Journal of Edge Computing, 2026, Vol. **5**, Iss. 1, pp. 90–104 10.55056/jec.749

Object detection method based on instance segmentation of satellite image obtained in the conditions of cloud cover

Serhiy V. Kovbasiuk[1], Mykola P. Romanchuk[2], Olena M. Naumchak[2] and Leonid M. Naumchak[2]

[1]*Zhytomyr Polytechnic State University, 103 Chudnivska Str., Zhytomyr, 10005, Ukraine*
[2]*Korolyov Zhytomyr Military Institute, 22 Myru Ave., Zhytomyr, 10004, Ukraine*

Abstract. Modern achievements in the space industry, combined with the continuous development of remote sensing technologies, form the basis for solving problems in various areas. Medium- and high-resolution satellite imagery often plays a key role in decision-making during crises in hard-to-reach areas. In the process of processing remote sensing data, a significant and still unresolved problem is the reconstruction of clouded images. This article analyses various approaches to cloud removal and data quality improvement. The traditional approaches considered have certain limitations associated with the loss of useful information. Particular attention is paid to deep learning methods, which are gaining popularity in solving cloud removal problems because they produce good results. The article discusses different DNN architectures (convolutional neural networks (CNN), conditional generative adversarial networks (cGAN)) and their modifications, identifies their advantages and disadvantages. A significant advantage of neural networks is their ability to adapt to various conditions and image types. The analysis of the disadvantages of fusing purely optical data led to the conclusion that the best solution to the problem of cloud removal from satellite images is to combine optical and radar data. As a result, the architecture of a model for removing clouds from optical satellite imagery using generative adversarial networks, combined with radar imagery, was developed. The theoretical hypotheses were confirmed by testing the model on the SEN12MS-CR dataset.[1]

Keywords: recognition, object detection, satellite images, instance segmentation, focal loss, reconstructing images, cloud removal, generative adversarial network

1. Introduction

The evolution of Earth remote sensing technologies using satellites enables the performance of a wide range of tasks in areas such as environmental monitoring, agriculture, cartography and geodesy, natural resource management, natural disaster monitoring, meteorology, ecology, security, and defence. In many cases, the primary challenge in crisis management for any type of facility is the lack of sufficient information to make informed decisions on how to mitigate the consequences. The inability to consider all factors in a crisis reduces the effectiveness of preventive measures and rescue operations. Accordingly, the key indicators during the elimination of a crisis are operational efficiency and objectivity in assessing hard-to-reach areas. Medium- and high-resolution remote sensing data, along with their integration with automatic detection, recognition, and classification methods for ground objects, are gaining increasing importance.

Multispectral remote sensing systems can provide data with a high degree of reliability. However,

[1]This paper is the further development of our work [15] presented at the 3rd Edge Computing Workshop.

0000-0002-6003-7660 (S. V. Kovbasiuk); 0000-0002-0087-8994 (M. P. Romanchuk); 0000-0003-3336-1032 (O. M. Naumchak); 0000-0002-7311-6659 (L. M. Naumchak)
klasik552008@gmail.com (S. V. Kovbasiuk); romannik@ukr.net (M. P. Romanchuk); olenanau@gmail.com (O. M. Naumchak); naumchak.leonid@gmail.com (L. M. Naumchak)
https://ieeexplore.ieee.org/author/37087014573 (S. V. Kovbasiuk); https://ieeexplore.ieee.org/author/37087013658 (M. P. Romanchuk); https://ieeexplore.ieee.org/author/37089181640 (O. M. Naumchak); https://ieeexplore.ieee.org/author/37089179498 (L. M. Naumchak)

Received	Accepted	Published	Version of record
2024-05-23	2025-12-02	2026-02-16	2026-05-21

achieving the necessary level of situational awareness in dynamic scenarios remains challenging, particularly due to the sensitivity of optical sensors to environmental factors and weather conditions. Among these, cloud cover presents a significant obstacle [26, 30].

The rapid development of space technologies, increased accessibility and growing participation of private entities in the space sector allow access to high and medium-resolution satellite data, sometimes even at no cost.

Given that information from various remote sensing platforms is often complementary, integrating data from multiple sources can significantly enhance the reliability and precision of the results. Nevertheless, the integration of multi-temporal data acquired from different sensors remains a largely unresolved challenge. An additional barrier to progress in this area is the limited availability of open-access multimodal datasets.

A further pressing issue is the presence of cloud cover in optical satellite imagery. Obstacles caused by smoke from various incidents and constant cloud cover significantly complicate the use of optical satellite data for rapid response to situations. Conversely, such atmospheric conditions have minimal impact on the performance of synthetic aperture radar (SAR) systems [51]. SAR sensors are capable of collecting data under all-weather conditions and during both day and night [31]. Therefore, due to the complementary nature of SAR and optical data, a critical research objective is the development of effective fusion techniques to produce enhanced and more informative imagery.

2. Problem statement

Modern visualisation systems facilitate the representation of vast volumes of information obtained from diverse spatial platforms, including spacecraft engaged in optical-electronic Earth surface monitoring and remote sensing. However, such data typically lack intermediate analytical interpretations, which complicates the forecasting of event sequences and hinders timely decision-making processes. To address this challenge in an automated manner, the acquired data must undergo thematic processing – specifically, the extraction of meaningful information from satellite imagery. Thematic processing and the integration of data from all the mentioned sources enable comprehensive situational assessments within specific geographical regions.

This approach to data acquisition requires the application of system analysis and synthesis techniques to handle heterogeneous data collected from different sources at varying times and with different parameters. To ensure the efficient transformation of disparate data streams into a unified, actionable result – especially under visually challenging conditions – it is essential to define the core phases of thematic processing, investigate the logical relationships in the structure of integrated data, and identify emerging issues along with potential solutions.

To address newly arising tasks involving satellite imagery, there is a pressing need for the development of an effective (i.e., rapid and sufficiently accurate) method for detecting fine-grained objects in conditions of cloud cover.

In earlier work [15], an object detection method was proposed based on instance segmentation of aerial imagery obtained by unmanned aerial vehicles under challenging visual conditions. This approach may serve as a foundation for developing object detection techniques applicable to satellite imagery acquired in cloudy environments.

The objective of this article is to analyse the application of neural network-based models for object detection in satellite images under cloud-covered conditions, and to explore their further development to improve the precision of object localisation and identification on the Earth's surface.

