GIMA
Geographical Information Management and Applications

An eye in the sky: a use-case for evaluating super resolution

Master thesis

Author: Yannick Bouten (5873452)

Email: y.a.m.bouten@students.uu.nl

Supervisor: Jesús Balado Frías PhD

External supervisor: Vincent Dieduksman MSc

Responsible professor: Prof.dr.ir. Peter van Oosterom

Date: November 9th 2022

Universiteit Utrecht
TU Delft
UNIVERSITY OF TWENTE.
WAGENINGEN UR
Defence Materiel Organisation
Ministry of Defence
JIVC
KIXS

Source image: https://www.esa.int/Enabling_Support/Operations/Sentinel-2_operations

Disclaimer

This research is written by me, Yannick Bouten, as both a student at Utrecht University and intern at JIVC/KIXS, part of the Defence Materiel Organisation of the Dutch Ministry of Defence. The Ministry of Defence formulated the initial needs for research into this topic, which were further developed by me in collaboration with my supervising team and with the support of the University.

This research is deemed unclassified ("ongerubriceerd") by the Beveiligingsautoriteit JIVC and therefore approved for publication in this form. No data or software has been used from the Ministry that would otherwise infringe upon the unclassified nature of this research or its scientific transparency. The research is entirely unrelated from current use of deep learning methodologies by the Ministry of Defence so any similarities are entirely coincidental.

Questions about the research can be directed to me personally and feel free to do so. Questions about the classification can be directed to the Beveiligingsautoriteit JIVC.

Preface

Dear reader,

I proudly present to you my thesis as part of the GIMA master programme, which I worked on form March until November 2022.

When I started the preparations for a thesis research in May 2021 I did not expect it to be accompanied with so many challenges, both organisational and personal, but the fact that I managed to overcome all of those and present this result feels like a true personal victory. In the 8 months I have been working on this thesis I learned a lot and gained experience in what for me was a relatively new field of work and I hope that by sharing this with you in this thesis research you can learn something as well that is usable in your own line of work or can contribute to your understanding about geo-information and deep learning.

Before diving into the research I do want to use this opportunity to thank everyone that contributed to my research. First I want to thank my supervisors. Thank you Bas for helping me get started with this research at the Ministry of Defence, both your knowledge and enthusiasm about geo-information were really catching and also inspired me to explore this topic. Unfortunate that our ways had to separate before the end of the project but I am really grateful for all the effort you put into it in the start-up phase. Vincent I really want to thank you for taking over the supervision from Bas. Although geo-information was not per se directly your own field of knowledge your input has been of great value and I definitely also learned things from you and your expertise in data science. I also enjoyed our almost weekly meetings as colleagues on a personal level. From the University I would like to thank Jesús for the scientific guidance and also sharing his insights on remote sensing, which formed a critical part of this research and supported the practical application of this research. Also thank you to Peter for being the responsible professor in overseeing the supervision and the proactive attitude in organising and enabling the execution of this research.

I would also like to thank all the colleagues at the Ministry of Defence who contributed to my work for both the research or to my employment at KIXS, without all of you that would not have been possible. Also many thanks to all the other colleagues at KIXS who I have got to know during the workdays at the Kromhout Kazerne, during lunch breaks, the weekly work meetings or any other KIXS-related events. It was great getting to know you all.

At last I also want to thank my friends, family and roommates. Just for the sometimes simple things in life like showing interest into what I have been doing, providing some (un)useful distraction or moral support when I was in need for it, thank you all.

That being said it leaves me nothing else to do than to hope you will enjoy reading this thesis and learning about the research I did. I also encourage you to take a look at the appendices to get a better understanding about the more specific technical and model related information, which served as input for visualisations, calculations or other operations that were done as part of this thesis (but were not included integrally in the core sections). If you are a Ministry of Defence employee I can also recommend taking a look at the KIXS SharePoint page, where you can find additional documentation and information focused on the use-case of this research for the Ministry.

Yannick

Abstract

The aim of this research was to execute a proof of concept on the added value of deep learning methodologies as part of remote sensing analysis. This was done in collaboration with the Dutch Ministry of Defence to improve the knowledge on this within the geo-domain. Deep learning constituted two different applications as part of this research; super resolution, which is to increase the spatial resolution beyond it is original limit (Yue et al., 2022) and feature extraction. The latter is also interesting as applying a geo-analysis task on super resolution data can prove to be a suitable methodology to evaluate the result and also attributes to the increase scientific interest in deep learning within the field of remote sensing (Yang & Newsam, 2013).

For super resolution models with varying amounts of input data have been tested and the metrical evaluation showed no significant issues although the models could be further optimised. Augmenting data to increase the usability of a dataset proved promising in performance but not conclusive in its added value to modelling super resolution. Visually the super resolution models showed more detail in comparison to a Sentinel 2 image of the same area but their differences in metrics did not result in apparent visual differences between models.

Feature extraction showed that all super resolution models outperformed a Sentinel 2 based extraction model in metrics. In comparison to the ground truth road network the model proved difficult and below expectation.

The conclusion is therefore that super resolution and deep learning based analysis methodologies can be of added value for remote sensing analysis and usable in an accessible and application-oriented manner. On an absolute scale however both the evaluation metrics and evaluation analysis in the form of road extraction showed that it could definitely benefit from further optimization to improve both performance and generalizability of the models.

The discussion touched upon several aspects of the research that could attribute to this, including other types of satellite data, open-source modelling software and alternative analysis tasks using super resolution data.

Contents

1. Introduction

Remote sensing as a discipline has been around for a long time, but has really matured in the post-World War 2 era by the development of satellite technology and the broader discipline of what is called Imagery Intelligence (Fernandez-Beltran, Latorre-Carmona & Pla, 2017). Satellite imagery nowadays provides a qualitatively good, reliable and frequently updated source of intelligence and information on areas where ground-based or lower-altitude data collection is not always possible or geopolitically desirable. For example geographically remote areas or areas where natural disasters occurred but also geopolitical and military events. A fitting recent example is the built-up of military forces by Russia at the Ukrainian border and the armed conflict that evolved from it. Social media as a new form to gather intelligence became really apparent but geographical information (in the form of satellite imagery) did not lose importance, if it only were for geo-referencing images or videos from social media. Satellites therefore quite literally can serve as “eyes in the skies”.

However, these eyes have their limitations. Where aerial photography can provide imagery with a resolution of sometimes just a few centimetres, this is relatively rare for satellite sensors. Satellites like Spot 6 and Superview operate in the (very-)high resolution segment of satellite remote sensing, being able to deliver a respectively 1,5- and 0,5-metre spatial resolution for the panchromatic band and respectively 2- and 6-metre resolution for the multispectral bands. These are commercially operated meaning that data can be either expensive or restricted in use. Military satellites can likely provide imagery at similar or even better resolution levels but have even more restrictions regarding use. There are however also public satellites that provide imagery in an open-access format (Gargiulo, Mazza, Gaetano, Ruello & Scarpa, 2019) like NASA’s Landsat programme or ESA’s Sentinel mission. These satellites deliver multispectral imagery up to respectively 30- and 10-metre resolution and the Landsat satellites provide 15-metre resolution for the panchromatic band (which Sentinel 2 is not equipped with). For a type of research that is heavily dependent on the use of the satellite imagery like detection, extraction, classification and monitoring change for an object of study like vegetation, urban environment or infrastructure one should consider if the spatial resolution is sufficient enough for the proposed research. Analysing small scale objects can be difficult on lower resolution data and the difference in resolution between open-source and commercial satellites can provide a reason to try and close this gap.

In the military the use of satellites and remote sensing technology is almost just as old as the technology itself as it can prove to be advantageous to have knowledge about an area of interest or operation, especially when a possible opponent does not have this or to a lesser extent. Important to consider when using satellite data as a source of intelligence is their two main technical constraints; the temporal and spatial resolution. Geospatial intelligence derived from satellite data can only be updated as often as the area is revised by the satellite, which is often at a constant interval but depending on the satellite can range from a revision time of just one day up to almost two weeks. Real-time data and intelligence gathering is therefore difficult which would shift the focus more to the need for intelligence about more solid objects like infrastructure and the built environment. As also pointed out earlier the spatial resolution of data influences what types of analysis you can do with it (and the proposed research influences the required resolution) so that is also important to consider as data gathering can be challenging depending on the needed spatial resolution. It can be assumed that the military and intelligence branches of many countries have very high-resolution sensors at their disposal but for geopolitical and security reasons those sources are not openly available for other research or their existence itself might not even be publicly disclosed. Research outside the actual intelligence domain is therefore more dependent on publicly available or purchasable

data or its own adaptability to overcome issues in regard to the resolution with technical solutions or innovative methodologies to accommodate the research needs.

An important and emerging trend is the use of SR (Super Resolution). Super resolution is the term for an algorithmic based approach to increase image resolution beyond the original boundaries of the sensor and original images, which results in higher resolution data that especially in the field of remote sensing is able to facilitate new developments and opportunities for using images (Bioucas-Dias et al., 2013; Li, Yang, Dong, Wang & Huang, 2020; Li et al., 2020). Super resolution is seated on the principle of being able to gradually "learn" over time by putting in large amounts of data and analysing it through a layered neural network. This is just one of the many fields and applications of deep learning and has become a recent trend in remote sensing in relation to super resolution but also for the purpose of analysis (detection, extraction, classification and regression) (Krizheysky, Sutskever & Hinton, 2012) because of its high-end performance compared to traditional methodologies. As a subset of deep learning it uses algorithms (the most well-known are CNNs (Convolutional Neural Networks)) and can be seen as a deepening to current and more manual analysis methodologies (Tian & Ma, 2011).

Figure 1.1: Sentinel 2 (left image) versus super resolution (right image)

Source: https://mdl4eo.irstea.fr/2019/03/29/enhancement-of-sentinel-2-images-at-1-5m/

Shown in figure 1.1 is a hands-on example of super resolution that has been performed on an area in southern France, with on the left the original and on the right the super resolution image as a result of a training model that used Spot 6 data. This illustrates that super resolution has been performed successfully in the geo-domain and it shows results in a practical manner without just looking at performance parameters. However, it could be an extra dimension to super resolution research to not only perform it, evaluate the parameters and show the output (of which the latter was already qualitatively well done in this example) but also to try and use super resolution data in an actual remote sensing analysis task. This could be a feature extraction or a detection assignment but the added value of it would be to show hands-on what super resolution could offer to a specific research objective and how it compares to data of a different resolution level. The latter would also illustrate the actual need of doing super resolution or if already existing data sources are sufficient for a task at hand.

