High Spatial Image Classification Problem: Review of Approaches.

Fedor Krasnov, Alexander Butorin

***Abstract*—Leading experts around the world analyzed geophysical images daily by and with the development of computer vision technologies, attempts should be made to automate this process. Image data can be acquired quickly using consumer digital cameras, or potentially using more advanced systems, such as satellite imagery, sonar systems, and drones and aerial vehicles. The authors of this article have developed several approaches to the automatic creation of seismic images. The amount of obtained images became enough to use algorithms of machine learning for their processing. In the last five years, computer vision techniques have evolved at a high rate and have advanced far from the use of Deep Neural Networks (DNN). It would be reckless to use in work only the latest developments without understanding how they appeared. Therefore, the authors reviewed the approaches of computer vision to determine the most appropriate techniques for processing high spatial images that differ from the most popular tasks of computer vision (face recognition, detection of pedestrians on the street, etc.). The main result of the paper is the set of research hypothesis for computer vision in Geoscience.**

***Keywords*— CV, seismic image, RGB blending.**

I. INTRODUCTION TO IMAGE CLASSIFICATION

The classification of images refers to the assignment of an image to certain categories (classes) defined in advance.

Classification of images is divided according to the type of classification problem into binary and multi-class. In the case of binary classification, the problem of assigning an image to a single class is solved and the answer can only be “yes” (1) or “no” (0). With multi-class classification, the answer is the presence or absence of each of the pre-selected classes on the image.

An example of a binary classification is the answer to the question: “Is there a channel in the image or not?” (Fig. 1).

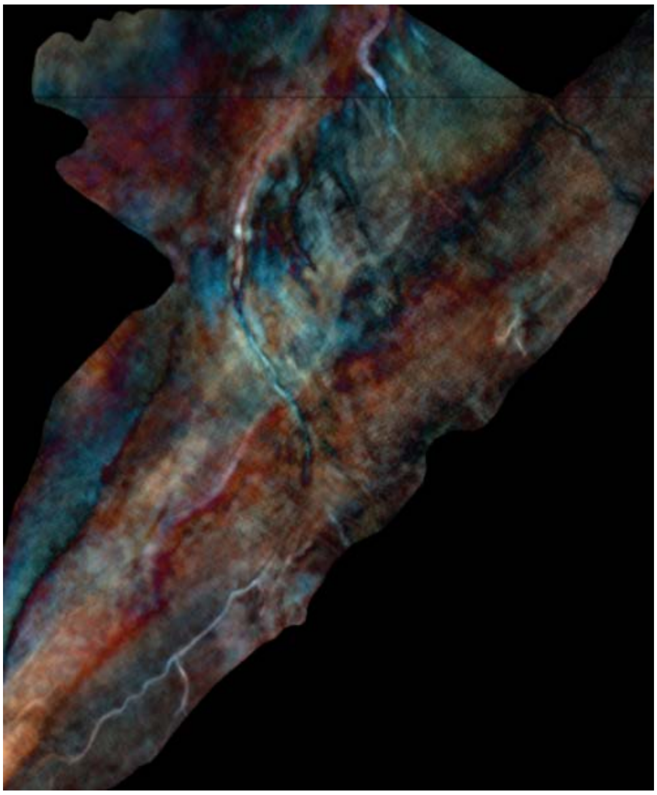

Fig.1. An Example of seismic image.

An example of a multi-class classification of an image may be the answer to the question: “What are the steps of the channel meandering process from the given dictionary presented in the image?” (Fig. 1).

To create the process of automatic classification of images, it is necessary to select features from the image.

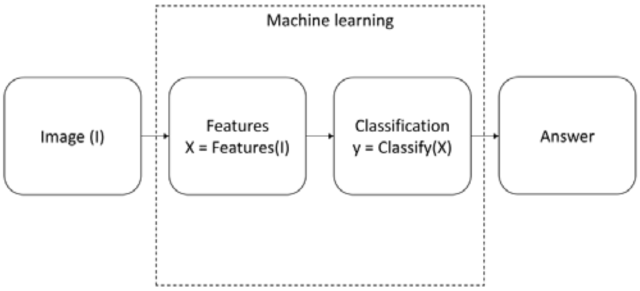

Fig. 2. The scheme of classification process.