3. Theoretical background

An analysis of atmospheric transmittance over Ukraine in 2024 [41], summarised in table 1, indicates that the average cloud cover across the country reaches approximately 62%. This significant level of cloudiness presents substantial challenges for the conventional visualisation and interpretation of

satellite imagery from various sources, thereby highlighting the need for dedicated image processing techniques that can operate effectively under such conditions.

Table 1
Analysis of cloud coverage over Ukraine in 2024.

Region of Ukraine	January	February	March	April	May	June	July	August	September	October	November	December
Cherkasy	83	92	73	62	22	40	33	17	43	58	67	95
Chernihiv	87	92	62	53	23	70	38	22	45	53	67	85
Chernivtsi	42	53	88	62	17	47	33	40	63	95	82	92
Dnipro	82	80	58	57	67	50	48	32	45	47	70	93
Donetsk	72	65	72	57	53	53	32	23	40	37	72	93
Ivano-Frankivsk	45	58	63	43	50	32	82	40	60	42	47	88
Kharkiv	90	73	78	58	32	42	13	12	13	45	87	82
Kherson	65	80	60	58	5	38	45	22	40	50	42	87
Khmelnytskyi	80	90	78	62	45	18	55	38	67	57	47	92
Kropyvnytskyi	82	85	63	65	33	58	35	38	62	47	68	95
Kyiv	78	93	67	70	27	63	40	25	43	50	77	98
Luhansk	85	73	75	60	22	48	53	38	15	60	90	97
Lutsk	88	80	70	53	60	72	57	37	62	72	68	95
Lviv	58	77	60	58	43	43	58	43	63	57	47	93
Mykolaiv	67	82	62	53	15	40	47	27	52	47	52	78
Odesa	73	65	72	53	15	40	37	35	48	52	37	82
Poltava	82	82	93	55	38	72	37	37	37	50	82	97
Rivne	92	87	48	72	57	60	55	45	62	67	80	93
Simferopol	65	67	70	33	45	40	75	22	57	52	58	73
Sumy	90	98	82	45	40	85	58	32	13	45	75	98
Ternopil	100	75	75	40	50	42	67	45	75	53	53	95
Uzhhorod	52	82	62	48	37	38	47	60	50	40	55	62
Vinnytsia	85	80	63	60	40	38	40	47	50	68	90	97
Zaporizhzhia	75	67	63	60	10	35	38	18	33	60	78	83
Zhytomyr	75	88	70	80	37	55	53	47	30	73	83	88
Ukraine	76	79	69	57	35	49	47	34	47	55	67	89

One of the persistent challenges in remote sensing is the detrimental effect of cloud cover on Earth observation data. The automated reconstruction of cloud-contaminated or noisy information remains an ongoing problem [28, 32]. A comprehensive review of traditional preprocessing techniques for remote sensing imagery affected by cloud cover [33] categorises the available methods into three primary groups: multispectral, multitemporal, and shading-based approaches. Most other techniques are typically hybrid variants combining elements from these core categories.

Multispectral methods are employed when clouds do not entirely obscure optical signals, allowing partial transmission depending on the wavelength of absorption and reflection. In such cases, information about the Earth's surface can be reconstructed using mathematical [43] or physical models [23]. These methods are advantageous because they rely solely on the original satellite data and do not require additional inputs. However, their effectiveness is limited to scenarios involving semi-transparent cloud cover.

Multitemporal methods utilise reference imagery acquired during cloud-free conditions to reconstruct affected regions. Their primary advantage lies in their simplicity and operational feasibility. Nevertheless, these methods inherently depend on the availability of recent and temporally close cloud-free observations. This requirement poses significant limitations, particularly in rapidly changing environments, where older data may no longer accurately represent current conditions.

Earlier approaches to cloud removal often relied on the assumption that cloudy and non-cloudy

regions shared similar statistical or geometric properties. Corrective methods restore occluded areas using information extracted from cloud-free sections within the same image [24]. While these approaches do not require external datasets, they are only effective in situations with minimal cloud presence. To address this, the process of selecting similar pixels for substitution is frequently supported by auxiliary data, including multitemporal [4] or SAR-based inputs [9].

Though these techniques are generally effective, they are complex and data-intensive, often requiring integration of multi-source or multi-temporal information. Related to them are interpolation methods, which reconstruct missing data by analysing spatial relationships between neighbouring cloud-free pixels. These include approaches based on nearest neighbors [34] and kriging interpolation [46]. However, when clouds cover extensive contiguous areas, proximity-based assumptions become ineffective.

Recent advancements in deep learning have created new opportunities for addressing the limitations of traditional methods. Deep neural networks (DNNs) have shown substantial promise in the task of cloud removal from satellite imagery, offering both flexibility and robustness.

In the context of fine-grained object detection and recognition, methods of semantic and instance segmentation are being actively developed, each with its own strengths and trade-offs.

Semantic segmentation methods based on convolutional neural networks (CNNs) address detection and recognition through multilevel feature aggregation [17] or structural prediction [22]. Enhanced CNNs, such as those utilising pyramid pooling modules (PPM) and feature pyramid networks (FPN) [37, 38], as seen in models like Mask R-CNN [12], enhance contextual understanding by preserving high-resolution features across layers.

Instance segmentation focuses on delineating object boundaries at the pixel level for each semantic class. Beginning with the regional CNN (R-CNN) framework [49], this approach typically involves a two-stage process: first generating candidate regions, then selecting the best match [3, 45]. Standard methods utilise a Region Proposal Network (RPN) to generate initial masks before classification. For example, InstanceFCN utilises a fully convolutional network to generate mask proposals [35, 45], whereas MNC [7] treats instance segmentation as a pipeline of three sequential tasks: mask localisation, prediction, and classification. In Mask R-CNN, this pipeline is extended by adding a parallel mask prediction branch to the Faster R-CNN framework [29], enabling accurate bounding box refinement and mask generation. PANet further enhances performance by introducing bidirectional information flow in the FPN [18, 20].

Semantic segmentation models built on FCNs [22] typically do not define object boundaries, whereas instance segmentation methods based on region proposals [5, 45] tend to ignore background objects. Their combination, however, enables robust scene understanding, image summarization [38], and comprehensive context extraction [37, 45].

To enhance the reliability of fine-grained object detection, two-stage detectors have been developed [11, 12, 29]. Compared to single-stage detectors [21, 27], they provide better optimisation and generate a larger number of high-level features. For instance, Multi-Region CNN [42] employs an iterative refinement mechanism, while AttractioNet [10] utilises an “Attend & Refine” module to iteratively improve region proposals. Detection models, such as CRAFT [44] and Fast R-CNN [21], incorporate cascaded RPN structures to enhance detection confidence.