The next chapter will discuss the goals and research questions, expanding and operationalising the introduction on remote sensing analysis and deep learning into a research design. The chapter will also discuss the research limitations and elaborate on what this research is not about.

2. Research goals

2.1 Research questions

Together with the Ministry of Defence, this Master thesis will serve as a proof of concept on the possibilities of using and applying super resolution, and also in a broader aspect about the use of deep learning as a methodology to enable super resolution in the first place and also to evaluate its performance. The research and knowledge enrichment this thesis will provide on this subject can be beneficiary to the different branches and departments of the armed forces that are involved in the geo-domain but it is also meant to go beyond interior purpose and to enable use by others. The Ministry of Defence can use its means and knowledge to contribute to the overall body of knowledge within the geo community and diving further into deep learning also attributes to a better understanding of its potential in geo-information science.

An important distinction to make is that deep learning as a methodology serves a dual purpose as part of this research. First and foremost, it is about applying a deep learning model that is able to improve the resolution of input data to a higher standard which is defined as super resolution. A second purpose is the use of a deep learning model as a way to do feature extraction. This distinction is important to make because in analysing an image the pixels in the image need to be firstly grouped based on spectral and spatial similarities (segmented) before the next step can be taken, which is grouping those segments into distinctive classes (classification) (Esri, n.d.). The added value of this analysis in relation to the research is to make a quantitative comparison between unmodified and super resolution data. Super resolution is an operation and eventually a result on itself and deep learning extraction is a methodology to evaluate the quality of this operation and the resulting data.

The main research question therefore focusses on the super resolution aspect of this research and also immediately touches upon the case for which this proof of concept will be tested. The sub questions dive deeper into the specifics of super resolution and how it is already being used and while also pay attention to deep learning in relation to feature extraction. As it is a practical proof of concept the sub-questions will also be strongly focussed on discussing the empirical results as this can be important knowledge and hands-on experience for future use of super resolution, for both the military and others.

The main research question is therefore as follows:

To what extent can super resolution be of significant added value to feature extraction on Sentinel 2 data?

To make answering this question more manageable and also improving the research structure, this question is broken-down in several sub-questions:

-What is super resolution and how is it being used in geographical information systems as a deep learning application?

-What are the needs and requirements for satellite imagery to be suitable for use in a super resolution model?

-To which degree does a super resolution model enhance the resolution of Sentinel 2 data?

-How can a deep learning model be applied to analyse Sentinel 2 data?

-What are the qualitative and statistical differences in a feature extraction analysis by using super resolution data?

2.2 Research limitations

This research is in the first and foremost place about assessing the added value a super resolution model could provide in comparison to unmodified (lower resolution) satellite data and how this influences the results of an actual analysis.

The deep learning analysis aspect of this research is definitely also an interesting and currently prominent development in the field of geo-information is subservient to the super resolution component of the research. Deep learning analysis as a development within the field of GIS (Geographical Information Systems) could be an entire subject for a thesis by itself but in this context, it is situated as a methodological component then the overall theme of the research. Applying a deep learning extraction model as a methodology is just one of the many already existing approaches for remote sensing analysis that are being used. It is a suitable choice for analysis in this research as it is a relatively shallow and accessible network architecture and super resolution itself is seated on the principles of deep learning and therefore the analysis builds on the same foundational model characteristics. This justifies its use as an evaluation methodology of the super resolution model but is in this research a mean and not the end.

The research itself can be characterised as explorative because practical research into super resolution is not that extensive. The goal is to show what can be done within the possibilities of super resolution and how that reflects in a subsequent deep learning analysis model, whether or not it is desirable or favourable for an organisation to adopt such a workflow is up to their own choice.

As a partner in this research the Ministry of Defence is interested in explorative research in this field as it is not widely applied in the line of work of the department. Especially in a military context the parameters like the area of interest, the conditions of the time when data is collected and also the requirements for analysis can be different comparison to this research. Gaining insight into the theme of super resolution and providing the Ministry with a practical application of the methodology mainly serves the purpose of knowledge enrichment on this topic. The goal is not to advise or decide if and how it could be incorporated in their own research and analysis workflows. The usage of super resolution and deep learning analysis go beyond the military domain of remote sensing and therefore it has been chosen not to incorporate the military aspect into the research design and focus the analysis on a subject that also has a non-military use. This approach also benefits transparency and usage by other researchers as it is not directly restricted by confidentiality agreements of any kind.

As also stated in paragraph 2.1 this research serves as a proof of concept about what the added value of using deep learning solutions such as super resolution and feature extraction could be as part of geospatial intelligence and remote sensing analysis. It is supposed to provide a practical insight and use-case about these technologies and the software that is being used without per se stating what is the ultimate best functioning solution for either the Ministry or other possible users. However the user experience in this specific case will be part of the discussion about this research.

3. Theoretical and conceptual framework

3.1 Deep learning in remote sensing analysis

Deep learning has been deemed as one of the major trends in big data and technology overall (MIT Technology Review, 2013), with its main characteristic being the use of neural networks in a network architecture of at least three layers (Zhu et al., 2017). Deep learning reduces the dependency on user input and knowledge as it is able to learn based solely on the input data. Much applied CNNs have shown to be suitable for analysing and extracting objects from images and segmenting semantics and RNNs (Recurrent Neural Networks) show a suitability for sequential data tasks as recognizing movements. This indicates the variety of tasks deep learning can be used for and the variance in different network architectures that are suitable for a certain task but also shows promise for application in the remote sensing field, although the characteristics of remote sensing and satellite imagery result in their own difficulties that need to be addressed.

(Zhu et al., 2017) address some geo-related specifications when it comes to using deep learning in RS (Remote Sensing):

-RS data can be collected by many different sensors which each have their own specifications and therefore data fusion might be relevant.

-RS data is geo-located, meaning that each object or pixel in the data is somewhere set in space and enables data fusion with other geographical data and also allows for linking images, location and routing services and even reality (by augmentation) to a geographical place in space.

-RS data use geodetic measurements, which are set and clearly defined parameters that can be achieved with a certain confidence. Knowledge about a sensor's quality and parameters are however important.

-Time and temporal resolution are essential as the surface captured by different sensors can change over time and nowadays satellites can provide new data in a short timeframe.

-Also, RS is faced with the challenges of big data, as single sensors can have a collection of large quantities of data. However, handling the data is rather streamlined due to sufficient annotations and metadata that accompanies the actual data.

-Often RS aims at gathering information on the physical geographical or biological quantities on earth then per se objects. But dependent on the actual purpose for a research expert knowledge might still be needed to interpret results.

The interest for deep learning within the field of RS can also be seen in statistics, as the quantity of papers written on this topic increased fast over time, especially since 2014. This is far from generic as shown by the sub-field on how to interpret high resolution imagery which again has further specifications like scene classification, object detection and image retrieval (Yang & Newsam, 2013).

Scene classification is an important concept in the analysis of satellite imagery as it includes several specific tasks in RS, with detecting objects and changes as some of the most important ones and as it is in the name also classifying them (Bhagavathy & Manjunath, 2006; Chen, Zhao, Li & Yin, 2006). The possibility of using high resolution imagery however forms a major challenge in this as it makes small scale objects better detectable in the analysis but can therefore be different to assign a specific scene to them as they are rather similar (Zhu et al.,

2017). Because of this the methodology for this type of analysis is two-folded; first extracting features and subsequently classifying those using different deep learning models. Three important categories of deep learning models exist:

-Pre-trained models, for example a CNN that relies on the original image data resulted in qualitatively good classification results. This by building the feature representation directly from the features out if the intermediate layers (Ayala, Sesma, Aranda & Galar, 2021; Castelluccio, Poggi, Sansone & Verdoliva, 2015; Hu, Xia, Hu & Zhang, 2015; Marmanis, Datcu, Esch & Stilla, 2016; Penatti, Nogueira & Dos Santos, 2015).

-Pre-trained models that are made fitting for specified conditions. This can be beneficial on small scale labelled sets of data (Castelluccio et al., 2015; Nogueira et al., 2017) and result in a better fit for a specific area but can be difficult to accomplish due to a relatively small scale of publicly available datasets (Xia, Yang, Delon, Gousseau & Sun, 2010; Yang & Newsam, 2010; Zou, Ni, Zhang & Wang, 2015).

-Models that need to be trained completely. A network is then trained with only the already existing dataset of satellite imagery but will likely be far less accurate in comparison to the already trained model categories. It can however be favourable to do so due to the complexity of the pre-trained models in especially the amount of needed parameters to be learned and would result in a better local fit for images but will need a new training for each new dataset (Luus, Salmon, Van den Bergh & Maharaj, 2015; Volpi & Tuia, 2017; Zou et al., 2015).

3.2 Super resolution

In the scientific literature the relevance of super resolution focusses on its benefits; it is a vivid field of research (Huang, He, Wu & Gou, 2021), it enables research methodologies that rely on high spatial resolution images which are originally difficult and costly to come by (Wang et al., 2019), apart from spatial also the temporal resolution of a certain data source can be important to monitor environmental changes (Pouliot, Latifovic, Pasher & Duffe, 2018) and also enhancing the data and therefore the research possibilities fits in the broader trend of open data (Galar, Sesma, Ayala, Albizua & Aranda, 2020).

Overall the goal of super resolution, independent from the specific technical operations undertaken to do so, is to achieve a spatial resolution that is in the end higher than the resolution offered by the original image. A classical method to achieve this is to perform an interpolation (either bi-cubic or bi-linear) which defines new pixel values based on a calculation of the adjacent pixels (Yue et al., 2022). It is however not ideal as it can lead to loss of finer spatial characteristics in the original image.

