

# Object detection method based on aerial image instance segmentation received by unmanned aerial vehicles in the conditions rough for visualization

Serhiy V. Kovbasiuk<sup>1</sup>, Leonid B. Kanevskyy<sup>1</sup>, Mykola P. Romanchuk<sup>1</sup>,  
Serhiy V. Chernyshuk<sup>1</sup> and Leonid M. Naumchak<sup>1</sup>

<sup>1</sup>Korolyov Zhytomyr Military Institute, 22 Myru Ave., Zhytomyr, 10004, Ukraine

## Abstract

The article analyses the possibilities to use the unmanned aerial complexes in the system of decision making process for the crisis situations that require the object detection at aerial images received by the unmanned aerial vehicle under the conditions of atmospheric fog and smoke over the territories. For image sharpening we used Pansharpening method for injecting the dimensional details from panchromatic image to multispectral image. In order to increase the operational efficiency and accuracy of automotive vehicles detection at aerial images received by the unmanned aerial vehicles for more efficient use of received information in the system of decision making support it was selected Hybrid Task Cascade for Instance Segmentation model. This model is more appropriate for solving the tasks of small-sized object multiclass classification and detection at aerial image using the indirect signs.

## Keywords

recognition, object detection, aerial photo-images, Pansharpening, instance segmentation, focal loss, unmanned aerial vehicles

## 1. Introduction

Some ten even five years ago the unmanned aerial vehicles (UAVs) were regarded skeptically as the ex-pensive toys for entertainment – to film the landscapes, animals, make photos from a bird's perspective over the reserved areas and so on. It was interesting only for quite few devoted people.

The contemporary situation all over the world concerning COVID-19 (SARS-CoV-2) epidemics placed new demands on mankind for communication, behavior and living [1, 2]. In general, we are talking about noncontact communications and various service rendering. First of all it touches upon assistance and danger identification in the cities and hard-to-access areas. The first

---


*doors-2023: 3rd Edge Computing Workshop, April 7, 2023, Zhytomyr, Ukraine*

✉ klasik552008@gmail.com (S. V. Kovbasiuk); leo10k10@ukr.net (L. B. Kanevskyy); romannik@ukr.net (M. P. Romanchuk); rekryt2002@gmail.com (S. V. Chernyshuk); naumchak.leonid@gmail.com (L. M. Naumchak)  
🌐 <https://ieeexplore.ieee.org/author/37087014573> (S. V. Kovbasiuk); <https://ieeexplore.ieee.org/author/37087014236> (L. B. Kanevskyy); <https://ieeexplore.ieee.org/author/37087013658> (M. P. Romanchuk); <https://ieeexplore.ieee.org/author/37088397045> (S. V. Chernyshuk); <https://ieeexplore.ieee.org/author/37089179498> (L. M. Naumchak)

🆔 0000-0002-6003-7660 (S. V. Kovbasiuk); 0000-0002-3298-5866 (L. B. Kanevskyy); 0000-0002-0087-8994 (M. P. Romanchuk); 0000-0003-2859-3306 (S. V. Chernyshuk); 0000-0002-7311-6659 (L. M. Naumchak)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

steps in this direction were made in November 2019 when COVID-19 (SARS-CoV-2) pandemic was in the initial stage but China already used UAVs to detect the isolation trespassers, potential sick people and even to monitor the body temperature by thermal imaging scanning.

Another important UAV task was the drugs and other important items delivery to the people on self-isolation (food, hygienic stuff, essential goods). There was also carried out the monitoring and control over the fires and other hazardous objects in hard-to-access areas [3, 4].

One of the most important elements affecting such tasks is the visualization system (information display) and information processing technologies. Often, one of the reasons making impossible using the visual control principles of UAV landing or information gathering concerning the objects at the Earth and the very Earth as bottoming surface is the atmospheric fog, smoke over the ground and imperfect (not adapted for such conditions) methods of object detection at the aerial photo-images received by the UAVs.

In the conditions of low possibility to take into account all factors concerning the UAV (visual confirmation) it may cause the task failure or flight safety violation. Accordingly, the key markers for delivery or situation monitoring using the UAVs in the conditions rough for visualization are the abilities to assess fast and reliably the area where the automatic object detection, recognition and classification means above the Earth ground may prove justifiable.

The contemporary visualization systems enable to represent huge information volumes from various sources of spatial basing: spaceships as the Earth surface optical-electronic monitoring and remote sensing, and UAVs. Usually, information from such sources does not contain the intermediate conclusions concerning monitoring that complicates the sequence of events forecast and executive decision making. To solve such problem in the automatic mode the gathered information processing is carried out – thematic aerial image processing. The thematic processing and data complexation from all aforementioned means enables the overall situation assessment in the given Earth area.

Such method of information gathering requires using system analysis and synthesis method of different time and parameter data from physically different means of information gathering. For qualitative incoming traffic transformation process of separated data from all the sources of spatial basing into a single final result fit for using under the complicated visual conditions it is necessary to determine the main components (phases) of thematic processing, logical links of various structural data complexation study along with determination of evolving problems and possible means of their solution.