A promising strategy for improving both detection reliability and recognition accuracy is the use of cascaded network architectures. Cascade R-CNN [1] operates through multiple stages, where each subsequent stage receives input refined by the previous one and applies progressively higher IoU thresholds. While the direct combination of Cascade R-CNN with Mask R-CNN provides only marginal improvements, it enhances detection precision through better localisation of bounding boxes, which leads to more accurate mask predictions.

Therefore, constructing a multi-stage pipeline for satellite image analysis that integrates detection, instance segmentation, and semantic segmentation offers a powerful approach for improving object recognition accuracy and contextual understanding.

4. Results

4.1. Cloud removal using SAR-optical fusion and generative adversarial networks

Cloud cover remains a critical limitation in the use of remote sensing data. Although existing neural network-based models for cloud removal have shown the ability to handle various types of clouds and residual atmospheric interference, the core challenge – reconstructing low-quality (cloud-obscured) input imagery – has driven the development of convolutional neural network based methods [47].

A notable example of such methods is the conditional generative adversarial network (cGAN) [14], which is trained to remove cloud-covered regions from visible (RGB) satellite imagery (e.g., WorldView-2) using auxiliary data from the near-infrared (NIR) spectrum. A similar model, McGAN [16], builds on the pix2pix architecture to correlate visible and NIR spectral data for cloud removal. Another noteworthy approach, CycleGAN [48], differs in that it does not require paired cloudy and cloud-free images for training. Despite their innovation, these GAN-based approaches generally suffer from high computational complexity, unstable training dynamics, and reduced prediction performance, particularly when processing low-quality inputs.

A further limitation of these deep learning-based techniques lies in their reliance on large, high-quality real-world datasets. Most existing models are trained on narrow geographical areas and synthetic cloud simulations, which hinders their generalisation to diverse regions and real-world operational conditions.

Given these limitations, integrated approaches combining synthetic aperture radar and optical data have attracted increasing attention. In [9], for instance, radar and optical imagery are fused using a nearest spectral matching method [25]. However, a fundamental constraint of SAR-optical fusion approaches is that SAR data cannot fully capture all the spectral characteristics of the Earth's surface, limiting their effectiveness in certain contexts [40].

SAR2OPT methods offer two main configurations:

- Global fusion, which leverages spatial relationships across the entire optical image to reconstruct cloudy regions while preserving the overall structural coherence with the surrounding cloud-free areas.
- Local fusion, which incorporates only SAR-derived information corresponding to the cloudy regions. This approach better preserves the texture and local detail of the occluded areas, while avoiding the degradation of quality due to irrelevant or dynamically filtered features.

In the proposed model, optical satellite images obscured by clouds are reconstructed using a continuous-valued cloudiness mask that distinguishes between cloudy and cloud-free pixels. Cloud-free regions remain unchanged, while cloud-covered areas are corrected with auxiliary SAR data. Unlike earlier methods that rely on binary cloud masks or are less sensitive to cloud extent, this approach leverages fine-grained cloud coverage maps. It does not require exact pixel-level alignment between clouded and reference images during training.

To enable efficient learning with a large receptive field and aggressive convolution, the neural network is structured around a cyclic-sequential 7-ResNeXt GAN module [50] (figure 1). The architecture consists of three major components:

- Encoder – extracts hierarchical features from the input image;
- Decoder – reconstructs the target image from encoded features;
- Generator – a series of residual blocks with long-skip connections [13], which enhance gradient flow and allow the model to retain both high- and low-level contextual information during training.

During the training phase, the encoder receives paired inputs: SAR and optical satellite images along with their corresponding cloud coverage masks. The decoder then produces a predicted cloud mask, which is combined with the original cloudy optical image. To correct for the non-linearity

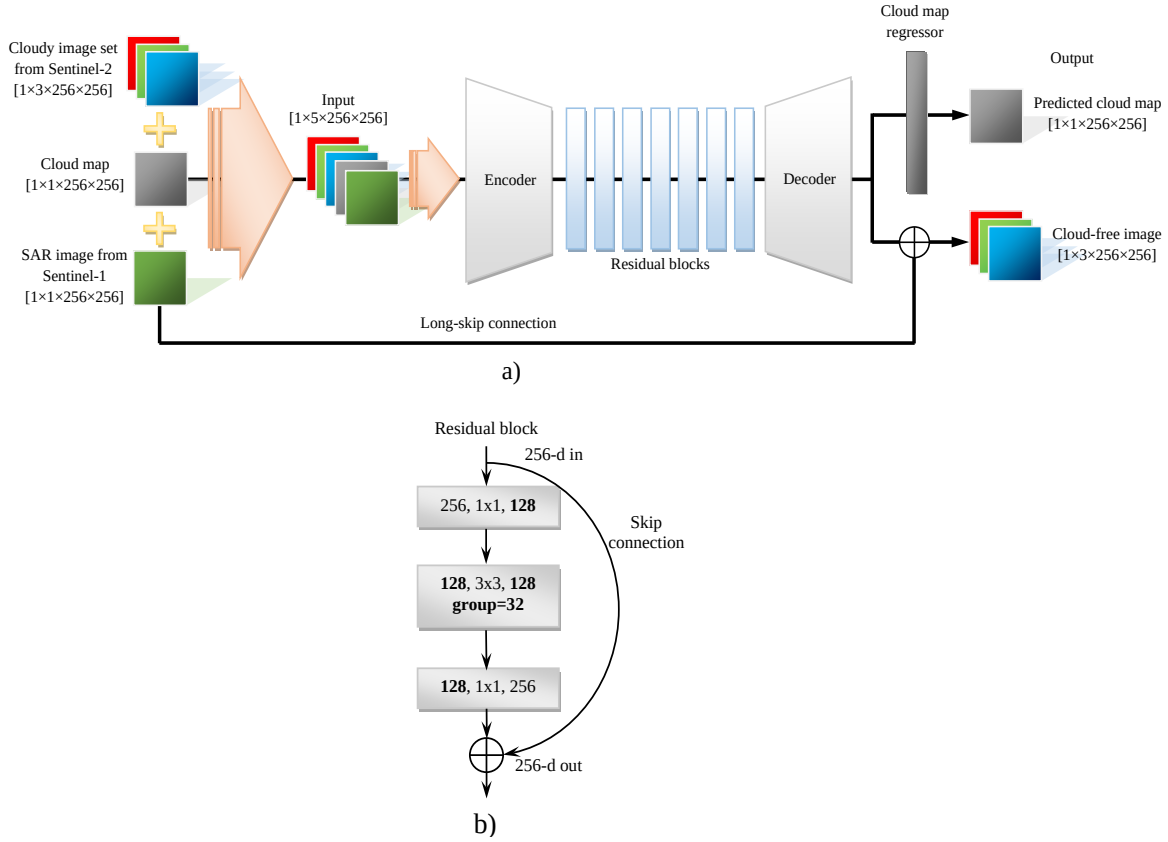

Figure 1: Model for cloud removal in optical satellite imagery by generative adversarial networks: model architecture (a); residual block (b).

introduced by heavily obscured (long-missing) pixels, an inverse hyperbolic tangent transformation is applied. This step effectively demodulates the cloud-induced distortion. As a result, the trained model can reconstruct a cloud-free optical image from previously cloud-contaminated input.