Achieving super resolution imagery is to some extent possible with the already present sensors and data of a singular satellite platform. Apart from RGB (Red, Green, Blue) and Infrared bands many satellites also have a panchromatic band, which is distinctive from multispectral bands as it captures light in a broader wavelength range which results in an image that shows brightness instead of colour, resulting in a higher spatial resolution then would be able to achieve in multispectral bands (Müller et al., 2020). Measured in resolution ratio the ratio in ground sampling distance of individual sensors is 1:2 (for the Landsat satellite) or even 1:4 (for Spot) between the individual multispectral and the panchromatic band (Ehlersa, Klonusa, Åstrandb & Rossoa, 2010). This difference in resolution opens up the possibility for data fusion, which in this case is called panchromatic sharpening, that fuses a lower resolution multispectral with the higher resolution panchromatic band to increase its resolution while preserving the multispectral image characteristics (Ehlersa et al., 2010).

Although resolution is a prevalent challenge in remote sensing it is not solely a geo-information related issue as it is also embedded in the broader field of computer vision; a discipline which also pursues the achievement of higher image resolution but traditionally focusses on the use of algorithms (Müller et al., 2020). CNNs, due to their similarity with already applied sparse-coding methods in in super resolution, provided a simple model architecture with three convolutional layers that provided end-to-end mapping from Low to high resolution images as what is known as Super Resolution CNNs (SRCNNs) (Goodfellow et al., 2015). SRCNNs can process three different colour channels (and are therefore suitable for RGB data) and pre-process data via bi-cubic interpolation and can be evaluated in performance with multiple widely used metrics, therefore both relying on already established methodologies but also being able to outperform most of them in the process (Müller, Ekhtiari, Almeida & Rieke, 2020).

Liebel & Körner (2016) emphasize the usability of SRCNNs, without making alterations to the architecture of the model and applying it directly on the three different spectral bands. It outperformed bi-cubic methods and showed it is applicability in remote sensing science, while it also pointed out possible improvements (Müller et al., 2020). Processing time was considered lengthy and in what is known as fast SRCNN the qualitative performance was similar to traditional SRCNNs while processing time decreased by a fortyfold. Also the input data was altered in multi-channel SRCNNs which enabled more than the original single input images, which addressed a common problem in super resolved imagery which was a lack in image frequency.

CNNs are considered a relatively shallow type of network architecture, meaning there are also deeper forms of network design available and researched. VDSR (Very Deep Super Resolution) and AE (Auto-Encoder) networks are considered to be more deep forms of network architecture compared to the traditional CNNs as described above and rely on the principal that the architecture allows to surpass layers (Müller et al., 2020). Although their deeper network structure they are able to reach better results compared to other CNNs and do show better values for PSNR (Peak-Signal to Noise Ratio) and SSIM (Structural Similarity Index) while it comes at the downside of a more extensive and complex network structure.

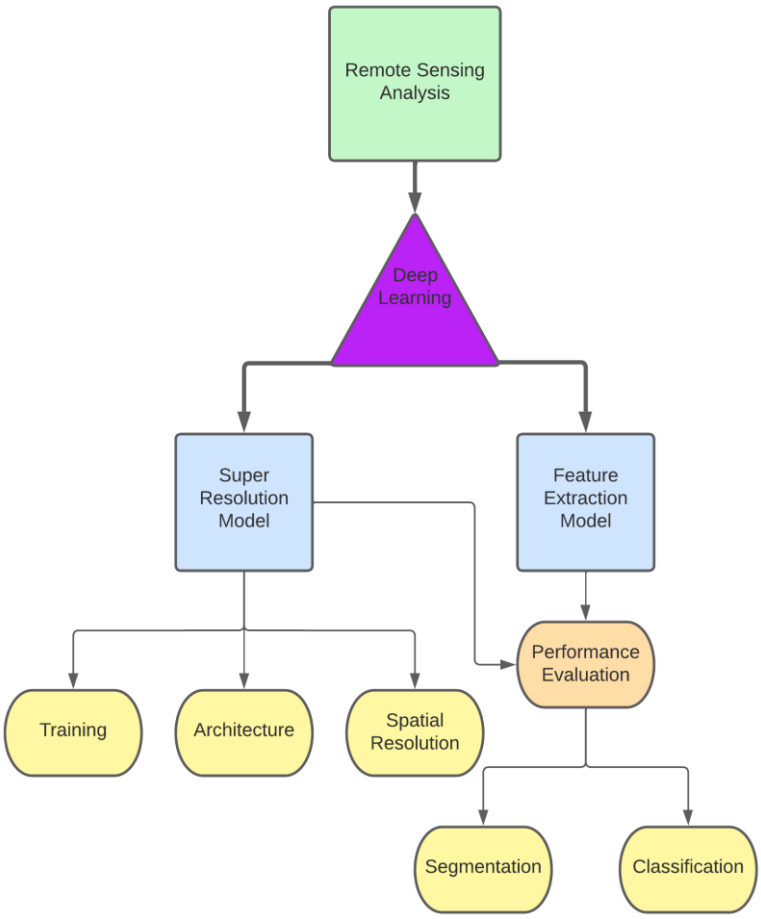
CNNs can also be distinguished based on their specific architectural design, where LeNet can be considered as one of the first architectures which was developed to detect visual patterns in images and initially performed well on handwritten digits (LeCun, Bottou, Bengio & Haffner, 1998). A more recent and for Remote Sensing purposes interesting CNN-architecture is the ResNet (Residual Network), developed by He, Zhang, Ren & Sun for the specific purpose of classifying images with an as low as possible error rate (2016). Data is augmented and down sampled (Simonyan & Zisserman, 2015) when handled by the convolutional layers but residual networks exploit the fact that in deeper layered networks the dimensions of data (in this case images) can be the same for the input and eventual output to allow the network to skip over layers. Counter-intuitive the deeper network structure does not lead to an increase in complexity because of this principle and also counters traditional saturation and eventual degradation of model accuracy and inflation of training error (He & Sun, 2015). ResNets address this optimization difficulty and were able to outperform less deep network architectures in classifying, detection and localization tasks on the ImageNet reference test dataset (He et al., 2016).

A final group of network architectures are the SRGANs (Super Resolution Generative Adversarial Networks), which are seated on the principle of trying to learn the structure of the input in order to recognize the consistency and patterns that are present in the input data. The

model then tries to generate sample cases as if they originated from the actual source data while actually, they were made by the model based on its knowledge of the input data. A SRGAN is divided into two separate models where one is used to generate samples and the second one tries to classify whether they are real or not, meaning that the generative aspect of the model can be classified as good performing if it is able to let the classifier believe that a generated sample is real. In super resolution the added value of SRGANs mainly lies in its ability to cope with a relatively low quantity and quality of data (Romero, Marcello & Vilaplana, 2020) while also containing a competitive element in its functioning which results in a high image fidelity (Müller et al., 2020). Its downside lies in that it is still outperformed by VDSR and AE types of architectures in PSNR and has a relatively quantity output and perceptual attractiveness compared to other methods as the output is more photorealistic as such (Ledig et al., 2017).

3.3 Conceptual model

Figure 3.1: Conceptual model



Made by: Yannick Bouten

The two preceding theoretical paragraphs in combination with the practical insights of the introduction accumulated into the conceptual model as shown in figure 3.1. The fundamental concept for performing remote sensing analysis in this research is the application of deep learning. The colour purple depicts versatility in a military context and is therefore also used here for deep learning, as it serves two different modelling tasks within this research. For

super resolution there are the three concepts of training, architecture and spatial resolution which based on the literature can be deemed as the most important components to consider in the methodological continuation of this research. Training is an important aspect as using a pre-trained model especially on a small scale dataset as it can benefit both model performance and processing time (Castelluccio et al., 2015; Nogueira et al., 2017). In the extent of the training component lies the specific architectural design where one that is suitable and preferably favourable for image analysis (He et al., 2016) can achieve the best result for the geo-information domain and also benefit performance and usability. Spatial resolution and to be specific the difference in resolution the model will try to overcome determines the requirements for data to train and test the models with but also whether or not it is feasible or that other data solutions might be more purposeful (Ehlersa et al., 2010). The feature extraction model is there to accommodate a performance evaluation of the super resolution model's output and as also explained in the introduction the extraction task can be sub-divided where pixels in an image are first segmented and then classified.

These are the main theoretical concepts that this research will be founded on. The upcoming methodology chapter will dive into the actual practical research design and the different aspects that are important to enable the actual modelling operations. It can be seen as a practical breakdown of the theoretical components of the conceptual model as depicted in figure 3.1.

4. Methodology

4.1 Features and label data

Before discussing the research area and the different needs regarding satellite data for this research, it is useful to first consider the feature that is chosen as subject for the feature extraction model. Although it is a methodology to evaluate the performance of the more important super resolution model it is still relevant to discuss as it partly forms the cause of this research and elaborates on why it is relevant to extract this type of feature from satellite imagery.

The feature chosen to classify is road infrastructure. From a military perspective infrastructure (and therefore roads) are important logistical corridors to enable and support all kinds of warfare. Infrastructure is a relevant factor in what is known as the operating environment and also the civil assessment which are both analysed in the preparation phase of a possible military action (Ministerie van Defensives, 2019). It is a crucial feature in a deployment area to have and gain intelligence about, both for your own but also to predict how your opponent will make use of the area's infrastructure (Ministerie van Defensie, n.d.). Therefore, having a working extraction model beforehand that can segment and classify an area's road infrastructure is more efficient and less time consuming while also providing information on it early or even before actual deployment. Besides a military purpose analysis can also be useful for remote areas in the world where doing this on site is not always possible. This could prove to be crucial information in for example the case of a natural disaster in an area in order to coordinate emergency assistance (which can depend on physical infrastructure).

The source for label data about road infrastructure to use in the extraction model of this research is OSM (OpenStreetMap), an open-source geographic data service with almost worldwide coverage. It is community based and for the Netherlands the overall quality and completeness of the database is very high. As it is open-source the classification scheme used for objects that are registered as part of the service is also transparent and therefore usable by others.

Table 4.1: OSM road classification

Road Type	**Description**
Motorway	A major highway with at least 2+ divided lanes (Autosnelweg [in the Netherlands])
Trunk	Second most important type that are not highways (Autoweg)
Primary	Third most important type of road that connects cities (Provincial road that is not an Autoweg)
Secondary	Type of road that connects settlements (local roads)
Tertiary	Level of roads that connect settlements and villages (small scale local roads)
Link*	Slip roads/ramps interconnecting all the road types above

**Link roads are a different road label for all these road types in OSM but are in this classification merged together*

Source: OpenStreetMap wiki

Table 4.1 provides an overview of the classification of road infrastructure as used by OSM which can also be used as part of the feature extraction model via label data. The benefit of using OSM is that there is already segmented and classified data for road infrastructure available for the whole research area which is available for download in Shapefile and other common geodata formats.