In the framework of solution of the new tasks for noncontact communications using the UAVs it is necessary to search and develop an efficient (operative and sufficiently reliable) detection method of fine-grained objects at aerial images received by the UAVs both in simple conditions and in the conditions rough for visualization.

The purpose of the article is to analyze the application of object detection neural network models for UAV image processing in conditions of atmospheric haze and smog, with their further improvement to increase the accuracy of localization and recognition of objects on the ground surface.

## 2. Related works

Based on the analysis of atmosphere transmission over Ukraine in 2019 given in table 1 as for classical visualization of image results at aerial photo-images from various sources it is possible to conclude that depending on the season 35 percent of daytime per year is clouded and require special methods of object detection at the aerial images received under such conditions.

**Table 1**

Analysis of cloud coverage over Ukraine in 2019.

Region	January	February	March	April	May	June	July	August	September	October	November	December
AR Crimea												
Vinnitsia												
Volyn												
Dnipropetrovsk												
Donetsk												
Zhytomyr												
Zakarpattia												
Zaporizhzhya												
Ivano-Frankivsk												
Kyiv												
Kirovohrad												
Lugansk												
Lviv												
Mykolayiv												
Odesa												
Poltava												
Rivne												
Sumy												
Ternopil												
Kharkiv												
Kherson												
Khmelnyskyi												
Cherkasy												
Chernivtsi												
Chernihiv												
UKRAINE												
Over 70% of time the sky was cloudy during that month												
From 25% to 70% of time the sky was cloudy during that month												
Up to 25% of time the sky was clouded, and 75% it was clear during that month												

Smoke is one of the emergency situations factors, which excludes the possibility of using detectors for processing aerial photographs from UAVs. The Pansharpening method, which is based on the use of spatial details injections from panchromatic image to a multispectral image,

showed better results for improving the original image in the presence of atmospheric haze or smoke during fires. Currently, the following injection models can be distinguished: the Gram-Schmidt projection model of orthogonalization, which was underlined the spectral sharpening [5] and context-oriented solution [6] methods; a model based on modulation, underlined the developing of high-frequency modulation [7], synthetic variable coefficients [8], and models of spectral distortions minimizing [9, 10]. The contrast-based model is inherently local, or context-adaptive [11], unlike the projection model, as the injection gain varies at each pixel [12].

For the task solution of efficient object detection and recognition at images there have been used the methods of image semantic and instance segmentation which are developing in parallel and which have their peculiarities, advantages and disadvantages.

The methods of semantic segmentation that use convolution neural networks (CNN) solve the task of detection and recognition from their multilevel aggregation or from through structural prediction [13]. Using the augmented CNN [14], as networks of pyramidal scenes analysis [13] that uses the phalanx pyramid module (PPM) and feature pyramid (FPN) [15, 16] that enable to keep high resolution till the last layer, has increased the efficiency of context receiving.

Instance segmentation allows solving the tasks of actual semantic class object identification related to an aerial image pixel. Starting from the regional CNN (R-CNN) [17] the instance segmentation is performed by two-stage principle: from the generated sequence of segmented proposals the comparison of the best one is carried out [18, 19]. The common for those methods of instance segmentation is segmentation by the regional proposal network (RPN) before the object classification. In InstanceFCN [20] mask proposals are received from full convolution network (FCN) [19]. MNC [21] uses sample segmentation as conveyor which work is composed of three subtasks: object mask localization, forecasting and categorization, and through cascade method it trains the neural network. InstanceFCN implementation is usage of full convolution approach for instance segmentation. Model Mask R-CNN adds additional branch based on Faster R-CNN [22] and uses common approach to forming the limits and masks when two target functions in parallel solve separate tasks that increases the accuracy of object localization and its recognition on the aerial image. PANet [23] uses bilateral information flow in FPN [24].

Background object classifiers that use semantic segmentation methods usually built on FCN with extensions [13] do not stipulate the sample limits for classes. The methods of instance segmentation based on detectors that usually use the object proposals based on offered areas [25, 19] ignore the background objects making impossible to use non-directs features. Their combination enables to solve the task of scene analysis [15], image review [16] or scene integral understanding [19].

To increase reliability of fine-grained objects detection two-stage detectors have been developed [14, 22, 26], which compared with the one-stage ones [27, 28] are characterized by optimization and possibility to generate sufficient number of high level features. In particular, in multi-regional CNN [29] the iterative mechanism of detection for specification of limits is used. Detector AttractionNet [30] uses module Attend&Refine for renewal of limiting places iteratively. Models CRAFT [31] and Fast R-CNN [27] for detection credibility growth include the cascade structure in RPN [22].