For the loss function, we use the cycleGAN approach [50] for both generator and discriminator networks, which can be expressed as:

$$L_{cGAN} = E_{x,y\ P_{data}(x,y))}[\log(D(x,y))] + E_{x,z\ P_{data}(x,y)}[\log(1 - D(x, G(x,z)))] \quad (1)$$

where D – the discriminator network, G – the generator network, and D tries to maximize and G tries to minimize an objective.

4.2. Multi-scale feature extraction and adaptive object localisation

The architecture of the proposed model is based on a convolutional neural network ResNeXt [42], integrated within the BiFPN (bidirectional feature pyramid network) framework [36]. ResNeXt, as a modular and high-capacity architecture, offers a large receptive field due to its use of aggressive convolutional operations. BiFPN, in turn, enables repeated, bidirectional multi-scale feature fusion – top-down and bottom-up – which allows efficient reuse of feature maps at multiple scales. This facilitates the capture of fine-grained, low-level features while preserving higher-level semantic information, thereby enhancing object recognition across a wide range of scales with fewer parameters compared to conventional augmented CNNs. The model is optimised for hardware efficiency, a critical factor in the joint training of semantic and instance segmentation models.

At the apex of the BiFPN structure, a deformable convolutional network (DCN) [39] is employed. DCNs adapt the target function to geometric variations of objects in aerial imagery by acknowledging that not all pixels within the receptive field equally contribute to the final response of the network.

The variation in these contributions is characterised by the effective receptive field, which is quantified as the gradient of the layer node response to pixel intensity perturbations. By incorporating learnable offsets and shifts into convolutional layers, DCNs can dynamically adjust their sampling patterns, making them more flexible in handling deformations and transformations of object structures. This significantly enhances the accuracy and robustness of object detection and recognition in remote sensing imagery.

To further improve object localisation and adapt bounding boxes to object shapes, a guided anchoring region proposal network (GA-RPN) [6] is applied after the BiFPN stage. The necessity of GA-RPN arises from the uneven spatial distribution of objects and the strong dependence of object scale and aspect ratio on contextual background elements. The GA-RPN module comprises two parallel branches: a location prediction branch that generates a probability map of potential object locations, and a shape prediction branch that determines the likely aspect ratios at these locations. Based on the outputs from both branches, a refined set of anchors is generated by selecting positions with probabilities exceeding a threshold and associating them with the most plausible object shapes.

Since object shapes may vary across image regions, a feature adaptation module is included to select appropriate anchor forms depending on the feature representation. This leads to a multi-level anchor generation strategy, where anchors are drawn from multiple BiFPN layers, ensuring accurate scale-aware localisation. As a result, each object is linked to a single dynamically predicted anchor rather than a predefined anchor set. The feature representations used for anchor generation are taken from the corresponding level of the BiFPN hierarchy.

4.3. Hybrid task cascade for instance segmentation

Although combining bounding box detection and mask generation can yield some improvements, using them in a cascade fashion proves more effective for enhancing localisation and recognition accuracy. A cascading approach ensures more refined hypothesis generation at each stage, mitigating issues such as vanishing positive samples and mismatch between training IoU thresholds and incoming predictions. However, standard cascade architectures suffer from a discontinuity in information flow between stages, which limits the refinement of mask predictions and results in only modest improvements in bounding box accuracy [1].

To address this limitation, the proposed model incorporates a hybrid task cascade (HTC) for instance segmentation [2]. The core innovation of HTC is the integration of spatial context and multi-task learning at each cascade stage, ensuring continuous information flow and improving both detection and segmentation performance. This approach significantly boosts the model's ability to distinguish foreground objects from complex, cluttered backgrounds by leveraging semantic segmentation as a context-aware filter.

As a result of these enhancements, the improved hybrid segmentation cascade model demonstrates increased effectiveness in multi-stage aerial image processing, delivering more accurate and reliable object recognition. The structural diagram of the model is shown in figure 2, where I denotes the input image, $Pool$ is the regional feature pooling, and B_t and M_t represent the bounding box and mask outputs at stage t, respectively.

To enhance object detection performance, the model utilises scene context as guidance for integrating semantic branches, facilitating accurate categorisation and scale determination. Features extracted from each BiFPN level are passed to a pyramidal pooling module [20], which performs semantic segmentation of background and contextual elements at the pixel level. This operation ensures that semantic information from different subregions of the scene is preserved and leveraged to distinguish between foreground and background areas, thereby minimising context loss.

The PPM is responsible for aggregating features into a unified representation that combines both local and global contextual information. It incorporates feature maps from five BiFPN layers. The topmost semantic layer performs global pooling to generate a single feature vector that captures the overall scene context. Subsequent levels in the pyramid divide feature maps into increasingly finer spatial subregions, forming localised feature representations for different areas of the image.

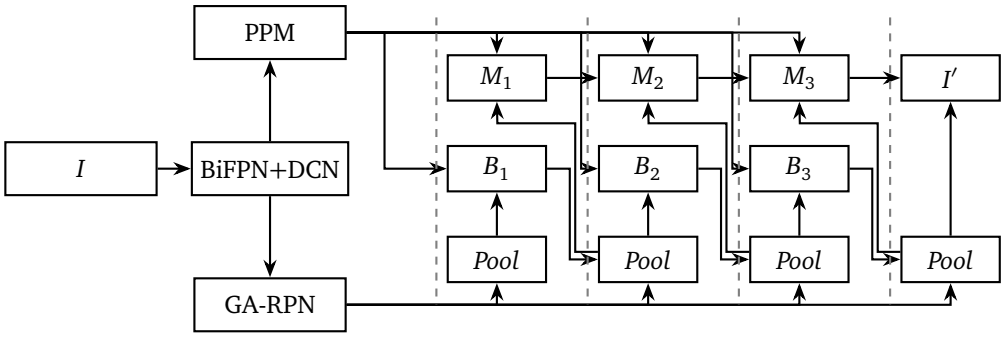

Figure 2: Improved model of hybrid segmentation cascade.

