Object Detection Method Based on Aerial Image Instance Segmentation in Poor Optical Conditions for Integration of Data into an Infocommunication System

Serhiy Kovbasiuk, Leonid Kanevskyy, Ihor Sashchuk, Mykola Romanchuk
Research Department
Zhytomyr Military Institute named after S. P. Korolov
Zhytomyr, Ukraine
klasik552008@gmail.com, leo10k10@ukr.net, sashchuk_im@ukr.net, romannik@ukr.net

***Abstract*— The article analyses the possibilities to use the unmanned aerial complexes in the system of decision making process in crisis situations that require the object detection at aerial images received by the unmanned aerial complexes under the conditions of atmospheric fog. The Pansharpening method was used for image correction to inject spatial details from panchromatic image to multidimensional image. In order to increase the operational efficiency and accuracy of automotive vehicles detection at aerial images received by the unmanned aerial complexes for more efficient use of received information in the system of decision making support it was selected the Cascade Mask R-CNN model. This model is more suitable for task solution of multiclass classification and small-sized object detection at the image. To improve this model it is suggested using the small-sized anchors making into account the aspect ratio to more classes, function focal loss for model training that along with test time augmentation use enabled to increase mean Average Precision (mAP).**

Keywords—object detection, aerial photo-images, instance segmentation, focal loss

I. INTRODUCTION

For provision of society activity security during the recent decade the use of unmanned aerial complexes (UAC) has been actively developed. For example, next areas for UAC use may be outlined: monitoring of objects hazardous for humans, information support during fires at hazardous industry, chemical and biological treatment of agricultural objects. A special attention is paid to relief operations and prohibition of accident propagation in crisis situations (accidents, emergencies, fires, natural disasters presenting a life or health threat to population or harming the property).

The main problem for crisis situation solution in many cases for any management objects is insufficient volume of information for making the best decisions for relief of consequences [1, 2]. And low capability to take into account all the crisis situation factors reduces the efficiency of preventive measures or rescue operations. Accordingly, the key indices during the crisis situation relief are operational efficiency and evaluation objectivity of hard-to-reach territories where UACs with automatic means of detection, recognition and classification of ground objects become more and more prospective.

II. PROBLEM STATEMENT IN GENERAL TERMS

One of the main components affecting the operational efficiency and quality of decision making in crisis situations is visualization system (information presentation) and such information processing technologies. The relevance of UACs use for such task solution is stipulated by the capabilities of information acquisition concerning the objects on the Earth and the very Earth as underlying terrain. Many studies have been conducted on the processing of images of UACs received: object detection [3-7], action detection [8], visual object tracking [9], object counting [10] and road mining [11]. Besides, the perspective uses of UACs are monitoring improvement of simple and complex objects condition that happened to be in the thick of events, operational efficiency improvement of crisis situation development forecasting and optimization of management decision making.

The modern visualization systems enable to produce huge volumes of information from various sources including UACs. Basically, such information does not contain the monitoring intermediate conclusions complicating scenario forecasting and management decision making. To solve that problem in automated mode the acquired information processing is additionally performed. One of such examples is quantity and quality analysis of infrastructure assets, cars at the fire epicenter to plan the fire relief measures and large special purpose machines maneuverability forecasting, presence of cars in the parking places in the crisis event epicenter where a requirement emerges to remove the cars or create corresponding passage.

Solution of that problem using UACs requires searching and development of efficient (expedite with sufficient accuracy) method of small objects detection at aerial images received by the UACs.

The purpose of the article model analysis of neural networks as a tool to increase the operational efficiency and credibility of small-sized object detection at aerial images received by the UACs with further refinement to increase localization and recognition under the condition of low visibility, search for such approaches that would enable to increase the efficiency of detection method usage during the fire response.

978-1-7281-4184-8/19/$31.00 ©2019 IEEE

III. Review of the Latest Researches and Publications

The system of object detection implementation at aerial images requires solving two main tasks. First, it is necessary to solve the problem of recognition – distinguish the forefront objects from the background and assign them corresponding marks of object class. Second, the detector should solve the problem of localization – assign the accurate bounding boxes for various objects.

In emergency situations the main factor that undermines the object detection at aerial images is the atmospheric fog or smoke during fires. The best practice analysis shows that the main approach is Pansharpening method of using the models of spatial detail injection from panchromatic image to multispectral image for the source image correction. However, the most popular injection models are: projecting model that may be removed from Gram-Schmidt orthogolization being the basics of spectral escalation [12] and context-based solution [13]; multiplicative or contrast model; model based on modulation being the basis of such methods as high-frequency modulation [14], synthetic transient coefficients [15] and model of spectral distortion minimization [16, 17].