One of the ways to increase the detection credibility and object recognition is usage of cascade structure of neural network structure. In particular, Cascade R-CNN [32] is composed of several

stages where the previous stage sends the data to the next one with metrics IoU threshold values increase to increase the quality of data processing trainings. Direct combination of Cascade R-CNN and Mask R-CNN provides an insignificant improvement due to mask foresight at further stages that receive higher accuracy of detection and recognition only from more qualitatively localized bounding boxes without direct combination. So, the creation of multi-stage conveyor of aerial image pro-cession that uses the combination of detection, instance segmentation and semantic segmentation for receiving the context, will enable to increase the accuracy of object detection and recognition.

### 3. Method

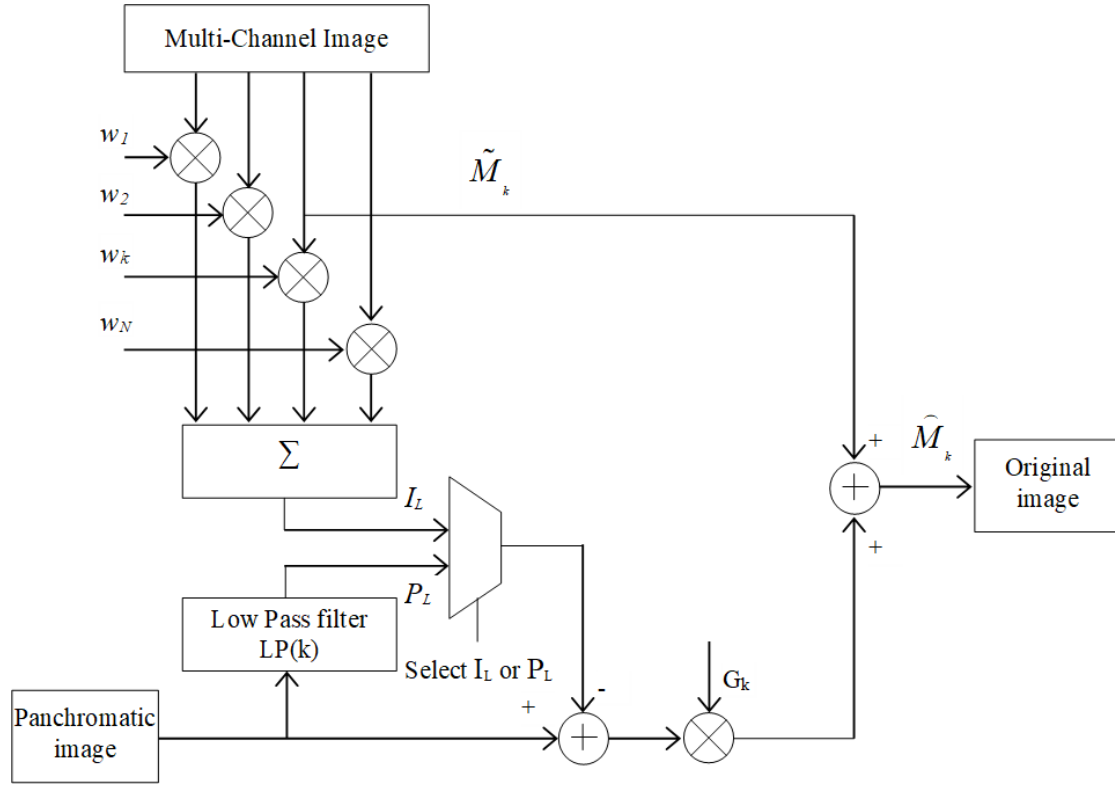
Information from various sources of spatial basing will stipulate complexation of various structural data within onetime interval (during the first day half). So, the data about the same object are received by UAV – aerial image in visible range, and from spaceship – multispectral image. Such approach will enable the connection with spatial and spectral analytical models and in case of the library of spectral etalons availability it may enable to use spectral-spatial (sub-pixel) analysis which should result in automatic ground object identification at the Earth surface (table 1).

It is also important to harmonize the images in one format, so that they had the same resolution. Then, the main principle of data acquisition construction about the object of monitoring may become the optimal effort resolution among the means of various spatial basing sources. In this case it is necessary and sufficient is the task of multi-criteria task solution for choosing sufficient means of intelligence data gathering and sequence of their use determination. The optimization approach stipulates mathematical models use and optimization criterion explicitly. The basis for such task solution is the best alternative search by some criterion. Such approach enables to increase the solution quality through such factors:

- enables to find the variants of task solution at various values of real limits to variables and various initial conditions;
- enables to simplify the best solution selection procedure thanks to using the analytical criteria; several criteria may be used simultaneously;
- presence of multitude of methods of dynamic optimization task solution enables to select the best alternative.