Geofabrik is a German foundation of OSM contributors that provides access to OSM data via their own download server in different data formats and in a structured manner with a worldwide coverage. However smaller data packages for specific countries and also provinces can be acquired. Data is also updated frequently (at the moment of writing the latest version is from 08-11-2022) and accompanied by metadata to maximize transparency and usage.

Important to note is that the choice for road infrastructure puts the analytical focus of this research on features and not on objects. This also steers the scope towards how to use super resolution as a data source to map a network of features (as roads are interconnected) rather than discrete objects in the geographical space of a satellite image. That also means that an extraction model should not be assessed based solely on its numerical output but how it compares to the ground truth data and the network as a whole. Also important to consider is that the roads corresponding to the labels as shown in table 4.1 are restricted to motorised traffic only, resulting in an analysis that form a military perspective would focus on the logistical aspect as that is dependent on motorised transport nowadays and roads also need to be suitable for this purpose. That means it leaves out the small scale infrastructure which could also be relevant from a military aspect (as it could be used for enemy movement) but for a proof of concept and logistical aspect would make the analysis unnecessarily extensive.

4.2 Research area for model training

It is important to designate the area in which this research will take place for a multitude of reasons. First it narrows down the data need as many satellites in principle provide worldwide coverage but that would put a great burden on the space needed to store data and also dramatically increase modelling time. For a proof of concept the amount of data needed and the time modelling takes should not be too extensive as the focus is on the experimental aspect of the methodology that is being tested. Secondly the spatial characteristics of the area (heavily urbanised/countryside) indicate what kind of features can be expected and also to what extent the results and analysis can be representable for similar regions. Adding to this are the geographical location and the local climate as these can influence the conditions under which data is collected and illustrate possible representation for other areas.

The Netherlands as an area is geographically well documented and data is relatively easy to access, which makes it optimal for a proof of concept. The North of the Netherlands has been chosen as the main research area as can be seen in figure 4.1. This area has several advantages in regard to data needs, as there is a large amount of high resolution images available. Also these images are relatively cloud-free and therefore also well illuminated, partially due to the proximity to sea. From a model training perspective this area is favourable in comparison to for example the highly urbanised Randstad area as it is a more rural area and training a model in this will be more representable and usable for other areas of interest instead of when it is trained on a largely dominant urban centre. The area borders the sea which reduces although not completely prevents the presence of cloud cover during data capture as a cause of stronger winds which is in this case advantageous. The North of the Netherlands is a sole statistical area (NUTS1-region, used for regional statistics) which constitutes out of the

three provinces of Friesland, Groningen and Drenthe, which makes discussing and quantifying the area size more transparent.

The area is characterised by a relatively low population density, with the provincial capitals being the only three municipalities with more than 100.000 inhabitants. This also translates to a relatively less extensive high-capacity road infrastructure and of which the most important inter-city roads are shown as red lines in figure 4.1.

Figure 4.1: Map of the Research Area

Made by: Yannick Bouten

4.3 Satellite data

For this research two different types of satellite data will be of importance; the medium-resolution data on which a super resolution model will be applied and tested and the high-resolution training data which will be used to train the model. High resolution data is used to train a super resolution model how to predict data on a higher resolution level. Subsequently a trained model will be applied to medium-resolution data to try and predict super resolution. It should be kept in mind that this is originally medium-resolution data but which will result in an image predicted at the high resolution at which the model has been trained.

This super resolution data will then be used to train a feature extraction model to recognise roads in the super resolution data based on the feature label data as described in paragraph 4.1. A similar extraction model (with the same label data) will be trained directly on medium-resolution data to be able to compare analysis results between what can be considered as original unmodified data and predicted super resolution data.

The research area, as defined in paragraph 4.2, is the area from which data will be used to train both types of models. Testing will be done on a different area in the Netherlands where cloud free remote sensing data is available from a similar moment in time as the data which was used for training. This to prevent different weather conditions at the moment of data collection but also to try and eliminate model bias and to illustrate how the model performs on new data. However some overlap might be possible as images are relatively large in geographical size and the Netherlands is a relatively geographical small country. To give an idea of possible model bias the models will also be tested on an area with a completely different physical environment. A trained model will inherently be biased because of the type of environment it is trained on but this will serve the purpose of providing insight about to what extent that will interfere with the model generalizability.

Table 4.2 compares multiple important operating and technical aspects of seven widely used remote sensing satellites. For this purpose, it is necessary to know for which period data is available, what the spatial resolution of the data is, at what local time images are made and what the inclination is of the satellite's orbit and therefore the angle under which images are made and lastly also who operates a sensor. The earliest launched satellite of the ones shown in table 1 is the GeoEye-1 satellite, although still in operation by Maxar technologies it also launched new satellites later in time. Also not unique is the operation of multiple satellites in the same orbit, phased 180 (with two) or 90 (with four) degrees from each other to significantly reduce the temporal resolution and therefore the revision time. In regard to the spatial resolution Sentinel and Landsat operate on a medium resolution where Worldview, Spot, Pleiades, GeoEye-1 and Superview can be considered as high-resolution sensors. For segmenting and classifying road infrastructure one should consider that a too high resolution would not per se be necessary as the features of study do not require such resolutions to be classified. High resolution data for relatively lower resolution features would only unnecessarily put a burden on data storage and processing time. In orbital inclination the satellites can be considered as almost identical with only small differences but an important deviation can be seen in local overpass time. Worldview captures local images at around 13:30 in the afternoon while Spot does the same at 10:00 and the other three high resolution satellites at 10:30. This difference in time is crucial as using images captured at other times of day will cause a different shade because of the positioning of the sun and also the weather conditions at the times can be different from each other.

Table 4.2: Remote sensing satellites

Sensor Name:		***Sentinel 2 (A and B)***	***Landsat 8***	***Worldview 3****	***Spot 6-7***	***Pleiades 1 (A and B)***	***GeoEye-1***	***Superview***
Launch date:		June 2015 (A), March 2017 (B)	Feb. 2013	Aug. 2014	Sept. 2012 (6), June 2014 (7)	Dec. 2011 (A), Dec. 2012 (B)	Sep. 2008	Jan. 2018
Resolution:	**MSI.**	10m	30m	1,24m	6m	2m	1,84m	2m
	PANCH.	/	15m	0,30m	1,5m	0,50 m	0,50m	0,50m
Local Overpass Time:		10:30	10:00	13:30	10:00	10:30	10:30	10:30
Orbital inclination:		98,62 Degrees	98,2 Degrees	97,97 Degrees	98,2 Degrees	98,2 Degrees	98 Degrees	97,49 Degrees
Data Quantization:		12 Bits, stored as 16 Bits	16 Bits	11 Bits (14 Bits SWIR)	12 Bits	12 Bits	11 Bits	11 Bits
Operator:		ESA	NASA/USGS	Maxar Technologies (USA)	Airbus Defence (Spot 6), Azercosmos (Spot 7)	CNES (France)	Maxar Technologies (USA)	Beijing Space View

** The Worldview-4 was also launched but was lost due to mechanical problems*

Source: Satellite Imaging Corporation, ESA & USGS

Satellites like GeoEye-1 which are launched a very long time ago would mean that the database to select a period from would be relatively extensive. A more recently launched satellite like Superview would mean the timeframe to get data from is rather limited. Sentinel 2 and the Spot mission offer a good combination for this research as their respective local overpass time are similar to each other. This ensures that the conditions under which data is collected will be similar and also that the images are revisited at a comparable interval, again benefiting parallel capturing conditions but also an adequate renewal of current data. Pleiades would in this aspect be even more suitable but the very high resolution of this satellite would be somewhat too high for the purpose of segmenting and classifying roads and mainly put a large pressure on storage capacity.

Spot is a commercially owned and operated satellite service which means data is not per se openly and free of charge accessible. However, the Netherlands Space Office (NSO) provides a collection of Spot 6/7 images of different parts of the Netherlands, taken in several months over the years and in multiple formats (8bits, 12 bits and unmodified). These are easily transferrable to a local drive via their ftp-client, making retrieving this data fairly accessible.

Data is available for the years 2014-2016 with images being of all regions around the Netherlands. Sentinel 2 data is openly accessible via a multitude of services like the Sentinel Access Hub and Google Earth Engine, with data available over a longer period of time and free of charge. As Spot data is only available via the NSO from the period 2014-2016 Sentinel 2 data used for the super resolution model and subsequent extraction model should be of the same time period. As the Spot data is provided as a database without accurate metadata images need to be downloaded before their quality and usability can be assessed but for Sentinel 2 portals like the Sentinel Access Hub allow to preview and also filter data based on for example cloud cover so the quality can be determined beforehand.

4.4 Hardware and software

4.4.1 Data storage

A technical requirement that needs to be thought about is the need for storage space. This comes from the need for many different and also high-resolution satellite imagery and also super resolution and feature extraction data which requires storage as well. The Ministry of Defence provides an external SSD disk with 2 terabytes of storage capacity which should be enough to allocate all the data needed for this proof of concept.

4.4.2 Hardware

The hardware requirements of this research will mainly concern the scale of the proposed analysis and therefore the processing power and accompanied by that also processing time. Although there is time to conduct the empirical part of this research it is expected that applying a super resolution and deep learning extraction model on multiple images of parts of the Netherlands will take up a lot of processing time to be accomplished. For reference the average satellite image as found on the Netherlands Space Office serve can be up to twenty gigabytes of data already for Spot 6 images which means that for training the model already a lot of data needs to be processed before the actual input data is even considered, which although lower in resolution will take up a lot of storage space and therefore processing time as well.

The Ministry of Defence provided a heavy duty graphical laptop for this purpose, which is given on loan for the duration of this research. As processing power and especially processing time can be an important factor for training and running models the technical specifications of the laptop are noted in table 2.2

Table 4.3: Specifications graphical laptop

Brand:	Hewlett-Packard
Type:	ZBook Fury G8
Disk space:	500 GB
CPU	11th Gen Intel Core i7-11850H
GPU:	NVIDIA RTX A5000
RAM memory:	64 GB
Release date	June 2021

Source: Hewlett-Packard

4.4.3 Software

How to perform the technical operation of super resolution and subsequent analysis tasks is not per se an issue of what is the best software package (as there are many available) to do that but what serves the research goal and the own user needs.

