# LSADN: Large Scale Automatic Detection Network
# A proposal for a Survey-Focussed Architecture

Martin Olivier

June 24, 2019

# Contents

# 1 Introduction

In the recent years, much progress has been made in the field of automated detection in Geophysical Surveys, but also surveys in general, by the introduction of newly developped Deep Learning techniques. Those techniques, especially Convolutional Neural Networks (CNN), allowed researchers to create more accurate and reliable models faster than ever before. We were now capable of analysing the vast swaths of data generated by surveying the landscape, and pick out the interesting features. During the recent CAA conference in Krakow, we have seen that automated detection in Archaeological Sureys using novel techniques has brought a lot of interest. Most approaches used very popular architectures, such as Mask-R-CNN[1][2] or YOLOv3 [3] to detect and segment objects.

While those techniques have proved themselves extremely capable and reliable in normal conditions, we have seen hints that they struggle in detecting objects in large scale images. Moreover, we have seen that those systems have trouble understanding the archaeological properties of detected objects.

In this short document, we will first expose some of the limitations of traditional architectures when applied to geophysical surveys, and surveys in general. We will then propose a set of specifications designed to combat those limitations.

# 2 Shortcomings of popular automatic detection systems

## 2.1 Differences between Natural Image and Geophysical Survey Processing

Deep Learning techniques, especially the ones that are used in what we will call "Natural Images" or images that are taken from a camera, and look similar to what we see with our own eyes, struggle to understand the complexity of large scale surveys. We believe that it is because those techniques are constructed using assumptions that don't hold in a world seen using geophysics. While a traditional network might not often times be asked to detect and classify objects that makes up most of the image for example, this might be the case in surveys.

This is due to the fact that in a photograph, if two objects look the same (i.e. they send a similar signal to the camera sensor), they will generally be the same object. In Geophysical surveys, this is not necessarily true. Two look a like objects can, and often will, have different classifications. For example a funeral mound and a roundabout look very similar seen in LiDAR, yet they are simply different objects.

A part of the problem could be explained by the fact that our classification came from our human senses. We see the world using visible light, so we construct a classification system that is heavily based on this. However, when we see the world using a different wavelength, our traditional classification system fails, and we cannot differ objects based purely on their appearance.

This is the case with geophysical surveys. In those, we see objects using a very different set of physical properties. In geomagnetic surveys, metallic objects, even tiny ones will have an important response to a magnetic field, and thus appear large on the final image. However, massive trenches, which are just moved earth, do not have such a response, and will only appear very faintly.

## 2.2 Scale

The problem of scale, and having a system that is robust to it comes up often. While we need to detect small objects, such as tombs, and small construction work, we also need to be able to detect massive objects, that may span

hundreds of meters. The types of surveys that are used in Archaeology are usually extremely large, not only in resolution but also in the vastness of terrain covered.

It then becomes extremely complicated to treat the entire survey using a simple convolutional network with a moving window for example, as it will be extremely slow and/or will miss out on most objects. Reducing the resolution by scaling down is a possibility, however this comes at the cost of making small objects disappear, and this cost becomes even more prevalent the larger the original image is.

Automated object detection systems, such as Mask-R-CNN, YOLOv3 or else, significantly scale down the input image size. This is done to reduce the impact on memory. However this scaling destroys the finer details of the image. While this downscaling can be done up to reasonable limits on natural images without too much of an impact on accuracy, this is not possible on images with a very high resolution, such as surveys.

## 2.3 Precise Labeling

Another problem that arises often in detection with geophysics is labeling, and the limits of an object. Especially with geomagnetism, objects produce a signal that goes beyond their physical borders. A nail might be small, but when seen on a geomagnetic survey, they create an enormous signature. This kind of problem is not seen in Natural Images, where the physical borders of an object usually reconciles very well with its signal produced and detected by the camera sensor.

## 2.4 The difficulty of inferring archaeological nature

Another point that needs to be clarified is the fact that inferring and understanding the archaeological nature of objects in surveys automatically is a very hard task. At its core a survey is simply an image. While it might contain very useful information about the location, size and general shape of objects, it does not hold information about its past. This information is contextual, and is not necessarily contained in the survey itself; it is often brought over by the human observer who knows information about the area and its history.

Moreover, Convolutional Neural Network are not capable of understanding

the relation between objects or parts of objects very well. A CNN understand an image by understanding smaller and simpler parts of it, and is focussed on the shape of those objects. A CNN will have a lot of trouble to understand the relation between the objects that it is able to detect and classify. To reuse the same example: a funeral mound and a roundabout look very similar. However, a human observer, even barely trained in analysis will understand that a roundabout is placed at the ends of objects that looks like roads, while funeral mounds are placed in middle of fields or forests. This information does not come purely from the object themselves but rather *from the relation between them.*

Understanding the relation between objects in an image is a very complicated task, and is still part of a very active area of study in Deep Learning[4][5].