Pansharpening methods synthesize images with the same number of spectral channels as the input multispectral image and resolution as in the input panchromatic image. After interpolation from the multispectral image into the panchromatic space, elements are extracted from it and added to the corresponding bands of the multispectral image using the injection model. Panning is pre-selected with a histogram, that is, radiometrically transformed by constant gain and offset. The injection model defines the combination of the multispectral image low frequency image with the spatial details of the panchromatic image. This approach is applied to each resampled band of the multi-spectral image and the low-frequency version of the panchromatic image. In this approach, the bandwidth of the panchromatic image covers four spectral bands (figure 1). This provides the advantage that the removal of the estimated path radii for the calculation of

the injection model is more consistent in terms of spectral quality (color hues) in relation to spatial characteristics [33]. This approach is the basis for the decision regarding the survey of infrastructure objects in the epicenter of the fire to improve image quality [34].



**Figure 1:** Flowchart of CS/MRA-switchable pansharpening.

Based on the results of the detectors application [35, 36, 37, 38, 39, 40], the problems of the impact of deformation, occlusion, changing image size in the picture and frequent background changes are determined. A promising approach to their solution is the application of a cascade of elemental and semantic segmentation models that use a deep trunk net-work generating sufficient representations of features.

As the model basis CNN ResNeXt [29] in BiFPN [41] is used. ResNeXt is high-module network architecture with great receptive field due to aggressive convolution. BiFPN usage enables to carry out the contribution research of various original feature cards with simultaneous repeated usage of multi-scale synthesis of features “from top downward” and “bottom upwards”. It enables to capture the features from the lower level of highway neural network and as a result it enables to recognize the objects in broader scale range using fewer parameters than augmented CNN. It solves the problem of hardware restrictions that usually exists both for semantic or for instance segmentation and their combined education. At BiFPN pyramid top deforming CNN (DCN) [42] is used that adapts the target function to the object geometric variations at aerial image using dependence that not all pixels inside the receptive layer filed of

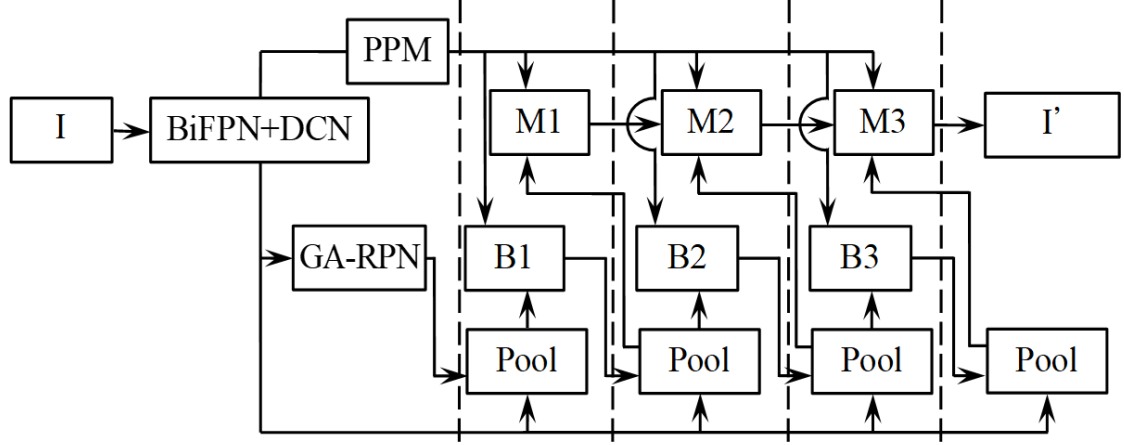
neural network make contribution into the neural network work result. The differences in those contributions are presented by efficient receptive field, which values are calculated as gradient of layer node response to in-tensity of each image pixel disturbance. DCN implementation that broadens the selection spatial placement in CNN additional layers by shifting and shift education, enables to adapt the target function reflection to object configuration as affected by possible transformations, deforming its selection structure and combination that fit the object structure. The suggested approach increases the detection credibility and object recognition at the aerial image.

To increase the credibility of object detection and recognition through object image localization increase at the aerial image and bounding boxes adaptation to the object forms the guided anchorage regional proposal network is used (GA-RPN) [43] used after BiFPN. GA-RPN usage is determined by two factors: the objects at the image are located unevenly, form (object scale and aspect ratio) are close related with its content and location as to the back-ground elements. The neural network placed in the guided anchorage module basis is composed of two branches for prediction of possible location regions and object form and feature adaptation component. The predictive branch determines the probability card that directs at possible objects locations, but the form predictive branch stipulates depending on the object location – aspect ratio. According to the results of both branches the anchor set is generated which predicted location possibilities surpass the given threshold and the most possible forms of each of selected places. As far as the anchor form may change the features in various places have to be captured in various scales. For that feature adaptation module is used additionally that selects the anchor forms according to the feature presentation. Thus, the multilevel anchor generation scheme is applied that enables to form the anchor set of several feature cards taking into account BiFPN architecture. As a result, each object location is related to only one anchor of dynamically predicted form instead of a set of predetermined anchors. The features for the anchor forming are received from the original feature card of BiFPN corresponding level.