*Table 4.4: Examples of super resolution modules**

Name	Geo pre-trained	Coding needed	Licensing
ArcGIS.learn API	Yes	No	ESRI
SR4RS	Yes	Yes	Open-Source
Super Resolution API	No	Yes	DeepAI
Open CV	No	Yes	Open-Source

**These are a few examples but in no way represent an all-encompassing overview of what is available in software solutions*

Source: Own research

The examples as shown in table 4.4 represent a simplified comparison on some of the relevant attributes to consider for this research. SR4RS was used to perform super resolution as shown in the example of figure 1.1 in the introduction of this research and proves that it is suitable for geo analysis. Super Resolution API and Open CV are just some of the non geo pre-trained solutions available and especially Open CV is often applied as it is open-source just like SR4RS. For this research the use of ArcGIS.learn API is purposeful but that definitely not makes it the best choice in any case. In this case The Ministry of Defence also makes use of ArcGIS and as the purpose of this research is to provide a proof of concept and also improve the knowledge on this topic within the geo domain a complementarity in used software would be a desirable goal. The availability of information and tutorials on the possibilities of using super resolution and deep learning within the ArcGIS environment makes it a favourable solution but also it makes using it by other researchers who want to do something similar insightful and accessible. As ArcGIS Pro is a user interface that accesses specific scripts to perform operations it eliminates coding (almost completely) by the user. It should however still be noted that Esri is a commercial provider and having a license for their software is therefore a crucial requirement.

Since end 2021 the ArcGIS Deep Learning methodologies are integrated in ArcGIS Pro, which makes it possible to perform all the required operations within one and the same environment instead of needing to import/export data between separate environments. Important to state in regard to literature is that ArcGIS Pro enables deep learning using ResNet, which is pre-trained on over 1 million images of the ImageNet dataset and outperformed many other model architectures for image analysis (Hu et al., 2016).

4.5 Empirical design

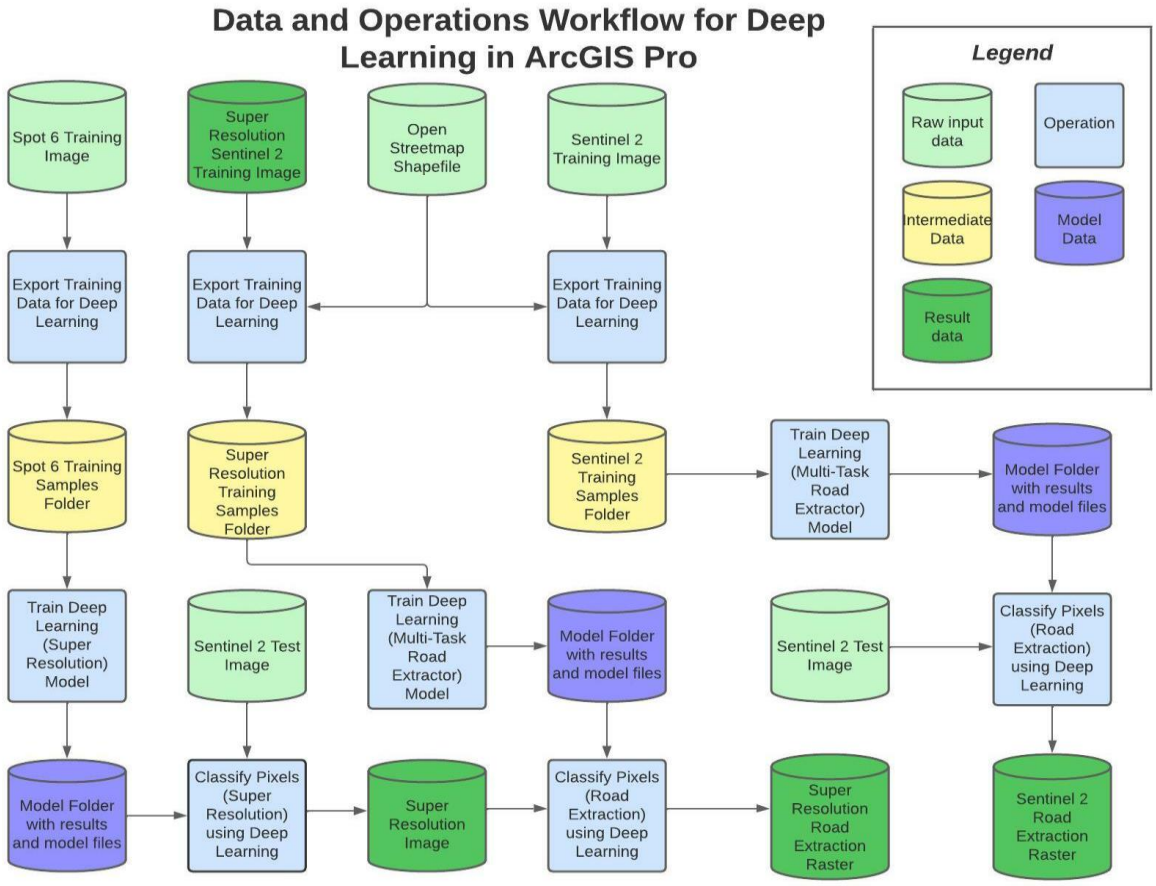
The previous paragraphs of this chapter elaborated on methodological aspects of the required data, the area from which training data is extracted, the source from which data can be acquired and also the resources needed to perform the proposed deep learning analysis. Combining all these aspects attribute to the foundation of this research and what is needed to go from initial input data to the actual results that will provide insight into the model performance and will eventually contribute to answering the research question.

Figure 4.2 provides a schematic overview of the steps needed to transform the initial input data needed to train the deep learning models towards the test data for which the models are

applied and provide the research results using Esri's ArcGIS Pro. The sequence starting at the Spot 6 training image and ending in the super resolution image comprises the super resolution modelling section while the rest of the schema is about the steps needed to perform road extraction analysis. Operations represent actual geo-processing tools as how they are named in ArcGIS Pro. Special attention should be paid to the super resolution training image. A trained super resolution model is in this case applied on the research area to create image data that is used to train the super resolution road extraction model. Although seemingly similar to the super resolution image that is the result of the super resolution model it is from a different area (training instead of test area) and serves a different purpose in the analysis.

Appendix A contains a guide with the in-detail technical description of how to prepare data for training the models and the parameters that are relevant as part of both the model training and model application. Although these are all left on default settings as part of the proof of concept they are added for reference purposes. These contain a level of detail not attributing directly to the research questions and therefore not elaborated on in this research integrally but were added as appendices for scientific transparency. Wherever relevant as part of the results chapters the information on technicalities will be discussed.

Figure 4.2: Data processing workflow



Made by: Yannick Bouten

4.6 Place of research

Different institutions are involved in organising and supervising this research.

In the first and formal place this research forms the thesis of the master Geographical Information Management and Applications. The scientific knowledge on the thesis process and this specific topic is at the University and the guidance as part of this research is also a task of the University.

Because the Dutch Ministry of Defence is interested in the possibilities super resolution could provide for their tasks and branches that are involved in the geo-domain, they have a stake in this research as well although their formal role is limited. Their role in the research is more practical and hands-on in the sense that they try to facilitate and make this research possible by providing resources. They provide office space to work at, they provide hardware to enable the actual empirical phase of this research and also their practical knowledge and contacts on this topic and within the geo community can be beneficiary to this research. Being employed as an intern at the Ministry as well as being a student at Utrecht University enables to make use of all those resources for the duration of the research.

The Ministry's involvement is not via an operational branch but is accommodated via the Defence Materiel Organisation, an executive organisation that operates in services of the regular branches of the armed forces. As part of the Defence Materiel Organisation the Joint-IV (Informatie-Voorziening)-Commando (JIVC) is the Ministry's main IT-branch. Residing under the IT-branch is KIXS (Kennis, Innovatie, Experimenten en Simulaties), a department focused on practical and scientific experiments involving different technology applications, again in service of other branches and services within the Ministry and the armed forces. KIXS is the department for doing research that can be of interest and benefit of other and more operational branches of the Ministry and is the department where this research will be done as well.

This chapter discussed the methodological design for both super resolution and feature extraction, by elaborating on the features of interest, the area where the models will be trained on, the source of satellite imagery and how other resources like hardware and software are organised. These aspect all come back in the empirical design on how the actual modelling will be done and how data and the different methodological operations relate to each other in the overall workflow. The paragraph above also briefly discussed the context of where this research will be done and which actors are involved and what their responsibilities are.

The next two chapters will discuss the results of the actual training and application of the deep learning models, both super resolution and road extraction. The split in two chapters was done to maintain structure as super resolution data will be used to train the super resolution road extraction model so this makes sense from a chronological perspective and also the split is useful as metrics to evaluate the models can be similar to each other. In both chapters first the results of the training and evaluation metrics will be presented and the second paragraph will be a visual evaluation by applying the models on a test area.

5. Results super resolution

Training a super resolution model (or any deep learning model for that matter) is a research in itself as it is an experiment into how the combination of model type, the training data and training parameters lead to the most optimal model result.

Before going into the results and quality assessment of when super resolution is applied on a test area of interest it is first important to evaluate the metrics in regard to how a super resolution model is trained. The final metrics of each trained model are summarised in a metrics report together with samples from the validation dataset, as shown by an example in appendix B. The raw data on all the evaluation metrics at each epoch can be found in appendix C.

ArcGIS Pro displays multiple metrics in regard to the training process; LR (Learning Rate), training and validation loss, pixel colour value, PSNR (Peak-Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). The super resolution models were trained with different amounts of input data to assess model quality with these metrics and to determine if there was a need for more input data to boost the quality. This led to models being trained with 1000, 2000, 3000, 5250, 10500* and 21000** images. As described before labels for the images were created by down sampling the original images by a factor four to create images at 10 metre resolution. Important note is that for the amount of images up to 5250 the images are completely unique. 5250 is the maximum amount of input image samples that can be created from the research area with a chip size of 512 by 512 pixels. This is relatively large but is chosen to ensure that for the analysis the Sentinel 2 model can have the same image size (although the difference in resolution) as then a chip size of 128 by 128 pixels can be used but the geometrical size is still the same. 10500* images can be created by applying an image augmentation on the maximum amount of 5250 by rotating each image 180 degrees, therefore doubling the amount of input images that can be extracted from the research area. The same is done for 21000** images by applying a rotation of 90 degrees.