To preserve global semantic weighting, a 1×1 convolutional layer is applied to the outputs of each BiFPN layer. Bilinear interpolation is used to merge low-level BiFPN outputs, ensuring smooth spatial alignment across scales during the final global feature combination.

The cascade semantic branch – trained jointly with other branches – predicts per-pixel semantic labels for the entire image using a fully convolutional architecture. It enhances the model's capacity to distinguish foreground objects from cluttered or flooded backgrounds, thereby complementing the outputs of bounding box detectors and instance mask generators. This additional semantic segmentation stream enhances the reliability and precision of object recognition by incorporating spatial context information that is not captured by the bounding box or mask branches alone.

Unlike conventional cascade architectures that process bounding boxes and masks in parallel, the proposed approach introduces several key innovations. First, it performs sequential regression of bounding boxes and mask predictions, improving accuracy through progressive refinement. Second, it enables direct information transfer between mask branches across stages, preserving specific characteristics identified in earlier layers. Third, it incorporates an auxiliary semantic segmentation branch designed to learn deeper contextual features and align with both the bounding box and mask prediction branches.

This joint configuration enhances the overall information flow throughout the network, resulting in superior localisation and classification performance. By training the detector with progressively increasing intersection over union (IoU) thresholds, the model becomes more selective and robust to near-miss detections. The flow of information across cascade stages is summarised by the following set of expressions (figure 2):

$$x_t^{box} = P(x, r_{t-1}) + P(S(x), r_{t-1}) \tag{2}$$

$$x_t^{mask} = P(x, r_t) + P(S(x), r_t) \tag{3}$$

$$r_t = B_t(x_t^{box}) \tag{4}$$

$$m_t = M_t(F(x_t^{mask}, m_{t-1}^{-})) \tag{5}$$

where x_t^{box}, x_t^{mask} – detected by bounding box and feature masks; $P(x, r_{t-1})$ – align operation RoI align [12]; $B_t(x_t^{box})$, $M_t(x_t^{mask})$ – definition of bounding box and mask at stage t; r_t, m_t – prediction of bounding boxes and sample masks; S – head of semantic segmentation.

The training of the suggested cascade includes class predictions, bounding box and mask regression, and is performed in a mode that runs from beginning to end. The general loss function takes the form of multi-task training at each iteration and looks like this:

$$L = \sum_{t=1}^{T} \alpha_t (L_{bbox}^t + L_{mask}^t) + L_{seg} \tag{6}$$

$$L_{bbox}^t(c_i, r_i, l_i, s_i, \hat{c}_t, \hat{r}_t, \hat{l}_t, \hat{s}_t) = L_{reg}(r_t, \hat{r}_t) + \lambda_1 L_{loc}(l_i, \hat{l}_t) + \lambda L_{shape}(s_i, \hat{s}_t) \tag{7}$$

$$L^t_{mask}(m_t, \hat{m}_t) = BCE(m_t, \hat{m}_t) \tag{8}$$

$$L_{seg} = CE(s, \hat{s}) \tag{9}$$

where L – general loss function; L^t_{bbox}, L^t_{mask} – loss of bounding box prediction and mask at stage t; L_{cls}, L_{reg} – loss of classification prediction and object image regularization; L_{loc}, L_{shape} – losses of anchor localization and anchor form prediction; L_{segm} – loss of semantic segmentation prediction; CE – loss function of cross entropy; BCE – loss function of binary cross entropy.

4.4. Modified focal loss for imbalanced dataset training

During the creation of training samples for each object class in a newly constructed dataset of aerial images, a significant class imbalance often arises due to the insufficient number of instances for certain classes. When training the model using the standard cross-entropy loss function under such conditions, the loss values for underrepresented classes tend to diminish rapidly as the model becomes increasingly confident in the dominant classes. This leads to suboptimal learning for minority classes.

To address this issue, resampling techniques are commonly employed. However, based on recent research findings, a more effective solution is to adopt a modified version of the focal loss function, specifically designed to enhance model performance on imbalanced datasets. In this approach, the conventional cross-entropy loss function is replaced with the focal loss, which dynamically down-weights the contribution of well-classified examples and focuses learning on hard, misclassified instances. So, instead of the cross-entropy loss function:

$$CE(p_t) = -\log(p_t) \tag{10}$$

very often the function of focal loss [19] is used

$$FL(p_t) = -(1-p_t)\log(p_t) \tag{11}$$

where FL – focal loss, CE – loss function of cross entropy, p_t – probability of credible class, γ – focusing value.

Focal loss aims to reduce the impact of well-classified samples on the overall loss, thereby shifting the training focus toward harder, misclassified examples. Originally developed to address extreme class imbalance in one-stage object detection models, focal loss is particularly effective when there is a disproportionate distribution between frequent (dense) and rare (sparse) object classes.

However, this approach does not always yield optimal results in two-stage detectors, where background regions are filtered out at the initial stage. In such scenarios, focal loss may overemphasise complex examples, which can negatively affect training stability.

To mitigate this, we propose modifying the focal loss function to soften its sensitivity to complex examples. Specifically, the modification applies equal weights to all positive samples whose predicted probabilities fall below a certain threshold. Simultaneously, it continues to suppress the contribution of confidently predicted (well-classified) samples. This modified scaling mechanism maintains the essence of the original focal loss while introducing a threshold-based smoothing factor that enhances training robustness (figure 3).

The proposed formulation can be expressed as follows:

$$MFL(p_t) = -f(p_t, t_h)\log(p_t) \tag{12}$$

where $f(p_t, t_h)$ – rejection ratio that scales the loss function by next formula:

$$f(x) = \begin{cases} 1: & p_t < t_h \\ \frac{(1-p_t)^\gamma}{t_h^\gamma}: & p_t \geqslant t_h \end{cases} \tag{13}$$

where t_h – probability of fundamental truth class. The focal loss modification function helps improve the average accuracy of object detection mAP for sparse classes; however, mAP is decreased slightly for well-flooded classes. The function of the modified focal loss application reduces the effect of the class imbalance factor during model training.