Common communication use between the bounding box detection and masks gives limited prize, so their cascade application for improvement of detected object localization and their recognition is more efficient solution. The cascade procedure is applied during the conclusions of each stage that enables to coordinate the hypotheses more accurately. The cascade use enables to decrease the network retraining as a result of exponentially vanishing positive samples and stage conclusion non-conformity for IoU value, for which the detector is optimal, to incoming hypotheses. But there is a rupture in information flow between the branches of cascade various stages that results in mask separation at later stages and gives prize only in better localized bounding boxes [32].

To overcome the rupture between the stages the hybrid task cascade is used for instance segmentation [44]. The key idea is information flow improvement by cascade inclusion and multi-task feature at each stage and usage of spatial context for further object detection and recognition credibility increase. As a result of research the hybrid segmentation cascade model was improved that enables to increase the productivity of the aerial image multi-stage processing, recognize the various plan foreground from overwhelmed background due to spatial context using the semantic segmentation. The model structural scheme is given at fig. 2, where:  $I$  – incoming image,  $Pool$  – feature regional deletion,  $B_t$ ,  $M_t$  – detection of bounding box and mask at stage  $t$ .





**Figure 2:** Improved model of hybrid segmentation cascade.

To detect the objects, the scene context provides useful recommendations for semantic branches combination for receiving the categories and scales. Received from each BiFPN layer feature cards of various levels transform into pyramidal phalanx module PPM [23], that execute the background object semantic segmentation at pixel level that prevents information loss in the context among various scene sub-regions. PPM is used for feature card combination to form their final representation with both local and global information about the context. PPM combines the features from five BiFPN original layers. The highest (semantically strong) level is global combination for receiving a single output vector. Next pyramid level separates the feature card into various sub-regions and forms combined presentation for various locations. To preserve the weights of global features the convolution layer  $1 \times 1$  is used after each BiFPN level. For representing the features of such fragmentation as in the final global pyramid the feature combination from the lower level outputs of BiFPN feature cards the bilinear interpolation is used. The cascade semantic branch encodes the context information from the background regions as a result of foreground object distinction from the flooded background that supplements the bounding box and sample masks. This branch is designated for semantic segmentation of the whole image each pixel forecasting that has completely convolution architecture and trains together with other cascade branches. The semantic segmentation features are addition to the existing features of bounding box and masks at their combination to increase the object detection and recognition credibility.

This approach differs from the existing cascade solutions by regression of bounding box sequence and mask prediction instead of their processing in parallel, inclusion of direct way to augment the information flow between the mask branches, delivery of previous stage peculiarities to the mask, direction for study of more contextual information of additional semantic segmentation branch and its alignment with bounding box and masks branches (figure 2). Using the detector sequence that passed the training with the threshold values increase of IoU metrics to be consistently more selective against the close faulty actuations. The sequence of



information passing among the cascade stages is displayed by the formulae:

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

$$x_t^{mask} = P(x, r_t) + P(S(x), r_t), \quad (2)$$

$$r_t = B_t(x_t^{box}), \quad (3)$$

$$m_t = M_t(F(x_t^{mask}, m_{t-1}^-)), \quad (4)$$

where  $x_t^{box}$ ,  $x_t^{mask}$  – detected by bounding box and feature masks;  $P(x, r_{t-1})$  – align operation RoI Align [14];  $B_t(x_t^{box})$ ,  $M_t(x_t^{mask})$  – 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.

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

$$L = \sum_{t=1}^T \alpha_t (L_{bbox}^t + L_{mask}^t) + L_{seg}, \quad (5)$$

$$L_{bbox}^t(c_i, r_i, l_i, s_i, \hat{c}_t, \hat{r}_t, \hat{l}_t, \hat{s}_t) = L_{csl}(c_t, \hat{c}_t) + L_{reg}(r_t, \hat{r}_t) + \lambda_1 L_{loc}(l_i, \hat{l}_t) + \lambda L_{shape}(s_i, \hat{s}_t), \quad (6)$$

$$L_{mask}^t(m_t, \hat{m}_t) = BCE(m_t, \hat{m}_t), \quad (7)$$

$$L_{seg} = CE(s, \hat{s}), \quad (8)$$

where  $L$  – general loss function;  $L_{bbox}^t$ ,  $L_{mask}^t$  – 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.

While creating the training selections for each class of objects by their images for the new dataset from the aerial images a misbalance of classes arises because of lack of sufficient number of object images. When using the loss function of cross entropy during model training at such datasets the scale ratio goes to zero because confidence in correct class grows. To solve this problem various methods are used as resampling. According to the results of re-researches held this solution offers to modify the focal loss method designated to improve the model training at the original non-balanced data. So, instead of cross entropy loss function:

$$CE(p_t) = -\log(p_t), \quad (9)$$

very often the function of focal loss [45] is used

$$FL(p_t) = -(1 - p_t)\log(p_t), \quad (10)$$

where  $FL$  – focal loss;  $CE$  – loss function of cross entropy;  $p_t$  – probability of credible class;  $\gamma$  – focusing value.