While it might be possible to infer some part of the object's history using purely the survey, for example if the object has a very particular shape. This is not possible in the general case.

For all of these reasons, it seems that a dedicated architecture, one that is designed to analyse and understand large scale surveys is necessary. In the next section, we will define some general specifications for such an architecture.

# 3 General specifications for a survey-focussed architecture

## 3.1 Scale Robustness

The type of objects we will try to capture in this type of data takes up very varying scale. This scale variation makes it hard for the network to create a suitable internal representation. This is a know problem with Convolutional Units: while they usually create representations that are robust to translations, they often fail to accurately detect the same object with different scale, or when it is rotated[6, p. 341].

In a system that must analyse and infer from massively large scale surveys, both in resolution and in area covered, this is a major issue. While in most Natural Images, this issue can be fixed by scaling the input image down to a reasonable resolution, this is almost impossible to do in survey without massive loss of details, and thus, potential objects.

Moreover, the sheer size of the input surveys is an issue in of itself. We will need to find efficient ways of storing this type of information in memory, without sacrificing to much in accuracy. Scaling down the image is problematic due to the suppression of small objects, and segmenting it into small squares let us run the risk of segmenting over an object.

There are examples in the literature of architecture built to be more robust against scale variance. For example **Scale Invariant Convolutional Neural Network**, by Xu et al[7]̇This approach uses a multi-column architecture, where each columns focusses on a different scale of the image. Others, like **Multi-Scale Orderless Pooling of Deep Convolutional Activation Features** by Gong et al[8]üses a particular pooling type on the convolutional units activation to create a feature that can be used for detection.

Those networks are less sensible to scale variance, and implementing them in a automated detection system for large scale surveys would allow us to map out larger areas and detecting larger objects while remaining accurate with small objects.

## 3.2 Handling of Multiple Data types: Ensemble Methods

We believe that to maximise accuracy, a robust system should be able to analyze different types of data, be it LiDAR, Satellites Images or any kind of frequently used geophysical data type. It should not only be able to analyse this kind of data separately, as would be the case for a normal computer vision system, but it should be able to use the inference made on each data type and combine them to further enhance detection accuracy.

This can be done in various ways, but the most direct would be to simply use different networks, each trained on a specific data type, and using the feature maps produced in combination for further inference. This approach would allow us to use each network separately, in the case where only one type of data is available for example.

Figure 1 present a rough idea on how the different networks would interact. Note that this is a similar idea than the one presented in Xu et al̇paper where different columns of CNN focusses on different aspects of the input. In Xu et al̇paper the columns try to find representation of different scales, here we try to find representation of different data types as input features.

## 3.3 Region Proposals

As we are aiming to not only detect objects in a survey, but also infer their position, a region proposal module is necessary. A decision will have to be made on whether to use a simple region proposal module, which simply outputs bounding boxes, or something more sophisticated that allows pixel-wise segmentation, as the one used in Mask-RCNN.

A Region Proposal module would be in charge of analysing the feature maps produced by the Network and identifying Areas of Interest. This module produce bounding boxes where an object could be located, along with a probability. A Non Maximum Suppression is then applied onto the bounding box to only keep the boxes most likely to contain an object.