As against the projecting model that may be global as for Gram-Schmidt orthogolization, or local as for context-based solution – the model based on contrast ratio is local by its essence, or context-adaptive [18] because the injection reinforcement changes in each pixel [19].

Detectors used to detect objects in aerial photographs are divided into two types. The first type is used to obtain features in pre-processing histograms of oriented gradients (HOG) [20], internal channel features (ICF) [21], aggregated channel features (ACF) [22], deformable parts model (DPM) [23] and for object recognition and localization classic machine learning algorithms such as decision trees (ACF, ICF) or support vector machines (SVM) (HOG, DPM). The second type includes detectors that use deep neural networks.

Based on the analysis of aerial image processing that passed correction those had advantages that technologically were arranged for using the deep neural networks. It is explained by their properties as for adaptive training, generalization, possibilities of making calculations in real-time mode and resilience to failures. For the task solution the neural networks architectures were analyzed that have been used in Object Detection tasks. Thus, single-pass detectors single shot multibox detector (SSD) [24] and You only look once (YOLO) [25] bypass in speed of operation the family of two-pass models regions with convolution neural network (CNN) features (R-CNN) [26-29]. But SSD using networks visual geometry group (VGG) [30] optimized such way that the fine layers in the neural network do not generate sufficiently high functions to detect the small objects. YOLOv1 uses the layers for alignment that also reduces the possibility to detect the small objects. In the next versions of YOLO as a result of architecture change the small object detection accuracy is still lower than of Faster R-CNN [28]. The main drawback of R-CNN, Fast R-CNN is low speed of image processing: Faster R-CNN demonstrates good results during detection but as all the previous models uses the bounding boxes during image visualization of detected objects that may lead to multiple coverage of small object at close dispositions. One of the approaches to solve the detection tasks is instance segmentation. A lot of methods create the conveyors that use CNN including DeepMask [31], SharpMask [32] and InstanceFCN [33]. Multi-task Network Cascades [34] uses sample segmentation as conveyor which consists of three sub-tasks: sample localization, mask forecasting and object categorization, and it trains the whole network by cascade sequence. InstanceFCN implementation is usage of completely convolution approach for instance segmentation. Mask R-CNN [28] adds an additional branch founded at Faster R-CNN to receive the mask forecasts for pixel level. PANet [35] uses double-sized information flow in Feature pyramid network (FPN) [36]. Other methods use at first the binary segmentation to receive the segmentation card of pixel level from the image and then identify the objects. It is necessary to emphasize Zhang et al. work [37] which stipulates the sample marking based on local corrections and local result integration from Markov random field. Deep Watershed Transform method [38] uses the procedure that creates an energy map after segmentation and then distinguished the samples by watershed divide.

IV. Core Material Summary

Pansharpening methods use the advantages of additional spatial and spectral resolutions of multispectral and panchromatic data for image synthesis that has as many spectral bands as the input multispectral image with the same spatial permission as well as panchromatic image. After through interpolation based on multispectral image the panchromatic image was formed the spatial details are stretched and added to multispectral image bands according to certain injection model. The detail extraction stage may resemble the spectral approach initially know as component substitution or spatial approach that may refer to multilevel analysis.

Panoraming by preliminary selected histogram that is radiometrically transformed through permanent reinforcement and shifting. The injection model determines the combination of lower frequencies image of multispectral image with spatial details received from panchromatic image. Such model is described between each of additionally selected bands of multispectral image and version of lower frequencies of panchromatic image. Such approach is shown at Fig. 1, where the panchromatic image bandwidth embraces four spectral bands, demonstrates that advantages of assessed radiuses for injection model assessment route removal are more consistent in the context of spectral quality (color shades) concerning the spatial characteristics [39]. Such approach – taken as a basis for the solution of infrastructural objects analysis, cars in the fire ground zero – enables to reduce the atmospheric fog influence or smoke from the fire at the quality of input image of aerial photo images processing systems.

Based on researches and findings [12-19] it was found that using the aforementioned models applied in the tasks of Object Detection next problematic issues arise: object images are often subjects to deformation, occlusion, rescaling and often background change that reduces the level of object localization. To solve these issues the representation is required which would be simultaneously

steady to external view changes but would not lose the context information of object recognition at the overwhelmed background. The solution of such dilemma may be application of various models compilation method. Another prospective direction at the moment may be the model cascade application for object localization increase.