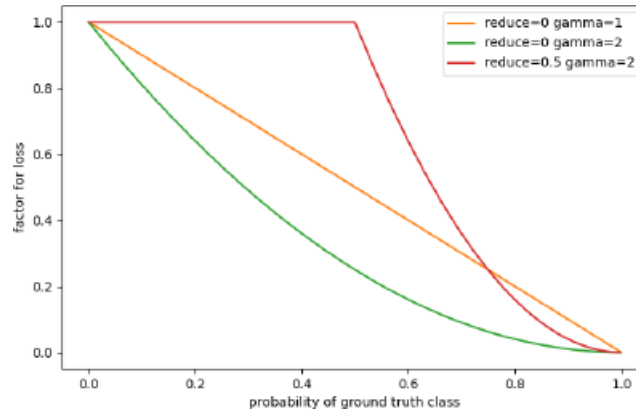
The focal loss minimizes the input of well classified samples and directs the focus at complicated samples. The function of focal loss is elaborated to solve the object determined detection scenario where an extraordinary balance exists between the full and sparse classes. But it does not show better results for two-passage detectors which separate the background at the first stage. It is offered to modify the focal loss function to soften the reaction for the loss functions to complicated samples. Accordingly, the same weights are used for positive samples with probabilities less than certain threshold as well as for minimization of well classified samples influence the focal loss approach is pre-served which scale reflects the threshold. The aforementioned may be described next way:

$$MFL(p_t) = -f(p_t, t_h) \log(p_t), \quad (11)$$

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

$$\begin{cases} 1 & : p_t < t_h \\ \frac{(1-p_t)^\gamma}{t_h^\gamma} & : p_t \geq t_h \end{cases} \quad (12)$$

where  $t_h$  – probability of fundamental truth class.



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

The focal loss modification function helps to improve the average accuracy of object detection mAP for sparse classes, however, mAP is decreased a little for well flooded classes. Function of modified focal loss application decreases the action of class misbalance factor in the process of model training.

## 4. Results

For approbation of improved model of hybrid segmentation cascade and in order to study the process according to the task DataSet with Vehicle Detection in Aerial Images was used. It contained 10 photos at height 1595-1600 m with resolution 5616x3744 pixels. As a result of object distribution 10 classes of transport vehicles were formed. The object class set is not

balances (number of object images in the classes varies from 7 to 2454), transport vehicle images differ much by dimensions, aspect ratio, distribution by brightness and color density.

Online augmentation was used for enlargement of object images taking into account the executing condition of photographing from UAVs (turns to  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$ ,  $270^\circ$ , adding Gaussian noise, contrast, sharpness, color density change). Transfer Learning approach was used through the trained models at COCO Detection dataset.

For the model work assessment metrics mAP was used that calculates mAP average score value for variables IoU to fine a great number of bounding boxes with incorrect classifications and it enables to avoid the maximum specialization in several classes at the account 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 600x600, 700x700 and turn ( $0^\circ$ ,  $90^\circ$ ,  $180^\circ$ ,  $270^\circ$ ), with augmentation to 800x800, 900x900, 1000x1000).

Model training was conducted from the end to the end of 18 epochs. The results obtained are shown in table 2.

**Table 2**

Dependence of mAP value depending on model improvements is applied.

Changes	Modified Hybrid Task Cascade				
DCN	nc	x	x	x	x
GA-RPN	nc	nc	x	x	x
PPM	nc	nc	nc	x	x
MFL	nc	nc	nc	nc	nc
mAP (at $\text{IoU} \geq 0.7$ ), %	63.2	64.6	65.6	65.9	66.2
No change + augmentation - (nc)					

As a result of Hybrid Task Cascade model improvement along with image set growth and post-processing the mAP accuracy was improved by 3%. It enables to increase the small-sized object detection credibility at aerial photos received by UAVs. As far as this approach has a little calculation complexity it enables to implement it on UAV board.

## 5. Conclusions and future work

The offered approach based on the results of existing approaches analysis of atmospheric correction based on injection model application of spatial details based on contrast highlighted from panchromatic image into interpolated multispectral band. For the outgoing image correction to solve the infrastructural object analysis task, cars in the fire epicenter enables to reduce the

atmospheric fog or smoke influence on the quality of incoming image of aerial photo processing systems for sufficient level to actuate the object detector.

As a result of image automatic processing method analysis during neural networks exploration to solve the task of scene analysis, aerial images review the influence of object deformations, occlusions was identified while receiving the aerial image and background change, where the object is located. It reduces the accuracy of object detection and recognition. Based on the review of contemporary neural network models in the framework of the task it was selected Hybrid Task Cascade for Instance Segmentation. It improves the information flow through cascade inclusion and multi-tasking at each stage that uses indirect signs from topographic elements of terrain to increase cred-ibility of object detection and recognition.