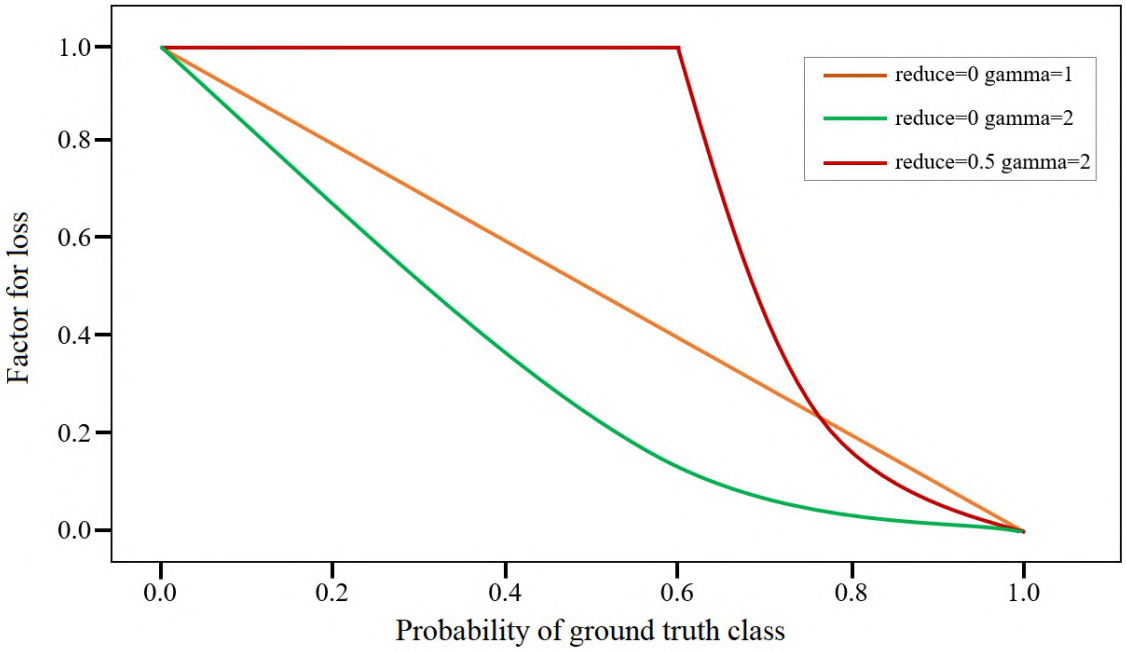

Figure 3: Dependence of rejection ratio from the class probability of validation set.

4.5. Experimental setup and performance evaluation

For the experiment, we used a large-scale dataset of globally selected orthocorrected, georeferenced cloudy and cloud-free 13-band Sentinel-2 multispectral images from triple-shot imagery, as well as the corresponding Sentinel-1 image from 175 globally distributed regions of interest (ROIs) evenly spaced across all continents and meteorological seasons, called SEN12MS-CR [8], for cloud removal. Cloud-free optical images of four ROI samples acquired in four different meteorological seasons. Each scene in the dataset is converted to the Mercator coordinate system and then divided into 256×256 px^2 plots with a spatial overlap of 50% between adjacent plots, resulting in an average of over 700 plots per ROI. The average cloud cover value is approximately 55% of the global share of cloud cover over land.

To test the improved model of the hybrid segmentation cascade and study the process in accordance with the task, a dataset was created to detect vehicles in satellite images. The modelling did not take into account the FPS latency, which is not a key factor in satellite image processing. As a result of the object distribution, 10 vehicle classes were formed. The set of object classes is not balanced (the number of object images in the classes varies from 7 to 2454), and the vehicle images differ significantly in size, aspect ratio, brightness distribution, and colour density.

Online augmentation was used to enlarge object images, taking into account the recording conditions (rotations to 0°, 90°, 180°, 270°), as well as adding Gaussian noise, contrast, sharpness, and colour density changes. The transfer learning approach was used with the trained models on the COCO Detection dataset. For the model work assessment metrics, mAP was used, which calculates the mAP average score value for variables IoU to identify a large number of bounding boxes with incorrect classifications, and enables avoiding maximum specialisation in several classes at the expense of weak projections in others.

To adapt the target function presentation for the object configuration the deforming convolution at BiFPN top was used that applies high level of feature synthesis; for fewer anchors use and taking into account of their possible form and size the guided anchorage method is applied; for further information loss reduction in the context among various sub-regions the hierarchical global previous

content is applied – PPM module enables to combine the features from five various FPN scales.

To improve the model operation quality, the approach of triple increase of testing time for aerial image pre- and post-processing (image compilation with resolution 600 × 600, 700 × 700 and turn (0°, 90°, 180°, 270°), with augmentation to 800 × 800, 900 × 900, 1000 × 1000).

Model training was conducted over 18 epochs. The results obtained are shown in table 2.

Table 2
Dependence of mAP value depending on model improvements is applied.

Improvements	mAP (at IoU ≥0.7)
Base model	63.2
DCN	64.6
DCN + GA-RPN	65.6
DCN + GA-RPN + PPM	65.9
DCN + GA-RPN + PPM + MFL	66.2

As a result of the hybrid task cascade model improvement, along with image set growth and post-processing, the mAP accuracy was improved by 3%. It enables the increase in credibility of small-sized object detection in satellite imagery.

5. Conclusions and further research

Cloud cover in satellite imagery remains a significant obstacle to the full exploitation of remote sensing data, substantially limiting both temporal and spatial availability of Earth surface observations. Based on an analysis of medium spatial resolution remote sensing data and a comparison of existing cloud removal techniques, a new model has been developed that reconstructs cloud-free optical images by leveraging the synergistic integration of synthetic aperture radar and optical data. The inclusion of SAR data provides supplementary information that enhances the robustness and reliability of the reconstruction process. The proposed attention-based neural network has been evaluated on the SEN12MS-CR dataset.

In the course of analysing neural network-based methods for automatic image processing and object detection, specific attention was paid to the effects of object deformation, occlusion, and variability in background context. To address these challenges, a hybrid task cascade approach for instance segmentation was adopted. This model enhances information flow through staged multitask processing and incorporates indirect features of topographic elements to improve the reliability of object detection and recognition.

The final model demonstrates a 3% increase in detection accuracy (measured by mAP) compared to commonly used baseline models when applied to satellite imagery.

Future work will focus on further refining the methodology for cloud removal in multispectral satellite images, particularly in the context of medium-resolution remote sensing data.

References

[1] Cai, Z. and Vasconcelos, N., 2018. Cascade R-CNN: Delving Into High Quality Object Detection. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, pp.6154–6162. Available from: https://doi.org/10.1109/CVPR.2018.00644.

[2] Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C.C. and Lin, D., 2019. Hybrid Task Cascade for Instance Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, pp.4974–4983. Available from: https://doi.org/10.1109/CVPR.2019.00511.