The diagram (Fig. 2) shows the process of image classification and the place of machine learning in this process. Unlike computer algorithms, people can recognize images with poor quality. For machine learning, it is

Manuscript received December 22, 2018.

Fedor Krasnov, LLC “Gazpromneft STC”, 190000 Saint-Petersburg, emb. R. Moyka, blds 75-79, (email: krasnov.fv@gazprom-neft.ru), orcid.org/0000-0002-9881-7371, SPIN 8650-1127
Alexander Butorin LLC “Gazpromneft STC”, 190000 Saint-Petersburg, emb. R. Moyka, blds 75-79, (email: butorin.av@gazpromneft-ntc.ru).

necessary to make manual labeling of images. Trained feature selection and classification functions should work with new images in the future.

II. Fine-grained classification

From the classification task, we can pick the sub-task of fine-grained classification. The fine-grained classification pays much more attention to details. Usually, separate intra-class and inter-class fine-grained classification. The fine-grained classification is intended for very similar objects according to [1].

For problems of the fine-grained classification, several classifiers are combined into an ensemble with subsequent collegial decision making. As, for example, in the work [2] proposed a model that identifies the shape of the bird's beak and the shape of the paws for further conjugation when referring to a particular subclass.

Thus, the selected features of different scales are combined into a single matrix of characteristics for classification. In the example above, it is shown that the methods for selecting features can be accurately tuned to certain parts of the object in the image. Such approach allows us to break the complex problem of the fine-grained classification into two more simple subtasks of multi-class classification.

III. Defining object attributes

Attributes are descriptive aspects of the object on the image. Concerning human faces in the images, attributes are age, sex, race, emotions, etc. The task of selecting the attributes of an object preceded by the work of finding this object on the image and selecting the boundaries of the object (Fig.3).

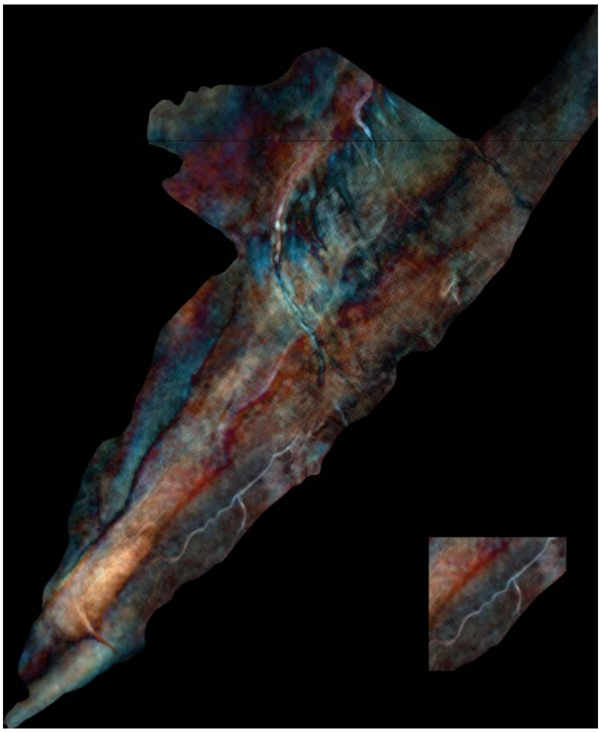

Fig. 3. Original image (on the left) and the object selected for the attribution (on the right).

For the object found, the task of multi-classification of the object's attributes from a given dictionary can be solved both as a global problem and as a task of the fine-grained classification of parts of an object (locally).

As noted in the study [3] for a global approach to solving the task of attribution, overfitting is a particular problem. Therefore, with a global approach, regularization, complex loss functions, and image augmentation are intensively used. On the other hand, with a local approach, a separate mechanism is usually used to select objects, and a set of classifiers trained for those type of objects [4]. Global and local approaches to the attribution of objects on images differ in the way of feature extraction.

IV. Identifying the key points of an object

Let us consider in more detail the problem of identifying the key points of an object on the example of human faces. The mathematical statement of this problem is made in the work [5] as a regression problem. Individual parts of the face (a nose, a mouth, an eye) can be selected with the help of a local approach, and then regression is performed to find the coordinates of the key points of these facial parts (X, Y). On the other hand, a global approach to finding the key points of an object also has the right to exist. In this case, based on the selected features, the regression problem is immediately solved for finding the coordinates of the key points. In work [6] the analysis of various models for allocation of the key points on the face is made. Among these models, SIFT [7] methods are used for accuracy and productivity, using statistics of the physical structure of the face, and models based on HOG descriptors [8].