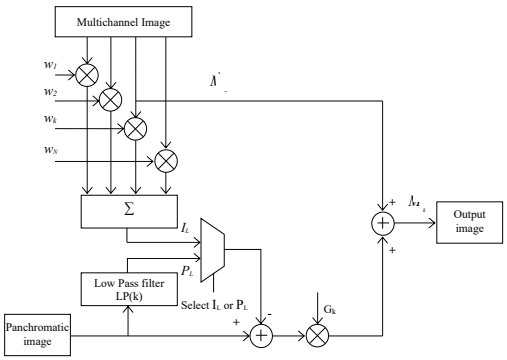

Fig. 1. Flowchart of CS/MRA-switchable Pansharpening.

Cascade Mask R-CNN [40] (Fig. 2) consists of several stages where output of each stage goes to the next one for the improvement of processing quality. Moreover, these trainings at each stage are selected with threshold values increase that by its essence handles various study divisions. First, the tasks are united at each stage including detection, mask forecasting creating such way a common multi-stage processing conveyor. As a result the adjustment at each stage benefits from the relationship among these tasks.

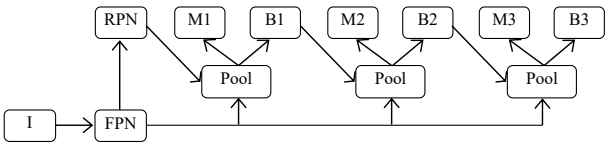

Fig. 2. Model Cascade Mask R-CNN.

The main idea for using the detector sequence that passed the training with the threshold values improvement intersection over union (IoU) to be consistently more selective against the close faulty actuations. The detectors are prepared by stages using surveillance that the detector output is a good distribution for the training of the next higher detector quality. The sequence of information passing among the stages is displayed by the formulae:

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

$$x_t^{mask} = P(x, r_{t-1}), \quad (2)$$

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

$$m_t = M_t(x_t^{mask}), \quad (4)$$

where x_t^{box}, x_t^{mask} are detected bounding box and sample masks: $P(x, r_{t-1})$ is align operation with region of interest (RoI Align): $B_t(x_t^{box})$, $M_t(x_t^{mask})$ are definition of bounding box and mask at stage *t*: r_t, m_t are anticipation of bounding boxes and sample masks.

Resampling of gradually improved hypotheses guarantees that all the detectors have a positive set of equivalent size examples that reduces the problem of retraining. The same cascade procedure is applied during the conclusions enabling more accurately to coordinate the hypotheses and detector quality of each stage.

As against FPN the Mask R-CNN head that obtains various peculiarities from the sample pyramid for various scales of this model for small-sized objects detection uses this function only from the lowest level. To overcome the large-scaled deviations the proposals received from the first stage are divided into several groups according to their scale and uniform sampling. 2048 proposals are selected for each group and they are used for the second stage. The model main architecture is registered residual network ResNeXt-152 that pulls the samples from the convolution layer of the first stage C1. For using of training transfer method layer C4, which weights were changed. For uniting of all the detected objects to exclude their repeated detection the post-processing algorithm of soft non-maximum suppression (soft-NMS) was applied.

While creating the training samples for each class of objects based on the object images for a new dataset from the aerial photo images a problem of class non-balance arises. Non-balanced data represent a substantial problem for machine training models. Various methods are used to resist this problem, such as repeated sampling and technique of selective sampling of synthetic minorities. This solution offers the upgraded method of focal loss designated for training process improvement at original non-balanced data. For instance, instead of cross entropy

$$CE(p_t) = -\log(p_t), \quad (5)$$

quite often the function of focal loss [41] is used of next type

$$FL(p_t) = -(1-p_t)^{\gamma} \log(p_t), \quad (6)$$

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

The focal loss minimizes the input of well classified samples and directs the focus at complicated samples. The function of focused loss is a soft approach to difficult examples acquiring. But it does not show good results for double-pass detectors. 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 preserved which scale reflects the threshold. The aforementioned may be described next way:

$$MFL(p_t) = -f(p_t, t_h)\log(p_t) \quad (7)$$

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

$$f(x) = \begin{cases} 1 & : p_t < t_h \\ \dfrac{(1-p_t)^{\gamma}}{t_h^{\gamma}} & : p_t \geq t_h \end{cases}. \quad (8)$$

The focal loss modification helps to improve the general model productivity. Such approach enables to improve the average accuracy of object detection mAP for rare classes using the model prepared thoroughly, however, it slightly decreases mAP for widespread classes.