[3] Cheng, M., Liu, Y., Lin, W., Zhang, Z., Rosin, P.L. and Torr, P.H.S., 2019. BING: Binarized normed gradients for objectness estimation at 300fps. *Computational Visual Media*, 5(1), pp.3–20. Available from: https://doi.org/10.1007/S41095-018-0120-1.

[4] Cheng, Q., Zhang, H.S.L., Yuan, Q. and Zhen, C., 2014. Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92, pp.54–68. Available from: https://doi.org/10.1016/j.isprsjprs.2014.02.015.

[5] Dai, J., He, K., Li, Y., Ren, S. and Sun, J., 2016. Instance-Sensitive Fully Convolutional Networks. In: B. Leibe, J. Matas, N. Sebe and M. Welling, eds. *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*. Springer, *Lecture Notes in Computer Science*, vol. 9910, pp.534–549. Available from: https://doi.org/10.1007/978-3-319-46466-4_32.

[6] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H. and Wei, Y., 2017. Deformable Convolutional Networks. *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp.764–773. Available from: https://doi.org/10.1109/ICCV.2017.89.

[7] Dollár, P., Tu, Z., Perona, P. and Belongie, S.J., 2009. Integral Channel Features. In: A. Cavallaro, S. Prince and D.C. Alexander, eds. *British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings*. British Machine Vision Association, pp.1–11. Available from: https://doi.org/10.5244/C.23.91.

[8] Ebel, P., Xu, Y., Schmitt, M. and Zhu, X.X., 2022. SEN12MS-CR-TS: A Remote-Sensing Data Set for Multimodal Multitemporal Cloud Removal. *IEEE Transactions Geoscience Remote Sensing*, 60, pp.1–14. Available from: https://doi.org/10.1109/TGRS.2022.3146246.

[9] Eckardt, R., Berger, C., Tiel, C. and Schmullius, C., 2013. Removal of Optically Thick Clouds from Multi-Spectral Satellite Images Using Multi-Frequency SAR Data. *Remote Sensing*, 5(6), pp.2973–3006. Available from: https://doi.org/10.3390/rs5062973.

[10] Gidaris, S. and Komodakis, N., 2016. Attend Refine Repeat: Active Box Proposal Generation via In-Out Localization. In: R.C. Wilson, E.R. Hancock and W.A.P. Smith, eds. *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press. Available from: https://bmva-archive.org.uk/bmvc/2016/papers/paper090/index.html.

[11] Girshick, R.B., 2015. Fast R-CNN. *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, pp.1440–1448. Available from: https://doi.org/10.1109/ICCV.2015.169.

[12] He, K., Gkioxari, G., Dollár, P. and Girshick, R.B., 2017. Mask R-CNN. *CoRR*, abs/1703.06870. Available from: http://arxiv.org/abs/1703.06870.

[13] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, pp.770–778. Available from: https://doi.org/10.1109/CVPR.2016.90.

[14] Isola, P., Zhu, J., Zhou, T. and Efros, A.A., 2017. Image-to-Image Translation with Conditional Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp.5967–5976. Available from: https://doi.org/10.1109/CVPR.2017.632.

[15] Kovbasiuk, S.V., Kanevskyy, L.B., Romanchuk, M.P., Chernyshuk, S.V. and Naumchak, L.M., 2023. Object detection method based on aerial image instance segmentation received by unmanned aerial vehicles in the conditions rough for visualization. In: T.A. Vakaliuk and S.O. Semerikov, eds. *Proceedings of the 3rd Edge Computing Workshop, Zhytomyr, Ukraine, April 7, 2023*. CEUR-WS.org, *CEUR Workshop Proceedings*, vol. 3374, pp.41–55. Available from: https://ceur-ws.org/Vol-3374/paper03.pdf.

[16] Li, X., Wang, L., Cheng, Q., Wu, P., Gan, W. and Fang, L., 2019. Cloud removal in remote sensing images using nonnegative matrix factorization and error correction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 148, pp.103–113. Available from: https://doi.org/10.

1016/j.isprsjprs.2018.12.013.

[17] Li, Y., Qi, H., Dai, J., Ji, X. and Wei, Y., 2017. Fully Convolutional Instance-Aware Semantic Segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp.4438–4446. Available from: https://doi.org/10.1109/CVPR.2017.472.

[18] Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B. and Belongie, S.J., 2017. Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp.936–944. Available from: https://doi.org/10.1109/CVPR.2017.106.

[19] Lin, T., Goyal, P., Girshick, R.B., He, K. and Dollár, P., 2017. Focal Loss for Dense Object Detection. *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp.2999–3007. Available from: https://doi.org/10.1109/ICCV.2017.324.

[20] Liu, S., Qi, L., Qin, H., Shi, J. and Jia, J., 2018. Path Aggregation Network for Instance Segmentation. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, pp.8759–8768. Available from: https://doi.org/10.1109/CVPR.2018.00913.

[21] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C. and Berg, A.C., 2016. SSD: Single Shot MultiBox Detector. In: B. Leibe, J. Matas, N. Sebe and M. Welling, eds. *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*. Springer, *Lecture Notes in Computer Science*, vol. 9905, pp.21–37. Available from: https://doi.org/10.1007/978-3-319-46448-0_2.

[22] Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, pp.3431–3440. Available from: https://doi.org/10.1109/CVPR.2015.7298965.

[23] Lv, H., Wang, Y. and Shen, Y., 2016. An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands. *Remote Sensing of Environment*, 179, pp.183–195. Available from: https://doi.org/10.1016/j.rse.2016.03.034.

[24] Meng, F., Yang, X., Zhou, C. and Li, Z., 2017. A Sparse Dictionary Learning-Based Adaptive Patch Inpainting Method for Thick Clouds Removal from High-Spatial Resolution Remote Sensing Imagery. *Sensors*, 17(9), p.2130. Available from: https://doi.org/10.3390/s17092130.

[25] Meng, Q., Borders, B., Cieszewski, C. and Madden, M., 2009. Closest Spectral Fit for Removing Clouds and Cloud Shadows. *Photogrammetric Engineering & Remote Sensing*, 75, pp.569–576. Available from: https://doi.org/10.14358/PERS.75.5.569.

[26] Pilkevych, I.A., Romanchuk, M.P., Naumchak, O.M., Fedorchuk, D.L. and Naumchak, L.M., 2025. Improved model for detecting randomly oriented objects on remote sensing images. In: T.A. Vakaliuk and S.O. Semerikov, eds. *Proceedings of the 5th Edge Computing Workshop (doors 2025), Zhytomyr, Ukraine, April 4, 2025*. CEUR-WS.org, *CEUR Workshop Proceedings*, vol. 3943, pp.118–126. Available from: https://ceur-ws.org/Vol-3943/paper26.pdf.