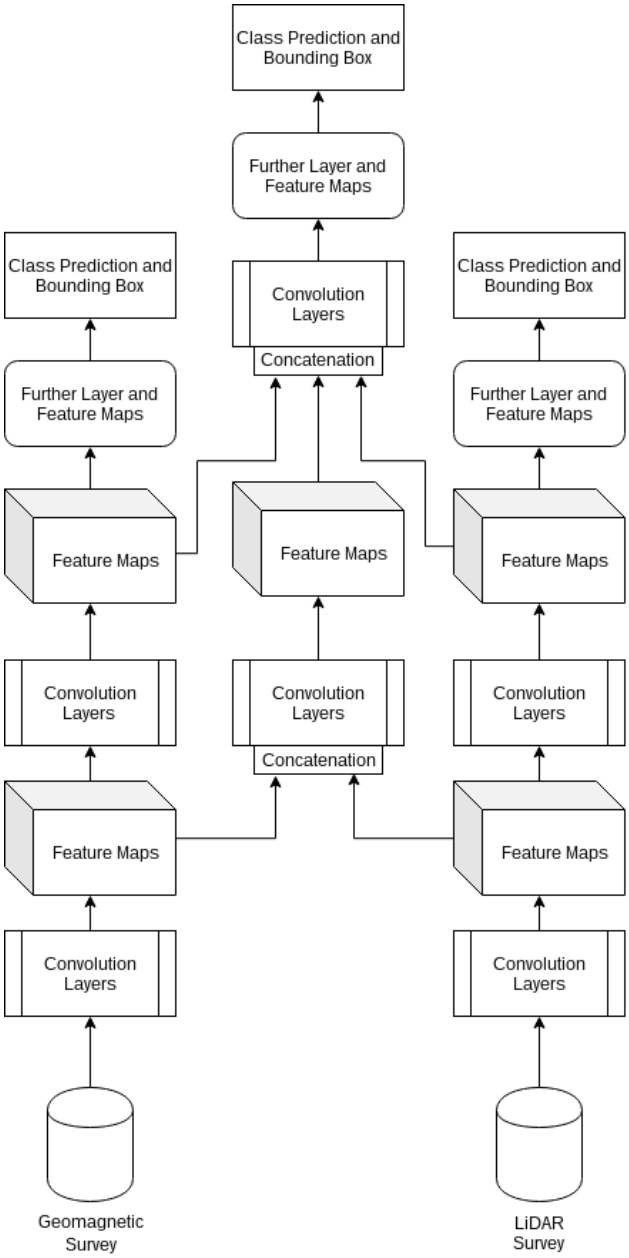


Figure 1: Rough Draft of the proposed architecture: each network can be used independently, or their feature map can concatenated and feed another network.

# 4 Conclusion

In this short document we have presented a rough overview on the current limitations of existent systems, and why they will fail at understanding and detecting objects in large scale surveys. We have exposed the inability of CNNs, but more generally Deep Learning techniques at understanding complex relationships structures, which is a crucial part in understanding the history behind an object, and the determination of its archaeological nature.

In passing, it should also be noted that the currently used architectures, such as YOLOv3 or Mask-R-CNN, while very accurate, are often optimized for speed. This is due to the fact that the end goals of those systems is to be able to do Real Time Detection, which is crucial in many "Real World" applications of Computer Vision (e.g. Passenger counting, object avoidance and recognition in autonomous robots, *etc.* While execution speed and optimisation should always be a concern, it is not as important in our use case. We do not have the need for a result to be ready immediately, especially if this increase in speed produces a notable decrease in accuracy.

It is also important to recognize the inability of those systems to perfectly distinguish between archaeological and non archaeological objects using information contained purely in the survey. However, we can achieve a good compromise by building a reliable and accurate object detection system designed for geophysical surveys. The automatic separation between archaeological and non archaeological objects is not a trivial task and should be its own subject, rather than be an "add-on" to automatic detection.

In conclusion we believe that, in order to reliably analyze and process large amount of surveys, a purposefully built architecture must be created. We cannot continue to simply use pre-existent systems in automatic detection and hope to have results that we can rely on. Those systems, while extremely efficient and robust in most contexts, are simply not made for this type of tasks. We have also presented a few characteristic that a system designed to analyze geophysical surveys should have. Those specifications are, for us, necessary but not sufficient conditions for a system to be reliable in its prediction in this particular environment.

# References


[1] K. He et al. "Mask R-CNN". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.

[2] W.B Verschoof-van der Vaart and Lambers K. "Learning to Look at LiDAR: The Use of R-CNN in the Automated Detection of Archaeological Objects in LiDAR Data from the Netherlands. Journal of Computer Applications in Archaeology". In: *Journal of Computer Applications in Archaeology* (2019), pp. 31–40.

[3] Joseph Redmon and Ali Farhadi. "YOLOv3: An Incremental Improvement". In: *CoRR* abs/1804.02767 (2018). arXiv: 1804.02767. URL: http://arxiv.org/abs/1804.02767.

[4] Adam Santoro et al. "A simple neural network module for relational reasoning". In: *CoRR* abs/1706.01427 (2017). arXiv: 1706.01427. URL: http://arxiv.org/abs/1706.01427.

[5] B. M. Lake et al. "Building Machines That Learn and Think Like People". In: *arXiv e-prints* (Apr. 2016). arXiv: 1604.00289 [cs.AI].

[6] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. Cambridge, MA, USA: MIT Press, 2016.

[7] Y. Xu et al. "Scale-Invariant Convolutional Neural Networks". In: *arXiv e-prints* (Nov. 2014). arXiv: 1411.6369 [cs.CV].

[8] Y. Gong et al. "Multi-scale Orderless Pooling of Deep Convolutional Activation Features". In: *arXiv e-prints* (Mar. 2014). arXiv: 1403.1840 [cs.CV].