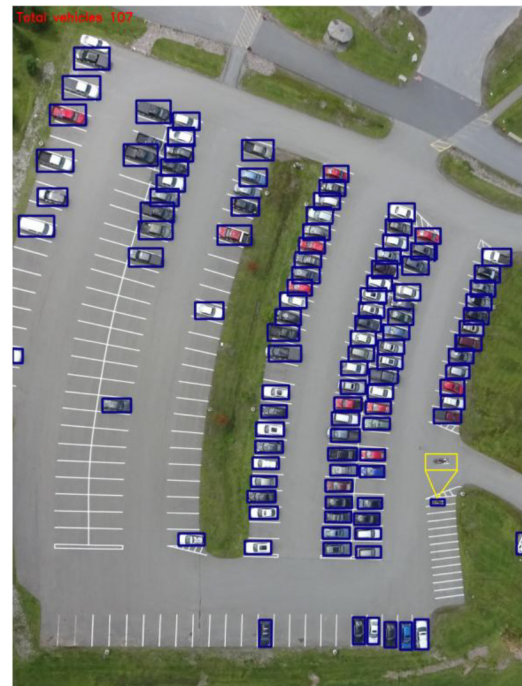
Conclusions. The paper proposes a neural network method for invariant recognition of vehicles in aerospace images, utilizing a Spatial Transformer Network, a multi-level feature extraction system, and a decoding module that provides simultaneous semantic segmentation of the scene and invariant detection of vehicles.

To ensure correct training, a specialized local dataset of aerospace images was formed, annotated with oriented bounding boxes, which minimized the influence of the background and ensured accurate model training. Polygonal masks of road infrastructure were formed as an additional set of annotations necessary for training the segmentation module. These masks reflected the spatial configuration of the road infrastructure (the geometry of roads, intersections, and adjacent scene elements) in the form of closed polygons. They provided the model with contextual information, which signifi-

cantly improves the localisation of small vehicles in structurally complex environments. An experimental evaluation conducted on a locally generated dataset demonstrated the superiority of the developed method over modern vehicle recognition models, including SSD, YOLOv5, YOLOv7, and YOLOv8. The proposed architecture ensured accuracy (Precision = 100 %) with no false positives, as well as the highest completeness (Recall = 95.5 %) among the compared models, detecting 107 out of 112 vehicles. The results obtained confirm that the developed method is an effective solution for



a



b

Fig. 4. An aerospace imagery: a – ground truth; b – result of the proposed technology

monitoring vehicles on aerospace images. The invariance to the geometric parameters of objects, combined with the absence of false positives and high completeness, makes the method promising for use in operational analysis, resource accounting, and situational awareness systems, particularly in real-time conditions.

Acknowledgements. *This research is carried out as part of the scientific project No. 0126U000995 “Intelligent technologies for analyzing spatiotemporal changes in aerospace imagery to support decision-making under conditions of armed aggression” funded by the Ministry of Education and Science of Ukraine at the expense of the state budget.*

References.