5.1 Metrics super resolution

5.1.1 Learning rate

Table 5.1: Learning rate

Model images	Minimum LR	Maximum LR
1000	2.29e-05	2.29e-04
2000	2.29e-05	2.29e-04
3000	9.12e-06	9.12e-05
5250	1.32e-05	1.32e-04
10500*	1.20e-04	1.20e-03
21000**	6.31e-06	6.31e-05

**: This model contains 5250 completely unique input images but the images have been rotated 180 degrees to double the amount of input images*

***: This model contains 5250 completely unique input images but the images have been rotated 90 degrees to quadruple the amount of input images*

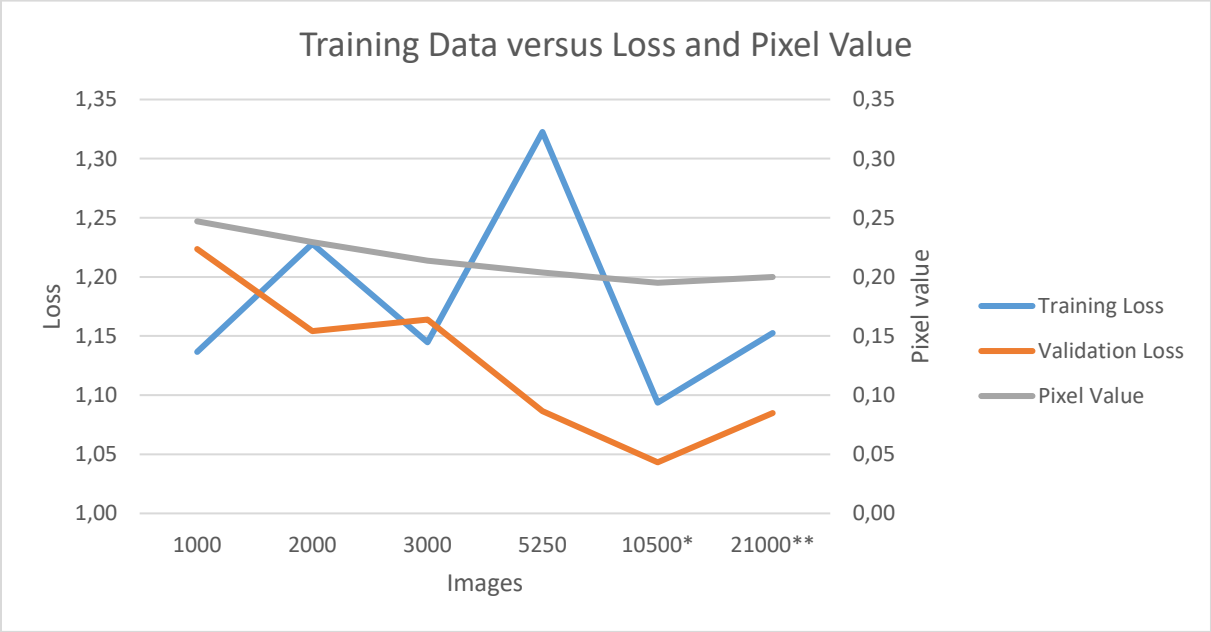
The learning rate as a metric displays the rate at which changes are made to the model at each iteration in order to eventually optimize the model towards a stable estimation of the model parameters leading to a minimal outcome in loss values. The learning rate is between

0 and 1 and the actual determination of the parameter given to the model can be described as an equilibrium between relatively large/small adaptations, shorter and longer processing time and the (in)stability of the model while learning from the training data. A sliced learning rate is used by the model to model the first layers with the minimum rate and the last ones with the maximum rate, where the layers in between will have a rate within that value range.

In table 5.1 the learning rate is displayed for each individual model. The learning rate was plotted by the training function and extracted the most optimal rate (learning rate in relation to model loss) and used that as the input variable. This indicates a stable training progress but also would result in a relatively extensive processing time to complete training. For 10500* images the learning rate is relatively higher compared to the other instances but overall the learning rate is comparable for the different amounts of input images used for each model. In each model the learning rate is increased by a tenfold after the first section is completed to make larger improvements to the model.

5.1.2 Loss and pixel value

Figure 5.1: Loss functions



**: This model contains 5250 completely unique input images but the images have been rotated 180 degrees to double the amount of input images*

***: This model contains 5250 completely unique input images but the images have been rotated 90 degrees to quadruple the amount of input images*

Loss as a metric depicts how precise the model is in its prediction for a single test case. In an optimal scenario the model should be able (because it learns over time) to reduce the loss by iterating through the dataset. For an individual model the loss function would be inversed linear and tries to approach zero, which if it occurs results in a perfect prediction by the model. Figure 5.1 displays the average loss value for each model after its final iteration, where the distinction is also made between training, validation and pixel loss. Although some fluctuations are visible with varying amounts of input images the loss stabilizes for all three types. Training loss and validation loss show similar values in the figure indicating that the model fits rather well on the training samples and is able to perform in a similar fashion on the validation dataset, assuming a good fit. However it should be noted that the loss for both

metrics on an absolute scale is still rather high and seems to stabilize on a level further from zero than what would be desirable. Overfitting might even be the case as the training loss is diverging from the validation loss at 21000** training samples in comparison to the instances with less input images although it cannot be concluded yet that it might be significant.

In the original Sentinel 2 image colours are displayed in a standard 8-bits RGB format. Average pixel value is a combination of red, green and blue with a value between 0 and 255 (as that is the maximum integer value for 8 bits). In the super resolution image these individual values and therefore the displayed average value are normalised on a 0 to 1 scale. As can be seen in figure 5.1 the pixel value is on average around 0,2. The original Sentinel 2 image had individual average RGB band values of 52, 69 and 78 and combining these band values into one value results in an average grey colour for the image. Normalising this value for Sentinel 2 results in a normalised value of 0,26. Evaluating the pixel values for the super resolution models with the average normalised value for Sentinel 2 means that the super resolution models predict a lower normalised value, meaning that the image is darker on average but the difference in value is not a large offset in comparison to the original.

Figure 5.1 shows that the model with 10500* amount of input images results in the lowest training and validation loss and also the difference between them is the lowest of the models trained, resulting in the best possible fit and also pixel loss is the lowest for this model. Therefore it would lead to the best results when evaluating based on these three metrics.

An important observation that should be made that although training loss is for a majority of the test models higher than the validation loss (indicating a slightly possible overfitting), in general both losses decreased when the amount of input images increased. A consideration to be made is that after the model with 5250 input images the amount is still increased but not necessarily the amount of completely unique images, as the amount is increased by image rotation and therefore some similarity between images will still exist. Evaluating the initial decrease as a result of image augmentation (although further augmentation might cause an increase in loss) one could reason that if the amount of completely unique images is further increased after 5250 the loss will decrease as well.

Figures 5.2 up to 5.4 further elaborate on the loss functions as shown in figure 5.1 by displaying values for the training loss, validation loss and pixel value at each iteration stage for the models. The loss is calculated as the MSE (Mean-Square Error Loss) by the following function;

$MSE = 1/n\sum(y-\hat{y})^2$

Where y = actual pixel value

And $\hat{y}$ = predicted pixel value

In a well-functioning model the loss should be reducing at each iteration step until a stable value is reached. As the MSE formula is quadratic this stability is not set to last and at some point continuing iterating through the dataset will lead to increase loss indicating overfitting of the dataset. The absolute differences in value for the losses between the models is relatively small but what is interesting to notice is that more input data does not necessarily reduce loss in this case. The training loss in figure 5.2 shows the same trend for all models but the initial training loss and also the progression for 10500* images is better than for the model with 21000** images. The same can be noticed for the validation loss in 5.3 where again the differences are small on an absolute scale and the loss curve is similar to the curve for the training loss indicating a good model performance. When the curves for both the training and validation loss stabilize the validation loss is below the training loss for each model indicating the model does not perform well on the training set but also on the validation data. Figure 5.4

plotting the pixel value curves show the same trends in improvement as can be seen for the training- and validation loss again indicating a desirable fit for the model on in this case the pixel level as differences in pixel values are reduce as good as possible. These plotted curves illustrate that performing image augmentation can be favourable for reducing training and validation loss. Pixel value also decreases over time but on an absolute scale the decrease is in a different order of magnitude. As also explained earlier in this paragraph the absolute difference in comparison to sentinel 2 imagery is limited. Pointed out earlier in the initial results on the models and reoccurring here is that the training loss is here higher than the validation loss, indicating a possible fitting problem. Underfitting is not per se the case as the loss decreased significantly during modelling but overfitting is also not what is directly depicted by figure 5.2 and 5.3, where although the training loss is higher than the validation loss the gap between them is stable and does not seem to increase over time.