V. Experimental Results

To research the suggested approaches and for the qualitative process modeling according to the assigned task DataSet with Vehicle Detection in Aerial Images was used that contained 10 photos made by Canon EOS-1Ds Mark III camera (focal distance 51 mm) at height 1595-1600 m with resolution 5616 x 3744 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.

The input image was magnified two times for object masks augmentation leading to better quality of small objects detection at the images. Online augmentation was used for enlargement of object images (D4 turns, adding Gaussian noise, contrast change, sharpness, color density). Transfer Learning approach was used through the trained models at COCO [42], ImageNet [43] datasets.

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.

For the areas of interest transfer the input of registered parameter network, where the convolution filter is used sized 7 x 7, use the anchors with sizes [4, 8, 16, 32, 64] that promote their better intersection and also takes into account the object aspect ratio using such anchor value ratio [0,33; 0,5; 1,0; 2,0; 3,0]. To take into account the place of the image location at the photo (minimization of object influence with more distorted signs at the image edge) the probability reduction of 0.8 for the correction coefficient was used. While finishing the model work we apply 3 test time augmentation (TTA) (negligible sampling of images 600x650, 700x750 and turn (0, 90, 180, 270), 800x850 to 1000x1000). The soft-NMS post-processing algorithm with a value of 0.7 is used to combine the detected detection objects. MFL was applied to prevent retraining due to class imbalance.

Model training was conducted from the end to the end of the 12 epochs. The received results are represented in table 1.

As a result of model Cascade Mask R-CNN setting along with image set growth and post-processing the mAP accuracy was improved at 7 %. It enables to increase the small-sized object detection credibility at aerial photos received by UACs. As far as this approach has a little calculation complexity it enables to implement it on UAC board improving the operational efficiency of received information in the decision making system.

TABLE I. Dependence of accuracy of MAP from change of hypermeters

Changes	**Cascade Mask R-CNN**						
anchor size change	No change + augmentation (nc)	x	x	x	x	x	x
value of Anchor ratio		nc	x	x	x	x	x
3 TTA		nc	nc	x	x	x	x
Soft-NMS=0.7		nc	nc	nc	x	x	x
Probability reduction at the edge of photo		nc	nc	nc	nc	x	x
MFL		nc	nc	nc	nc	nc	x
mAP (at IoU>=0.7), %	61.2	62.2	64.2	65.5	67	67.4	68.2

VI. Conclusions and Further Research

Outlined 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 output 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 input image of aerial photo processing systems for sufficient level to actuate the object detector. Based on the review of the contemporary neural networks in the framework of assigned task the best one was selected with corresponding architecture. Based on the researches it was suggested performing the dataset augmentation taking into account the conditions during object shooting, set the model hyper parameters to detect the transport vehicles and additionally to perform the post-processing that enabled to increase the object localization and classification accuracy. The direction for further research is focused on the improvement of small objects detection task on the images.

Further research should be directed at examination of FPN application in the main architecture Mask R-CNN and usage of additional semantic segmentation in Cascade R-CNN, Z-score normalization and model assembly usage for capability enhancement in the decision making support system. Besides, it is necessary to perform the research of capability enhancement of UACs application in the rough conditions of crisis situation and complicated spatial orientation.

References

[1] New US Geological Survey-led research helps California coastal managers prioritize planning and mitigation efforts due to rising seas and storms. – Available at https://www.preventionweb.net/news/view/64251.html.

[2] V.O. Alekseev, O.P. Alekseev, A.A. Vidmish, and V.O. Khabarov *Interactive monitoring of highways: monograph*, Vinnitsa: VNTU, 2012.

[3] J. Zhong, T. Lei, and G. Yao, "Robust vehicle detection in aerial images based on cascaded convolutional neural networks," in *Sensors 17(12)*, 2017, p. 2720.

[4] F. Li, S. Li, C. Zhu, X. Lan, and H. Chang, "Cost-effective class-imbalance aware CNN for vehicle localization and categorization in high resolution aerial images," in *Remote Sensing 9(5)*, 2017, p. 494.

[5] N. Tijtgat, W. Van Ranst, B. Volckaert, T. Goedeme, and F. De Turck, "Embedded real-time object detection for a UAV warning

system," in *ICCV2017, the International Conference on Computer Vision*, 2017, pp. 2110-2118.

[6] L. Sommer, T. Schuchert, and J. Beyerer, "Fast deep vehicle detection in aerial images," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conferenceon,* pp. 311-319.

[7] T.D. Stek, "Drones over mediterranean landscapes. the potential of small UAV's (drones) for site detection and heritage management in archaeological survey projects," in *Journal of Cultural Heritage,* vol. 22, 2016, pp. 1066-1071.