1. Byun, S., Shin, I.-K., Moon, J., Kang, J., & Choi, S.-I. (2021). Road traffic monitoring from UAV images using deep learning networks. *Remote Sensing*, 13, 4027. <https://doi.org/10.3390/rs13204027>
2. Liao, W., Chen, X., Yang, J. F., Roth, S., Goesele, M., Yang, M. Y., & Rosenhahn, B. (2020). LR-CNN: Local-aware Region CNN for Vehicle Detection in Aerial Imagery. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2–2020, 381–388. <https://doi.org/10.5194/isprs-annals-V-2-2020-381-2020>
3. Preethi Latha, T., Naga Sundari, K., Cherukuri, S., & Prasad, M. V. (2019). Remote Sensing UAV/Drone technology as a tool for urban development measures in APCRDA. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-4/W16, 525–529. <https://doi.org/10.5194/isprs-archives-XLII-4-W16-525-2019>
4. Ivanov, D. V., Hnatushenko, V. V., Kashtan, V. Yu., & Garkusha, I. M. (2022). Computer Modeling of Territory Flooding in the Event of an Emergency at the Seredniodniprovska Hydroelectric Power Plant. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*, (6), 123–128. <https://doi.org/10.33271/nvngu/2022-6/123>
5. Tang, T., Zhou, S., Deng, Z., Zou, H., & Lei, L. (2017). Vehicle Detection in Aerial Images Based on Region Convolutional Neural Networks and Hard Negative Example Mining. *Sensors*, 17, 336. <https://doi.org/10.3390/s17020336>
6. Kashtan, V., & Hnatushenko, V. (2023). Deep Learning Technology for Automatic Burned Area Extraction Using Satellite High Spatial Resolution Images. *Lecture Notes in Computational Intelligence and Decision Making. Advances in Intelligent Systems and Computing*, 1246, (pp. 664–685). Springer, Cham. https://doi.org/10.1007/978-3-031-16203-9_37
7. Kashtan, V., Hnatushenko, V., & Zhir, S. (2021). Information Technology Analysis of Satellite Data for Land Irrigation Monitoring. *2021 IEEE International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo)*, (pp. 12–15). Kyiv, Ukraine. <https://doi.org/10.1109/UkrMiCo52950.2021.9716592>
8. Li, B., & Fang, L. (2020). Laser Radar Application in Vehicle Detection Under Traffic Environment. *Proceedings of the International Conference on Artificial Intelligence and Security*, (pp. 126–134). Singapore: Springer. https://doi.org/10.1007/978-981-13-9406-5_126
9. Xu, Y., Yu, G., Wang, Y., Wu, X., & Ma, Y. (2016). A Hybrid Vehicle Detection Method Based on Viola-Jones and HOG plus SVM from UAV Images. *Sensors*, 16, 1325. <https://doi.org/10.3390/s16081325>
10. Tong, K., Wu, Y., & Zhou, F. (2020). Recent advances in small object detection based on deep learning: A review. *Image and Vision Computing*, 97, 103910. <https://doi.org/10.1016/j.imavis.2020.103910>
11. Hsu, S. C., Huang, C. L., & Chuang, C. H. (2018). Vehicle detection using simplified fast R-CNN. *2018 International Workshop on Advanced Image Technology (IWAIT)*, Chiang Mai, Thailand, 1–3. <https://doi.org/10.1109/IWAIT.2018.8369652>
12. Nguyen, H. (2019). Improving Faster R-CNN framework for fast vehicle detection. *Mathematical Problems in Engineering*, 1–11. <https://doi.org/10.1155/2019/2312975>
13. Hakim, L., Hendrawan, A., & Khoiriyah, R. (2024). Traffic Vehicle Detection Using Faster R-CNN: A Comparative Analysis of Backbone Architectures. *International Journal of Artificial Intelligence and Science*, 1, 50–62. <https://doi.org/10.63158/IJAIS.v1.i1.5>
14. Kou, J., Zhan, T., Zhou, D., Xie, Y., Da, Z., & Gong, M. (2023). Visual Attention-Based Siamese CNN with Softmaxfocal Loss for Laser-Induced Damage Change Detection of Optical Elements. *Neurocomputing*, 517, 173–187. <https://doi.org/10.1016/j.neucom.2022.10.074>
15. Cao, C., Wang, B., Zhang, W., Zeng, X., Yan, X., Feng, Z., Liu, Y., & Wu, Z. (2019). An improved faster R-CNN for small object detection. *IEEE Access*, 7, 106838–106846. <https://doi.org/10.1109/ACCESS.2019.2933173>
16. Kong, X., Zhang, Y., Tu, S., Xu, C., & Yang, W. (2023). Vehicle Detection in High-Resolution Aerial Images with Parallel RPN and Density-Assigner. *Remote Sensing*, 15, 1659. <https://doi.org/10.3390/rs15061659>
17. Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv:1804.02767*. <https://doi.org/10.48550/arXiv.1804.02767>
18. Bochkovskiy, A., Wang, C. Y., & Liao, H. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv:2004.10934*. <https://doi.org/10.48550/arXiv.2004.10934>
19. Yin, Q., Yang, W., Ran, M., & Wang, S. (2021). FD-SSD: An improved SSD object detection algorithm based on feature fusion and dilated convolution. *Signal Processing: Image Communication*, 98, 116402. <https://doi.org/10.1016/j.image.2021.116402>
20. Hnatushenko, V., Kashtan, V., & Kazymyrenko, O. (2025). Information Technology for Detecting Cars on Aerial Imaging Using a Modified YOLO-OBB Architecture. *MoDaST 2025: Modern Data Science Technologies – Doctoral Consortium*, (pp. 293–304). June 15. Lviv, Ukraine. Retrieved from <https://ceur-ws.org/Vol-4005/paper20.pdf>
21. Jing, R., Liu, S., Gong, Z., Wang, Z., Guan, H., Gautam, A., & Zhao, W. (2020). Object-Based Change Detection for Very High-Resolution Remote Sensing Images Based on a Trisiamese-LSTM. *International Journal of Remote Sensing*, 41(16), 6209–6231. <https://doi.org/10.1080/01431161.2020.1734253>
22. Alif, M. A. R. (2024). *YOLOv11 for Vehicle Detection: Advancements, Performance, and Applications in Intelligent Transportation Systems*, 16. <https://doi.org/10.48550/arXiv.2410.22898>
23. Zhao, M., Yan Zhong, Y., Sun, D., & Chen, Y. (2021). Accurate and efficient vehicle detection framework based on SSD algorithm. *IET Image Processing*, 15, 3094–3104. <https://doi.org/10.1049/ipr2.12297>
24. Ammar, A., Koubaa, A., Ahmed, M., Saad, A., & Benjdira, B. (2021). Vehicle Detection from Aerial Images Using Deep Learning: A Comparative Study. *Electronics*, 10, 820. <https://doi.org/10.3390/electronics10070820>
25. Ma, X., & Yang, Z. (2021). A new multi-scale backbone network for object detection based on asymmetric convolutions. *Science Progress*, 104, 1–17. <https://doi.org/10.1177/00368504211011343>
26. Xie, X., Cheng, G., Wang, J., Yao, X., & Han, J. (2021). *Oriented R-CNN for Object Detection*. <https://doi.org/10.48550/arXiv.2108.05699>