SIFT and HOG methods are used as detectors for the features by which the classification of images is based.

The key points of objects in the tasks of classifying objects on seismic images can be the shape and curvature of river beds (Fig. 4).

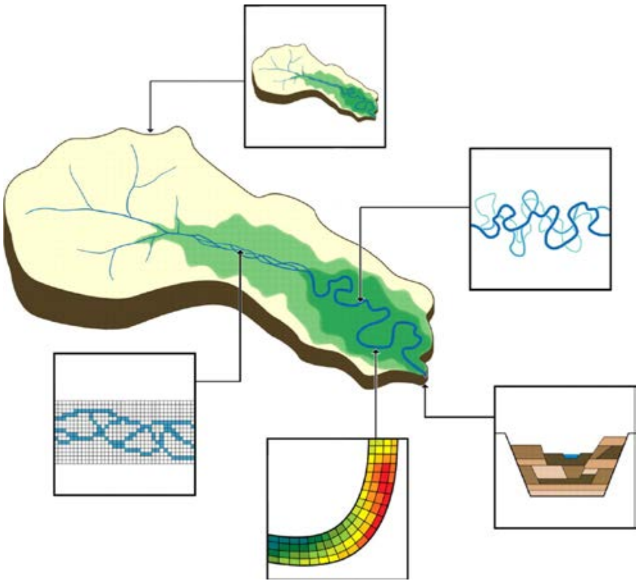

Fig. 4. An example of the key points of the seismic image.

V. Sliding window detectors

The task of detecting objects in an image can be divided into two subtasks: the detection of objects of a specific shape and size (man, cow, bicycle, etc.) and the discovery of regions of textures of uncertain size and shape (grass, clouds, road, etc.). In both cases, the output of the object detector will be a data structure from the predefined dictionary containing the coordinates of the object's bounding box in the image, the class of the object, and possibly the probability of belonging to this class when the classification is "weak".

In the study [9] in particular, the criteria for object detection are discussed. The object frame detected by the detector may differ from the correct object frame. The IoU metric (Intersection over Union) is designed to quantify the accuracy of the detector and is calculated as the ratio of the intersection of the frames to the area of the union of the two frames. The closer the metric IoU to one, the more accurately the object is detected.

According to work, [10], the accuracy of the automated pedestrian position prediction does not exceed the possibility of human evaluation, even with the use of modern architectures such as R-CNN [11].

Manually labeled images datasets ImageNet [12], Caltech-USA [13] and KITTI [14] are used to fit detectors.

The central principle for detectors with a sliding window is to select the window size, search through all window positions on the image and binary classification of the object for each window position. After this, a conclusion is made about the position of the object in the original image. Immediately one can see several problems in this approach - such as the object can be of different sizes, with different width and height ratios, objects can intersect, and different frames can contain the same object. Modern architectures of object detectors are looking for ways to solve the listed problems. For example, you can create a set of multi-scale image options for constant window size.

VI. Detectors of objects based on histograms of oriented gradients

Histograms of oriented gradients (HOG) was proposed in research [8]. The main idea of the HOG algorithm is that the image is divided into a grid and the direction of the color gradient is calculated in each cell. In general, the HOG-based detector algorithm is shown in the figure (Fig. 5).

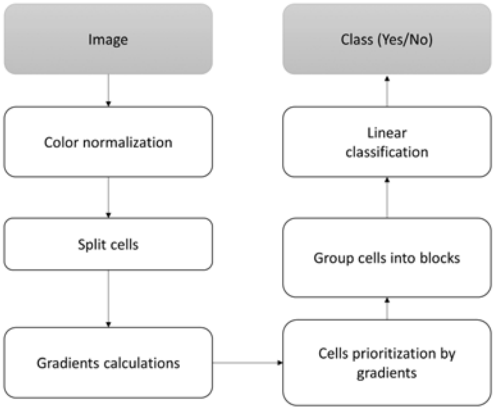

Fig. 5. The HOG algorithm

The task of fitting the object detector is asymmetric: the number of the objects in the image are much smaller than the "no objects". This imbalance is not acceptable for classification problems in machine learning. Also, the "non-object" is a reasonably complex class, so it is necessary to have enough different instances of "non-objects" in the sample to distinguish them from the object reliably.