[8] M. Barekatain, M. Marti, H. Shih, S. Murray, K. Nakayama, and Y. Matsuo "Okutama-action: An aerial view video dataset for concurrent human action detection," in *1st Joint BMTT-PETS Workshop on Tracking and Surveillance, CVPR*, 2017, pp. 1-8.

[9] P. Chen, Y. Dang, R. Liang, W. Zhu, and X. He, "Real-time object tracking on a drone with multi-inertial sensing data," in *IEEE Transactions on Intelligent Transportation Systems 19 (1)*, 2018, pp. 131-139.

[10] M. Hsieh, Y. Lin, and W. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *The IEEE International Conference on Computer Vision (ICCV)*, vol. 1, 2017.

[11] K. Kanistras, G. Martins, M. Rutherford, and K. Valavanis, "A survey of unmanned aerial vehicles for traffic monitoring," in *Unmanned Aircraft Systems (ICUAS), 2013 International Conference on IEEE*, pp. 221-234.

[12] B. Aiazzi, S. Baronti, and M. Selva, "Enhanced Gram-Schmidt spectral sharpening based on multivariate regression of MS and Pan data," in *Proc. IGARSS*, 2006, pp. 3806-3809.

[13] L. Alparone, L.Wald, J. Chanussot, C. Thomas, P. Gamba, and L. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data fusion contest," in *IEEE Trans. Geosci. Remote Sens*, vol. 45, 2007, pp. 3012-3021.

[14] R. Schowengerdt, "Remote Sensing: Models and Methods for Image Processing," in *Academic Press*, Orlando, FL, USA, 2nd ed, 1997.

[15] C. Munechika, J. Warnick, C. Salvaggio, and J. Schott, "Resolution enhancement of multispectral image data to improve classification accuracy," in *Photogramm. Eng. Remote Sens*, vol. 59, 1993, pp. 67-72.

[16] B.Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M.Selva, "An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas," in *Proc. 2nd GRSS/ISPRS Joint Workshop Remote Sens. Data Fusion URBAN Areas*, 2003, pp. 90-94.

[17] L.Alparone, B. Aiazzi, S.Baronti, and A. Garzelli, "Sharpening of very high resolution images with spectral distortion minimization," in *Proc. IEEE IGARSS*, 2003.

[18] R. Restaino, M. Dalla, G. Vivone, and J. Chanussot, "Context-adaptive pansharpening based on image segmentation," in *IEEE Trans. Geosci. Remote Sens*, vol. 55, 2017, pp. 753-766.

[19] G. Vivone, R. Restaino, M. Dalla, G. Licciardi and J. Chanussot, "Contrast and error-based fusion schemes for multispectral image pansharpening," in *IEEE Geosci. Remote Sens. Lett*, vol. 11, 2014, pp. 930-934.

[20] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, IEEE Computer Society Conferenceon*, vol. 1, 2005, pp. 886–893.

[21] P. Doll´ar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. of British Machine Vision Conference*, 2009.

[22] P. Doll´ar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[23] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, IEEE Conferenceon*, 2008, pp. 1–8.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single shot multibox detector," in *ECCV*, 2016.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *arXiv:1506.02640*, 2015.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[27] R. Girshick, "Fast R-CNN," in *ICCV*, 2015.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[29] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick, "Mask R-CNN," *arXiv:1703.06870*, 2017.

[30] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[31] P. Pinheiro, R. Collobert, and P. Dollar, "Learning to segment object candidates," in *Advances in Neural Information Processing Systems*, 2015.

[32] P. Pinheiro, T. Lin, R. Collobert, and P. Dollar, "Learning to refine object segments," in *European Conference on Computer Vision*, 2016.

[33] J. Dai, K. He, Yi Li, S. Ren, and J. Sun, "Instance-sensitive fully convolutional networks," in *European Conference on Computer Vision*, 2016.

[34] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[35] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[36] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[37] Z. Zhang, A. Schwing, S. Fidler, and R. Urtasun, "Monocular object instance segmentation and depth ordering with CNNs," in *IEEE International Conference on Computer Vision*, 2015.

[38] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[39] S. Lolli, L. Alparone, A. Garzelli, and G. Vivone, "Benefits of haze removal for modulation-based pansharpening," in *Image and Signal Processing for Remote Sensing XXIII. Proc. of SPIE*, vol. 10427, 2017.

[40] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[42] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C.L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV. Springer*, 2014, pp. 740–755.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L.Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge,"in *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.