[27] Redmon, J., Divvala, S.K., Girshick, R.B. and Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, pp.779–788. Available from: https://doi.org/10.1109/CVPR.2016.91.

[28] Rees, W., 2012. *Physical Principles of Remote Sensing*. 3rd ed. Cambridge University Press. Available from: https://doi.org/10.1017/CBO9781139017411.

[29] Ren, S., He, K., Girshick, R.B. and Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama and R. Garnett, eds. *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. pp.91–99. Available from: https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

[30] Romanchuk, M., Zavada, A., Naumchak, O., Naumchak, L. and Kosheva, I., 2024. Removing cloudiness on optical space images by a generative adversarial network model using SAR images. *Eastern-European Journal of Enterprise Technologies*, 5(2 (131)), pp.6–12. Available from: https://doi.org/10.15587/1729-4061.2024.313690.

[31] Saha, S., Bovolo, F. and Bruzzone, L., 2018. Destroyed-buildings detection from VHR SAR images using deep features. In: L. Bruzzone and F. Bovolo, eds. *Image and Signal Processing for Remote Sensing XXIV*. International Society for Optics and Photonics, SPIE, vol. 10789, p.107890Z. Available from: https://doi.org/10.1117/12.2325149.

[32] Schowengerdt, R., 2007. *Remote Sensing, Models and Methods for Image Processing*. 3rd ed. Elsevier. Available from: https://doi.org/10.1016/B978-0-12-369407-2.X5000-1.

[33] Shen, H., Li, X., Cheng, Q., Zeng, C., Yang, G., Li, H. and Zhang, L., 2015. Missing Information Reconstruction of Remote Sensing Data: A Technical Review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3), pp.61–85. Available from: https://doi.org/10.1109/MGRS.2015.2441912.

[34] Siravenha, A., Sousa, D., Bispo, A. and Pelaes, E., 2011. Evaluating Inpainting Methods to the Satellite Images Clouds and Shadows Removing. *Signal Processing, Image Processing and Pattern Recognition - International Conference, SIP 2011, Held as Part of the Future Generation Information Technology Conference FGIT 2011, in Conjunction with GDC 2011, Jeju Island, Korea, December 8-10, 2011. Proceedings*. Springer, *Communications in Computer and Information Science*, vol. 260, pp.56–65. Available from: https://doi.org/10.1007/978-3-642-27183-0_7.

[35] Sun, M., Kim, B.S., Kohli, P. and Savarese, S., 2014. Relating Things and Stuff via ObjectProperty Interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), pp.1370–1383. Available from: https://doi.org/10.1109/TPAMI.2013.193.

[36] Tan, M., Pang, R. and Le, Q.V., 2020. EfficientDet: Scalable and Efficient Object Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, pp.10778–10787. Available from: https://doi.org/10.1109/CVPR42600.2020.01079.

[37] Tighe, J., Niethammer, M. and Lazebnik, S., 2014. Scene Parsing with Object Instances and Occlusion Ordering. *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, pp.3748–3755. Available from: https://doi.org/10.1109/CVPR.2014.479.

[38] Tu, Z., Chen, X., Yuille, A. and Zhu, S., 2005. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, 63(2), pp.113–140. Available from: https://doi.org/10.1007/S11263-005-6642-X.

[39] Wang, J., Chen, K., Yang, S., Loy, C.C. and Lin, D., 2019. Region Proposal by Guided Anchoring. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, pp.2965–2974. Available from: https://doi.org/10.1109/CVPR.2019.00308.

[40] Wang, L., Xu, X., Yu, Y., Yang, R., Gui, R., Xu, Z. and Pu, F., 2019. SAR-to-Optical Image Translation Using Supervised Cycle-Consistent Adversarial Networks. *IEEE Access*, 7, pp.129136–129149. Available from: https://doi.org/10.1109/ACCESS.2019.2939649.

[41] Windy.com. Available from: https://www.windy.com/?50.452,30.529,5.

[42] Xie, S., Girshick, R.B., Dollár, P., Tu, Z. and He, K., 2017. Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp.5987–5995. Available from: https://doi.org/10.1109/CVPR.2017.634.

[43] Xu, M., Jia, X., Pickering, M. and Jia, S., 2019. Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149, pp.215–225. Available from: https://doi.org/10.1016/j.isprsjprs.2019.01.025.

[44] Yang, B., Yan, J., Lei, Z. and Li, S.Z., 2016. CRAFT Objects from Images. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, pp.6043–6051. Available from: https://doi.org/10.1109/CVPR.2016.650.

[45] Yao, J., Fidler, S. and Urtasun, R., 2012. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. IEEE Computer Society, pp.702–709. Available from: https://doi.org/10.1109/CVPR.2012.6247739.

[46] Yu, C., Chen, L., Su, L., Fan, M. and Li, S., 2011. Kriging interpolation method and its application in retrieval of MODIS aerosol optical depth. *2011 19th International Conference on Geoinformatics*. pp.1–6. Available from: https://doi.org/10.1109/GeoInformatics.2011.5981052.

[47] Zhang, Q., Yuan, Q., Zeng, C., Li, X. and Wei, Y., 2018. Missing Data Reconstruction in Remote Sensing Image With a Unified Spatial–Temporal–Spectral Deep Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8), pp.4274–4288. Available from: https://doi.org/10.1109/TGRS.2018.2810208.

[48] Zhang, X., Zhang, T., Wang, G., Zhu, P., Tang, X., Jia, X. and Jiao, L., 2023. Remote Sensing Object Detection Meets Deep Learning: A metareview of challenges and advances. *IEEE Geoscience and Remote Sensing Magazine*, 11(4), pp.8–44. Available from: https://doi.org/10.1109/MGRS.2023.3312347.

[49] Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J., 2017. Pyramid Scene Parsing Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp.6230–6239. Available from: https://doi.org/10.1109/CVPR.2017.660.

[50] Zhu, J., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp.2242–2251. Available from: https://doi.org/10.1109/ICCV.2017.244.

[51] Zhu, X.X., Montazeri, S., Ali, M., Hua, Y., Wang, Y., Mou, L., Shi, Y., Xu, F. and Bamler, R., 2021. Deep Learning Meets SAR: Concepts, models, pitfalls, and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 9(4), pp.143–172. Available from: https://doi.org/10.1109/MGRS.2020.3046356.