Further research should be directed at increasing the possibilities to use the UAVs in the complex conditions of the crisis situation and complex spatial orientation.

## References

- [1] A. L. Miller, Adapting to teaching restrictions during the COVID-19 pandemic in Japanese universities, *Educational Technology Quarterly* 2022 (2022) 251–262. doi:10.55056/etq.21.
- [2] V. Tkachuk, Y. V. Yechkalo, S. Semerikov, M. Kislova, Y. Hladyr, Using Mobile ICT for Online Learning During COVID-19 Lockdown, in: A. Bollin, V. Ermolayev, H. C. Mayr, M. Nikitchenko, A. Spivakovsky, M. V. Tkachuk, V. Yakovyna, G. Zholtkevych (Eds.), *Information and Communication Technologies in Education, Research, and Industrial Applications - 16th International Conference, ICTERI 2020, Kharkiv, Ukraine, October 6-10, 2020, Revised Selected Papers*, volume 1308 of *Communications in Computer and Information Science*, Springer, 2020, pp. 46–67. doi:10.1007/978-3-030-77592-6\_3.
- [3] P. Barnard, L. Erikson, A. Foxgrover, et. al, Dynamic flood modeling essential to assess the coastal impacts of climate change, *Scientific Reports* 9 (2019). doi:10.1038/s41598-019-40742-z.
- [4] V. Alekseev, O. Alekseev, A. Vidmish, *Interactive monitoring of highways*, VNTU, Vinnytsia, 2012.
- [5] B. Aiazzi, S. Baronti, M. Selva, L. Alparone, Enhanced Gram-Schmidt Spectral Sharpening Based on Multivariate Regression of MS and Pan Data, in: 2006 IEEE International Symposium on Geoscience and Remote Sensing, 2006, pp. 3806–3809. doi:10.1109/IGARSS.2006.975.
- [6] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, L. M. Bruce, Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data-Fusion Contest, *IEEE Transactions on Geoscience and Remote Sensing* 45 (2007) 3012–3021. doi:10.1109/TGRS.2007.904923.
- [7] R. A. Schowengerdt, *Remote Sensing: Models and Methods for Image Processing*, 3 ed., Academic Press, 2007. doi:10.1016/B978-0-12-369407-2.X5000-1.
- [8] C. Munechika, J. Warnick, C. Salvaggio, J. Schott, Resolution Enhancement of Multispectral Image Data to Improve Classification Accuracy, *Photogrammetric engineering and remote sensing* 59 (1993) 67–72.

- [9] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, M. Selva, An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas, in: 2003 2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, 2003, pp. 90–94. doi:10.1109/DFUA.2003.1219964.
- [10] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, Sharpening of very high resolution images with spectral distortion minimization, in: IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477), volume 1, 2003, pp. 458–460 vol.1. doi:10.1109/IGARSS.2003.1293808.
- [11] R. Restaino, M. Dalla Mura, G. Vivone, J. Chanussot, Context-Adaptive Pansharpening Based on Image Segmentation, IEEE Transactions on Geoscience and Remote Sensing 55 (2017) 753–766. doi:10.1109/TGRS.2016.2614367.
- [12] G. Vivone, R. Restaino, M. Dalla Mura, G. Licciardi, J. Chanussot, Contrast and Error-Based Fusion Schemes for Multispectral Image Pansharpening, IEEE Geoscience and Remote Sensing Letters 11 (2014) 930–934. doi:10.1109/LGRS.2013.2281996.
- [13] Y. Li, H. Qi, J. Dai, X. Ji, Y. Wei, Fully Convolutional Instance-Aware Semantic Segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4438–4446. doi:10.1109/CVPR.2017.472.
- [14] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, 2018. arXiv:1703.06870.
- [15] J. Tighe, M. Niethammer, S. Lazebnik, Scene Parsing with Object Instances and Occlusion Ordering, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3748–3755. doi:10.1109/CVPR.2014.479.
- [16] Z. Tu, X. Chen, Yuille, Zhu, Image parsing: unifying segmentation, detection, and recognition, in: Proceedings Ninth IEEE International Conference on Computer Vision, 2003, pp. 18–25 vol.1. doi:10.1109/ICCV.2003.1238309.
- [17] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid Scene Parsing Network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2017, pp. 6230–6239. doi:10.1109/CVPR.2017.660.
- [18] M.-M. Cheng, Z. Zhang, W.-Y. Lin, P. Torr, BING: Binarized Normed Gradients for Objectness Estimation at 300fps, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3286–3293. doi:10.1109/CVPR.2014.414.
- [19] J. Yao, S. Fidler, R. Urtasun, Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 702–709. doi:10.1109/CVPR.2012.6247739.
- [20] M. Sun, B.-s. Kim, P. Kohli, S. Savarese, Relating Things and Stuff via ObjectProperty Interactions, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 1370–1383. doi:10.1109/TPAMI.2013.193.
- [21] P. Dollár, Z. Tu, P. Perona, S. Belongie, Integral channel features, in: British Machine Vision Conference, London, 2009. URL: [https://pages.ucsd.edu/~ztu/publication/dollarBMVC09ChnFtrs\\_0.pdf](https://pages.ucsd.edu/~ztu/publication/dollarBMVC09ChnFtrs_0.pdf).
- [22] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS’15, MIT Press, Cambridge, MA, USA, 2015, p. 91–99. doi:10.5555/2969239.2969250.
- [23] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path Aggregation Network for Instance Segmentation,