The technique of image augmentation is used to eliminate the imbalance. Each picture of the object is slightly deformed: it rotates a small angle, reflects itself horizontally and scaled. Thus, the number of positive samples in the sample is significantly increased.

For "non-objects" in the sample, it is essential to divide by "exactly not objects" and "not objects with parts of objects". Both types are essential for training and are usually created as part of a separate procedure.

We note the resultant approach to the selection of features based on the Haar cascade, demonstrated in the work [15]. And also, the Viola-Jones object detector, proposed in the articles [16, 17], and using the cascading architecture of classifiers.

Cascade architecture classifiers for object detectors in images has been developed in the following works [18, 19].

VII. Detectors of objects based on neural networks

Before the appearance of the Viola-Jones detector 16, the best accuracy reached by detector based on neural networks, proposed in [20]. In work, [21] the convolutional neural network was used as one of the classifiers in the cascade. In addition, in the study [22] the cascade of the three classifiers and regression is already wholly built on the pre-trained convolutional neural networks. At each stage of the cascade, both the object classification and regression are performed to determine the object bounding boxes. Thus, the cascades combine three convolutional neural networks with different complexity.

Further development of object detectors was obtained in particular neural networks. In work [23] the architecture of a neural network consisting of two cascades for the proposal of object bounding boxes and classification (R-CNN) is proposed. The training of R-CNN consists of three stages (Fig. 6).

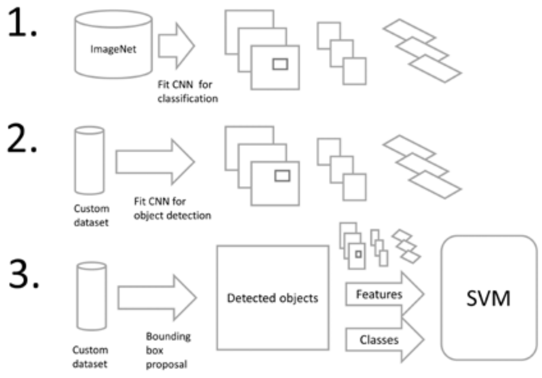

Fig. 6. R-CNN training steps.

The disadvantages of the R-CNN architecture include excessive calculations of features for each proposed bounding box, the need to bring all images to a single scale. An attempt to eliminate these shortcomings was made in the Fast R-CNN architecture [24].

For this purpose, a mechanism for the spatial pyramid pooling (SPP) is proposed. The essence of SPP is to calculate the attributes of objects within a bounding box using one convolutional neural network for the entire image Further improvement of R-CNN is suggested in the works [11, 25] and is to use softmax activation instead of the linear SVM classifier and combine the CNN to extract features and CNN to identify the bounding boxes.

The problems that have arisen in this approach are noted in the work [26] and consist of difficulties when working with objects in very different scales on one image. To do this, you have to train Faster R-CNN for different scales of the image.

Thus, according to the research [27] Faster R-CNN architecture combines speed and accuracy by creating multiple intersecting hypotheses for bounding boxes, attributes calculated independently for each bounding box based on a single convolutional neural network and a separate evaluation of the bounding boxes.

Further ways to speed up image processing were developed by the authors of the study [28] in the R-FCN architecture and are to replace fully connected layers (Dense) to convolutional layers (Conv2D) with a dimension of 1x1 (Fully Convolutional Networks, FCNs). Such a replacement allows not to convert images to the scale of the sample used to train the convolutional neural network for classification (224x224 for ImageNet).

Currently, the fastest architecture of object detectors is Single Shot Detectors (SSD). One of the striking examples of SSD is the YOLO (You Only Look Once) detector [29], developed by Google developers. Research [30] presents a new SSD architecture that showed higher performance than YOLO.

VIII. Scene labeling

Scene labeling strategy is a segmentation-based approach, an image is segmented and its various regions are classified, unlike classifying the individual pixels.

Scene labeling demands contextual information because of the labels tend to be dependent across pixels. Further, every image consists of information that is required to label pixels at several levels.

A. *Superpixel algorithm*

In order to enhance the segmentation, some pre-processing techniques are implemented. One of the widely used techniques is superpixel to segment the image.

The superpixels algorithm [31] produces compact and perceptual meaning to small regions of image. Each pixel in a superpixel symbolizes the essential piece of the same object.