Figure 5.2: Training Loss

Figure 5.3: Validation Loss

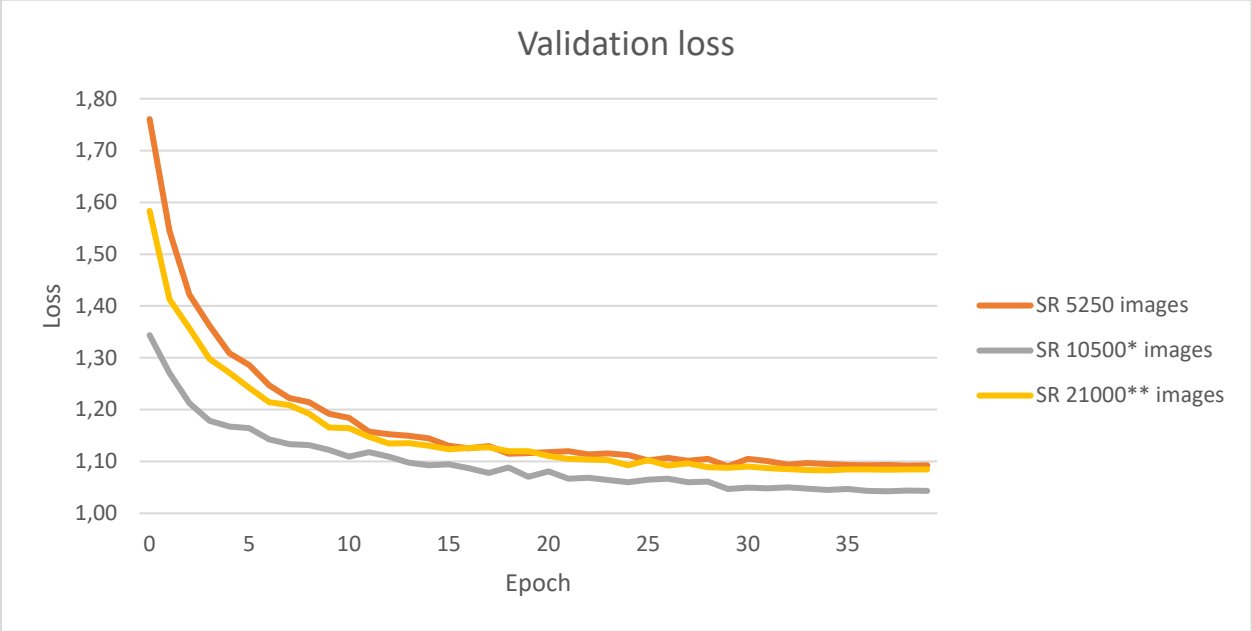
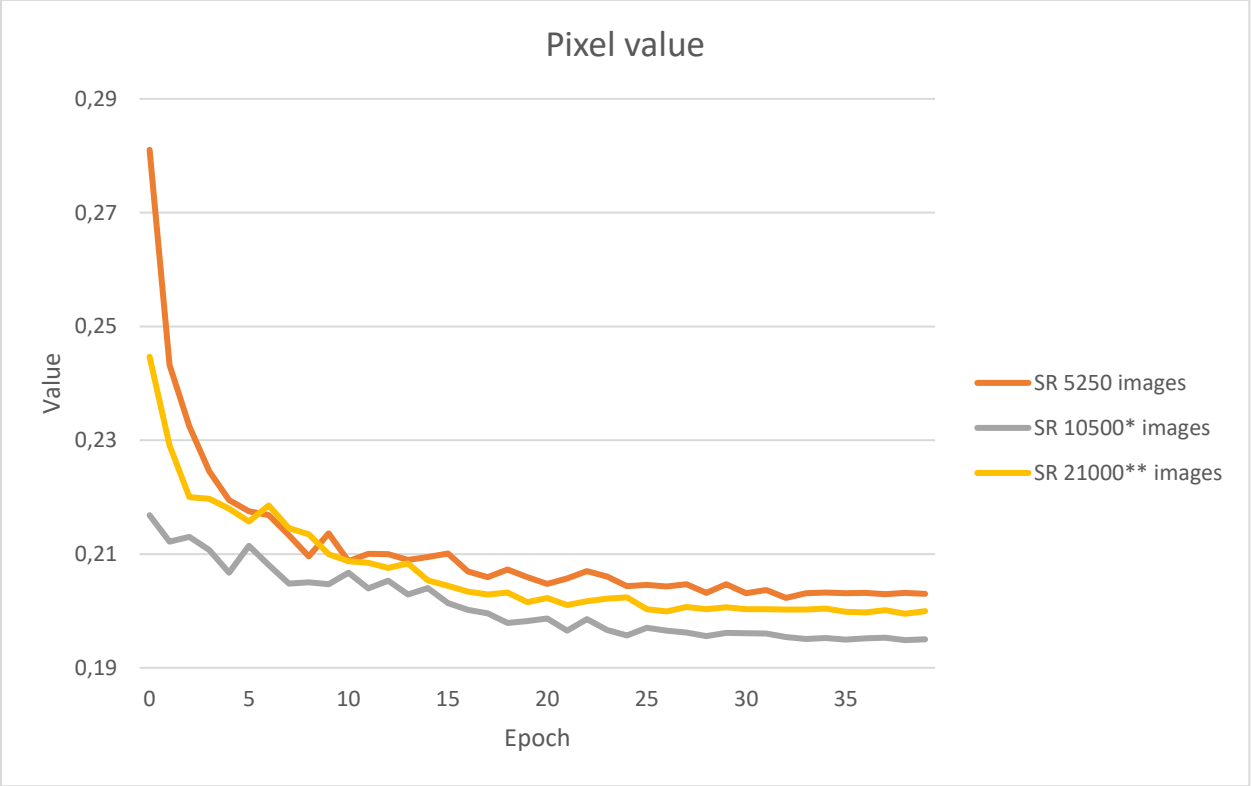


Figure 5.4: Pixel value



**: This model contains 5250 completely unique input images but the images have been rotated 180 degrees to double the amount of input images*

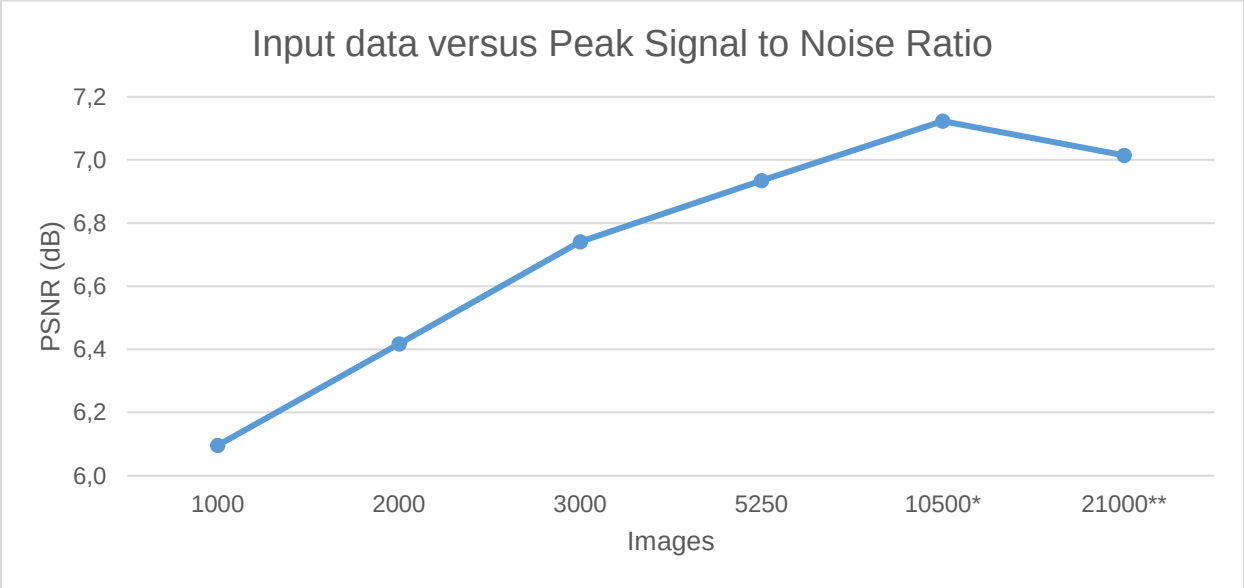
***: This model contains 5250 completely unique input images but the images have been rotated 90 degrees to quadruple the amount of input images*

5.1.3 PSNR and SSIM

PSNR and SSIM measure the overall quality between two images. While PSNR is an absolute measure that is displayed in dBs (decibels) the SSIM is a value between 0 and 1. Both are computed for the validation segment of the dataset.

PSNR in computer vision represents the ratio between the signal power and the corrupting noise, illustrating the efficiency of the processing task applied on the validation data. The maximum PSNR is defined by PSNR=20*log(max pixel value), which for 8 bit imagery is 255 resulting in a max PSNR of PSNR=20*log(255) = 48dB. The PSNR values displayed in figure 5.5 for each model indicate a relatively low PSNR which means that the noise power is relatively strong compared to the signal power. Increasing the amount of input images results in an increase in PSNR and therefore a better processing quality but the increase is limited in relation to the maximum possible value for PSNR, which although being an ideal scenario in which the signal power is as strong as possible compared to the noise leaves room for improvement in processing quality. Important to note is that for the maximum amount of input training images for the research area (21000** images) results in a small decrease in PSNR in comparison to the previous data step which would indicate that noise becomes more prevalent at this point when the amount of input images would be increased.

Figure 5.5: PSNR

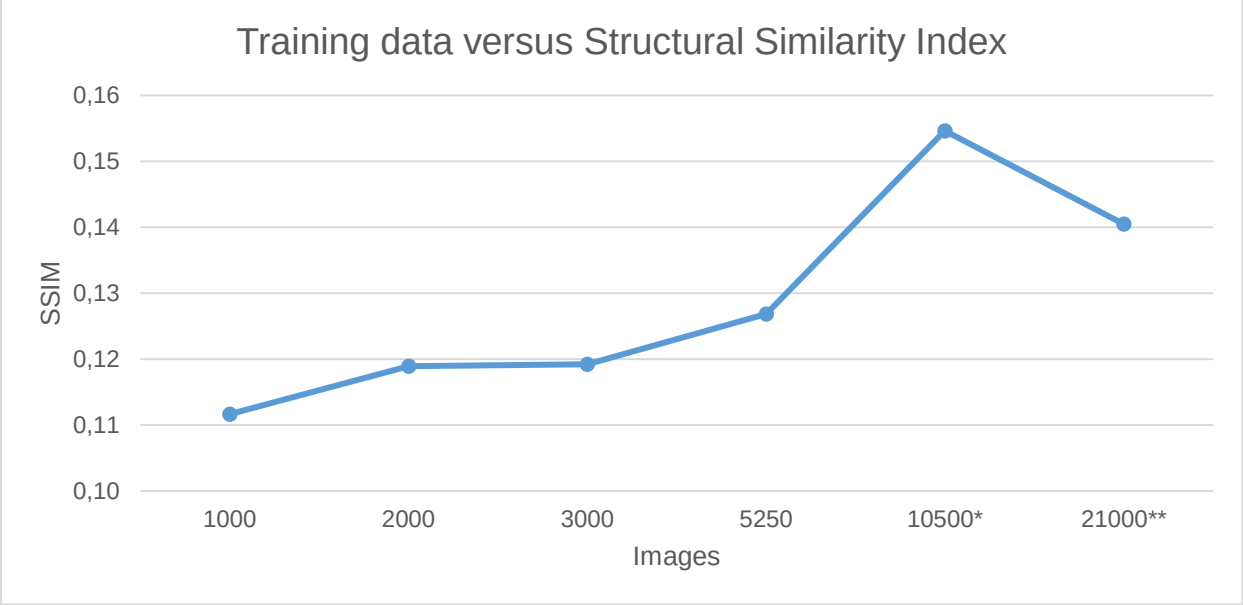


**: This model contains 5250 completely unique input images but the images have been rotated 180 degrees to double the amount of input images*

***: This model contains 5250 completely unique input images but the images have been rotated 90 degrees to quadruple the amount of input images*

SSIM is normalized between 0 and 1 (the actual value range is -1 to +1) and displays the ratio to which two images are completely similar (1) or completely not similar to each other (0). The calculation for SSIM can be broken down in three sub-parameters; luminance, contrast and structure. Although not individually reproducible as values in the metrics report they are relevant to take into consideration as they can be assessed visually, which will be the focus of the next subparagraph. SSIM values for the different models as shown in figure 5.6 shows a low and relatively stable SSIM for several models with a small increase towards the model with 10500*, but a decrease when the amount of input images is further increased to 21000**.