- in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768. doi:10.1109/CVPR.2018.00913.
- [24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature Pyramid Networks for Object Detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944. doi:10.1109/CVPR.2017.106.
- [25] J. Dai, K. He, Y. Li, S. Ren, J. Sun, Instance-Sensitive Fully Convolutional Networks, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, volume 9910 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2016, pp. 534–549. doi:10.1007/978-3-319-46466-4\_32.
- [26] R. Girshick, Fast R-CNN, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448. doi:10.1109/ICCV.2015.169.
- [27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, SSD: Single Shot MultiBox Detector, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, volume 9905 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2016, pp. 21–37. doi:10.1007/978-3-319-46448-0\_2.
- [28] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2016, pp. 779–788. doi:10.1109/CVPR.2016.91.
- [29] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated Residual Transformations for Deep Neural Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5987–5995. doi:10.1109/CVPR.2017.634.
- [30] S. Gidaris, N. Komodakis, Attend Refine Repeat: Active Box Proposal Generation via In-Out Localization, in: British Machine Vision Conference, York, 2016. doi:10.48550/arXiv.1606.04446.
- [31] B. Yang, J. Yan, Z. Lei, S. Z. Li, CRAFT Objects from Images, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 6043–6051. doi:10.1109/CVPR.2016.650.
- [32] Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving Into High Quality Object Detection, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162. doi:10.1109/CVPR.2018.00644.
- [33] S. Lolli, L. Alparone, A. Garzelli, G. Vivone, Benefits of haze removal for modulation-based pansharpening, in: L. Bruzzone (Ed.), Image and Signal Processing for Remote Sensing XXIII, volume 10427, International Society for Optics and Photonics, SPIE, 2017, p. 1042707. doi:10.1117/12.2279086.
- [34] S. Kovbasiuk, L. Kanevskyy, I. Sashchuk, M. Romanchuk, Object Detection Method Based on Aerial Image Instance Segmentation in Poor Optical Conditions for Integration of Data into an Infocommunication System, in: 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), 2019, pp. 224–228. doi:10.1109/PICST47496.2019.9061496.
- [35] N. Tijtgat, W. Van Ranst, B. Volckaert, T. Goedemé, F. De Turck, Embedded Real-Time Object Detection for a UAV Warning System, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 2110–2118. doi:10.1109/ICCVW.2017.247.

- [36] L. W. Sommer, T. Schuchert, J. Beyerer, Fast Deep Vehicle Detection in Aerial Images, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 311–319. doi:10.1109/WACV.2017.41.
- [37] P. Chen, Y. Dang, R. Liang, W. Zhu, X. He, Real-Time Object Tracking on a Drone With Multi-Inertial Sensing Data, volume 19, 2018, pp. 131–139. doi:10.1109/TITS.2017.2750091.
- [38] M. Hsieh, Y. Lin, W. H. Hsu, Drone-Based Object Counting by Spatially Regularized Regional Proposal Network, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2017, pp. 4165–4173. doi:10.1109/ICCV.2017.446.
- [39] Y. Tang, C. Zhang, R. Gu, P. Li, B. Yang, Vehicle detection and recognition for intelligent traffic surveillance system, *Multimedia Tools and Applications* 76 (2017) 5817–5832. doi:10.1007/s11042-015-2520-x.
- [40] X. Wen, L. Shao, W. Fang, Y. Xue, Efficient Feature Selection and Classification for Vehicle Detection, *IEEE Transactions on Circuits and Systems for Video Technology* 25 (2015) 508–517. doi:10.1109/TCSVT.2014.2358031.
- [41] M. Tan, R. Pang, Q. V. Le, EfficientDet: Scalable and Efficient Object Detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2020, pp. 10778–10787. doi:10.1109/CVPR42600.2020.01079.
- [42] J. Wang, K. Chen, S. Yang, C. Loy, D. Lin, Region Proposal by Guided Anchoring, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2019, pp. 2960–2969. doi:10.1109/CVPR.2019.00308.
- [43] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable Convolutional Networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 764–773. doi:10.1109/ICCV.2017.89.
- [44] K. Chen, W. Ouyang, C. Loy, D. Lin, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, Hybrid Task Cascade for Instance Segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2019, pp. 4969–4978. doi:10.1109/CVPR.2019.00511.
- [45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2020) 318–327. doi:10.1109/TPAMI.2018.2858826.