By now, publicly available superpixel algorithms have turned into conventional tools in low-level vision.

The authors are familiar with more than ten algorithms based on superpixel technique. We discovered that Superpixels from Edge-Avoiding Wavelets (SEAW) [32] is not yet popular in discussions so far.

An exclusive web portal [33] was created to compare the performance of various algorithms based on the superpixel algorithms.

IX. Computer vision in geoscience

With the development of spatial information technology, remote sensing imagery has become a vital data source for many geoscience domains.

Recent studies using space and aerial images to detect earthquake-induced building damage via image segmentation [34] using joint color and shape features. Damage detection problems such as crack detection [35] and structural damages [36] have been investigated using local contexts encoded by different methods. Vision-based bridge component extraction approach was developed in [37].

An urban vehicle detection algorithm was proposed via dictionary learning for aero photos in [38].

An exciting study of image segmentation for search textures such as sand ripple, hard-packed sand, and rock was conducted in work [39]. The sonar images have the same nature as seismic ones.

The focus on spatial attributes and its examination in a new application for seismic interpretation, i.e., seismic volume labeling was made in research [40]. For this application, a data volume was automatically segmented into various structures, each assigned with its corresponding label.

Spectral decomposition is one of the sources of seismic images. Hyperspectral images are the subject for classification with independent component discriminant analysis [41], Bayesian approach [42] and Generative Adversarial Network [43].

X. Research hypothesis

Geophysical high-resolution images serve as a source for obtaining new information using computer vision.

The authors identified two areas for further research that deserve attention:

1. Detection of geological objects. Determination of their shape, relative position, and volumetric characteristics,
2. Determination of the content of geological objects. Energy characteristics of geological objects.

These areas of research are not fundamentally new. But with the advent of modern tools and techniques for working with high-resolution images, it is advisable to re-evaluate their capabilities.

The main research hypotheses that the authors have accepted for themselves in subsequent works are as follows:

Do the methods of computer vision allow to create a fully automated process for the identification of geological objects by seismic volume?

References

[1] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based cnn with improved triplet loss function,” in Proceedings of the iEEE conference on computer vision and pattern recognition, 2016, pp. 1335–1344.

[2] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear cnn models for fine-grained visual recognition,” in Proceedings of the iEEE international conference on computer vision, 2015, pp. 1449–1457.

[3] G. Levi and T. Hassner, “Age and gender classification using convolutional neural networks,” in Proceedings of the iEEE conference on computer vision and pattern recognition workshops, 2015, pp. 34–42.

[4] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proceedings of the iEEE international conference on computer vision, 2015, pp. 3730–3738.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” Yale University New Haven United States, 1997.

[6] N. Wang, X. Gao, D. Tao, and X. Li, “Facial feature point detection: A comprehensive survey,” arXiv preprint arXiv:1410.1037, 2014.

[7] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” in Proceedings of the iEEE conference on computer vision and pattern recognition, 2013, pp. 532–539.

[8] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Computer vision and pattern recognition, 2005. cVPR 2005. iEEE computer society conference on, 2005, vol. 1, pp. 886–893.

[9] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, “Face detection without bells and whistles,” in European conference on computer vision, 2014, pp. 720–735.

[10] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, “How far are we from solving pedestrian detection?” in Proceedings of the iEEE conference on computer vision and pattern recognition, 2016, pp. 1259–1267.

[11] R. B. Girshick, “Fast r-cNN,” in 2015 iEEE international conference on computer vision (iCCV), 2015, pp. 1440–1448.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in neural information processing systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[13] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[14] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The kITTI vision benchmark suite,” in Conference on computer vision and pattern recognition (cVPR), 2012.

[15] R. Lienhart and J. Maydt, “An extended set of haar-like features for rapid object detection,” in Proceedings. international conference on image processing, 2002, vol. 1, pp. 900–903.

[16] Viola and Jones, “Detecting pedestrians using patterns of motion and appearance,” in Proceedings ninth iEEE international conference on computer vision, 2003, vol. 63, pp. 734–741.

[17] P. A. Viola and M. J. Jones, “Rapid object detection using a boosted cascade of simple features,” in Proceedings of the 2001 iEEE computer society conference on computer vision and pattern recognition. cVPR 2001, 2001, vol. 1, pp. 511–518.