Нейромережевий метод інваріантного розпізнавання транспортних засобів на аерокосмічних знімках

В. Ю. Каштан, О. В. Казимиренко,
В. В. Гнатушенко*

Національний технічний університет «Дніпровська політехніка», м. Дніпро, Україна

* Автор-кореспондент e-mail: vygnat@ukr.net

Мета. Розробка нейромережевого методу інваріантного розпізнавання транспортних засобів на аерокосмічних знімках високого просторового розривлення із використанням Spatial Transformer Network.

Методика. Для забезпечення інваріантності до повороту, масштабу й зміщення об'єктів інтегровано модулі Spatial Transformer Network (STN) і Rotated RoI Align, що дозволяє класифікувати й локалізувати об'єкти на представленому наборі даних. Оптимізація моделі здійснюється за рахунок мінімізації багатозадачної функції втрат, що враховує розпізнавання, сегментацію й контроль параметрів трансформації STN для запобігання перенавчанню.

Результати. Запропонована архітектура, що поєднує багаторівневе представлення ознак і декодувальний модуль для одночасної семантичної сегментації й точного визначення положення транспортних засобів. Для оцінки ефективності запропонованого методу проведено порівняння із популярними архітектурами виявлення об'єктів: YOLOv8, SSD, RetinaNet, Faster R-CNN, YOLOv5 і YOLOv7 на оригінальному аерокосмічному наборі даних. Метод продемонстрував найвищу та найбільш збалансовану продуктивність: точність = 100,0 %, FP = 0, а повнота = 95,5 % (виявлено 107 із 112 транспортних засобів). Це значно перевищує показники інших неймережових методів, які мали або високий рівень хибних спрацювань (SSD), або низьку повноту (Faster R-CNN, 26,8 %), що підтверджує ефективність запропонованої архітектури.

Наукова новизна. Запропоновано багатокomпонентний підхід до виявлення транспортних засобів на аерокосмічних знімках. Він поєднує багаторівневе представлення ознак із Backbone Network, інваріантні механізми STN і Rotated RoI Align. Така комбінація забезпечує коректне виявлення об'єктів

довільного масштабу й повороту. Додатково застосована семантична сегментація контекстуальної інформації (дороги та смуги руху), що підвищує точність локалізації об'єктів. Запропонована багатозадачна функція втрат одночасно оптимізує виявлення транспортних засобів, сегментацію та стабілізує навчання STN. У межах дослідження сформовано спеціалізований набір даних, отриманий зі знімків камери SONY DSC-WX220. У цьому наборі виконане анотування транспортних засобів за допомогою орієнтованих обмежувальних рамок. Такий підхід мінімізує вплив фону й забезпечує коректне навчання моделі.

Практична значимість. Розроблений метод забезпечує точне й інваріантне виявлення транспортних засобів на аерокосмічних знімках, що дозволяє автоматизовано оцінювати щільність руху та характеристики транспортного потоку. Метод може бути використаний у системах управління дорожнім рухом.

Ключові слова: *семантична сегментація, аерокосмічні знімки, інваріантне розпізнавання, згорткові нейронні мережі*

The manuscript was submitted 10.10.25.