[18] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 743–761, 2012.

[19] P. Dollár, S. J. Belongie, and P. Perona, “The fastest pedestrian detector in the west.” in Bmvc, 2010, vol. 2, p. 7.

[20] H. A. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23–38, 1998.

[21] J. H. Hosang, M. Omran, R. Benenson, and B. Schiele, “Taking a deeper look at pedestrians,” in 2015 iEEE conference on computer vision and pattern recognition (cVPR), 2015, pp. 4073–4082.

[22] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural network cascade for face detection,” in 2015 iEEE conference on computer vision and pattern recognition (cVPR), 2015, pp. 5325–5334.

[23] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in CVPR ’14 proceedings of the 2014 iEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.

[25] S. Ren, K. He, R. B. Girshick, and J. Sun, “Faster r-cNN: Towards real-time object detection with region proposal networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.

[26] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” european conference on computer vision, pp. 354–370, 2016.

[27] J. Huang, “Speed/Accuracy trade-offs for modern convolutional object detectors,” in 2017 iEEE conference on computer vision and pattern recognition (cVPR), 2017, pp. 7310–7311.

[28] J. Dai, Y. Li, K. He, and J. Sun, “R-fCN: Object detection via region-based fully convolutional networks,” neural information processing systems, pp. 379–387, 2016.

[29] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in 2016 iEEE conference on computer vision and pattern recognition (cVPR), 2016, pp. 779–788.

[30] W. Liu, “SSD: Single shot multiBox detector,” european conference on computer vision, pp. 21–37, 2016.

[31] D. Stutz, A. Hermans, and B. Leibe, “Superpixels: An evaluation of the state-of-the-art,” Computer Vision and Image Understanding, vol. 166, pp. 1–27, 2018.

[32] J. Strassburg, R. Grzeszick, L. Rothacker, and G. A. Fink, “On the influence of superpixel methods for image parsing.” in VISAPP (2), 2015, pp. 518–527.

[33] D. Stutz, A. Hermans, and B. Leibe, “Superpixels: An evaluation of the state-of-the-art,” Computer Vision and Image Understanding, vol. 166, pp. 1–27, 2018.

[34] S. Li, “Unsupervised detection of earthquake-triggered roof-holes from uAV images using joint color and shape features,” IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 9, pp. 1823–1827, 2015.

[35] Y.-J. Cha, W. Choi, and O. Büyüköztürk, “Deep learning-based crack damage detection using convolutional neural networks,” Computer-Aided Civil and Infrastructure Engineering, vol. 32, no. 5, pp. 361–378, 2017.

[36] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk, “Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types,” Computer-Aided Civil and Infrastructure Engineering, vol. 33, no. 9, pp. 731–747, 2018.

[37] Y. Narazaki, V. Hoskere, T. A. Hoang, and B. F. Spencer Jr, “Automated vision-based bridge component extraction using multiscale convolutional neural networks,” arXiv preprint arXiv:1805.06042, 2018.

[38] X. Zhang, H. Xu, J. Fang, and G. Sheng, “Urban vehicle detection in high-resolution aerial images via superpixel segmentation and correlation-based sequential dictionary learning,” Journal of Applied Remote Sensing, vol. 11, no. 2, p. 026028, 2017.

[39] A. Zare, N. Young, D. Suen, T. Nabelek, A. Galusha, and J. Keller, “Possibilistic fuzzy local information c-means for sonar image segmentation,” in Computational intelligence (sSCI), 2017 iEEE symposium series on, 2017, pp. 1–8.

[40] Z. Long, “A comparative study of texture attributes for characterizing subsurface structures in seismic volumes,” Interpretation, vol. 6, no. 4, pp. 1–70, 2018.

[41] A. Villa, J. A. Benediktsson, J. Chanussot, and C. Jutten, “Hyperspectral image classification with independent component discriminant analysis,” IEEE transactions on Geoscience and remote sensing, vol. 49, no. 12, pp. 4865–4876, 2011.

[42] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, “Active learning with convolutional neural networks for hyperspectral image classification using a new bayesian approach,” IEEE Transactions on Geoscience and Remote Sensing, no. 99, pp. 1–22, 2018.

[43] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, “Generative adversarial networks for hyperspectral image classification,” IEEE Transactions on Geoscience and Remote Sensing, no. 99, pp. 1–18, 2018.G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.