Figure 5.6: SSIM



**: This model contains 5250 completely unique input images but the images have been rotated 180 degrees to double the amount of input images*

***: This model contains 5250 completely unique input images but the images have been rotated 90 degrees to quadruple the amount of input images*

5.1.4 Performance matrix

The previous paragraphs discussed several metrics on evaluating super resolution but for reference purposes table 5.2 will provide a complete overview of the metrics for each model. This to assess which one performs best and to provide insight about the numerical values for each different model's metrics as these cannot be precisely dissected from the plotted curves for each model.

Table 5.2: Performance metrics matrix

Model	Training	Validation	Pixel	PSNR	SSIM
5250	1,32	1,09	0,20	6,93	0,13
10500*	1,09	1,04	0,19	7,12	0,15
21000**	1,15	1,08	0,20	7,01	0,14

The table again illustrates that the super resolution model with 10500* images results in the best performance based on these evaluation metrics. Training, validation and pixel value are reduced to the best extent by this model and the PSNR and SSIM are higher compared to the others indicating less noise interference and a better similarity in images based on contrast, luminance and structure. It also shows that augmenting images has a positive effect in performance as the result in evaluation metrics increased compared to the scenario without augmentation (5250 images). Relativizing is important as further augmentation (21000** images) might be on the other hand counter-productive towards model performance as the matrix shows a small increase in loss and reduction in PSNR and SSIM compared to 10500* images.

5.2 Visual evaluation super resolution

The metrics as discussed in the previous paragraph provide computational tools to evaluate the training, validation and test phases. Images are the result of the modelling based on these metrics but also serve a powerful purpose as visual examples that to some extent illustrate the influence of these metrics on the result.

All model trainings reserve maximum 10% of data for validation, and a small batch of it is attached to the training report to display input images (the labels), target images (the original images) and the prediction/validation images. A small selection is shown in figure 5.7 for the models trained with 5250, 10500* and 21000** images. The tile size for the target and prediction images is 512x512 pixels which would result in a metrical size of 1,6384 square kilometres. As the focus of the research is eventually on detecting road infrastructure the batch in figure 5.7 all show some form of infrastructure to already give insight into how a super resolution model handles this type of environment.

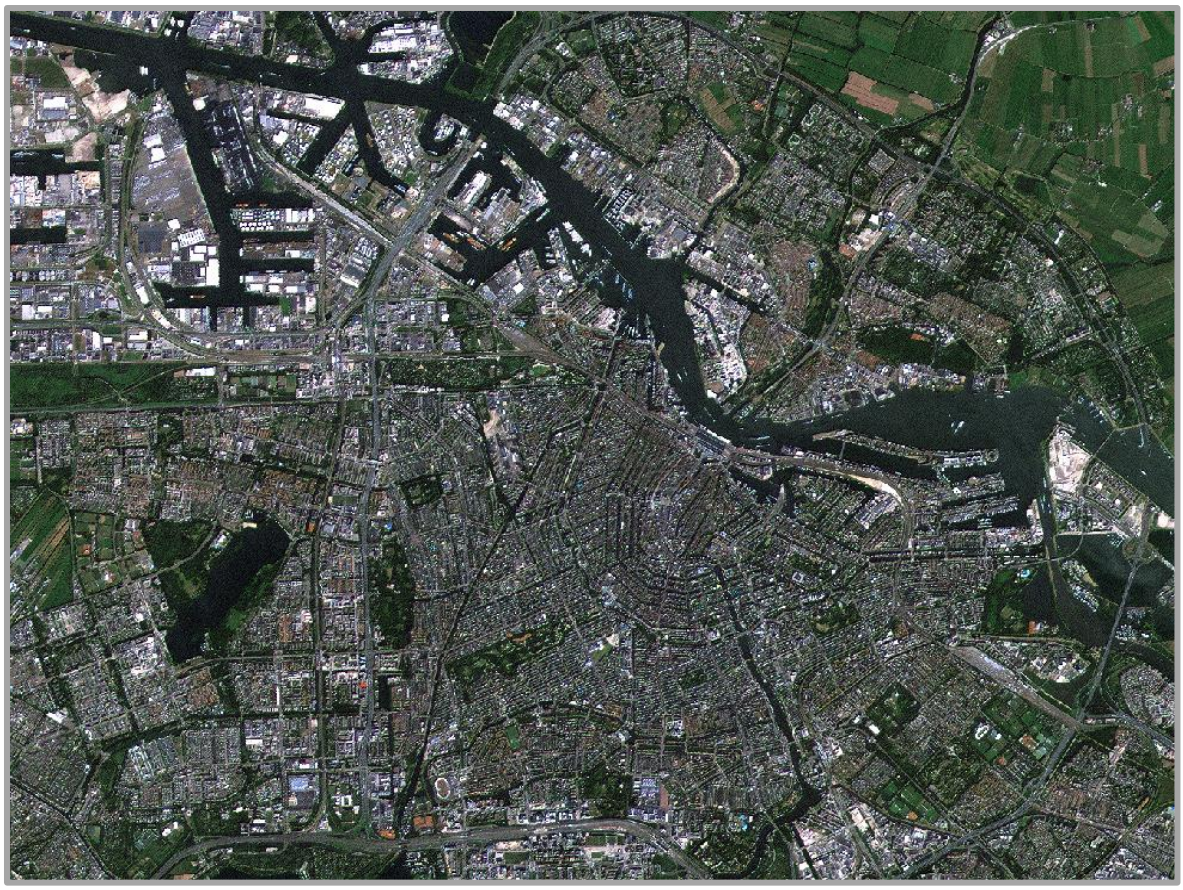
Figure 5.7: Input/Prediction/Target images for 5250, 10500 and 21000** images*

Input **Prediction** **Target**

What the batch shows is that their shapes can be preserved rather well by the model. Visually they are still recognizable as such and the same goes for other types of environment like neighbourhoods, agricultural land, vegetation and so on. What should be noted and was also pointed out by the metrics is the noise in the images. Achieving the level of detail and cleanliness of the target images might be a best case scenario but one can see that the prediction imagery falls somewhat short of that. This was also determined by the validation loss and the PSNR and SSIM but is visually supported by these images. Putting the focus on road infrastructure illustrates that outstanding pixel values within these segments like white lines and objects like roundabouts are somewhat difficult for the model to try and predict them as accurately as possible.

Figure 5.8: Ground truth Sentinel 2 (upper) vs 10500 (lower, next page) super resolution image*

Figure 5.9: Close up view of figure 5.8 on Amsterdam. Ground truth Sentinel 2 vs 10500 images (next page) super resolution imagery*



The previous figures 5.8 and 5.9 show the predicted super resolution imagery in comparison to the Sentinel 2 imagery of the same area, which serves as both the input data for the super resolution model and also as a ground truth reference dataset. What immediately shows up when comparing the imagery is the difference in contrast and luminance. The super resolution image shows a starker contrast between different features in the built up environment while the overall image has a lower luminance compared to the Sentinel 2 ground truth. Overall the image integrity seems to be preserved by the super resolution model as the distinct features like built up environment, road infrastructure and other features of interest (like the port area in figure 5.9 in the north-western part of the image) can still be recognized as such in the super resolution imagery. In general the super resolution model is able to predict super resolution correctly in a way that it preserves the general features as shown in the ground truth but the low luminance and high colour contrast and the metric evaluation which pointed out low structural similarity and high noise disturbance require a more in depth analysis.

Figure 5.10: Ground truth Sentinel 2 versus 10500 super resolution image*

*Figure 5.11: Pixel Level samples of 5250 and 21000** super resolution imagery*

Evaluating the visual result at a pixel level as shown in figures 5.10 and 5.11 supports the earlier result from the training phase, which showed that the super resolution model is able to preserve the shape and structural integrity of objects in the image. The colour composition in the super resolution image somewhat darkened in comparison to the Sentinel 2 image. What is clearly visible at this level is that although the model predicts well and reduces loss it is influenced by the noise and low structural similarity as was also pointed out by the metric evaluation. The resolution is increased by a factor four but for low scale objects a correct prediction might be difficult due as pixels representing it can have an offsetting pixel value that can impede an accurate representation in the image or one that is needed to analyse these objects correctly. To assess the influence of this on analysis tasks the next chapter will evaluate the road extraction task that has been executed on these imagery to assess their quality and fit for use.

6. Results road extraction

Training a road extraction model shows some similarities with training the super resolution model. The tools are similar, some evaluation metrics are the same but the parameters are different and several metrics specific to the extraction model are also introduced. Most important is the goal and purpose for which the model is trained and this

Before going into the results and quality assessment of when road extraction is applied on an area of interest it is first important to evaluate the metrics in regard to how a road extraction model is trained. Also it is important to be aware of how super resolution relates to the road extraction as the super resolution output is used as input for road extraction.

ArcGIS Pro displays five main metrics in regard to the training process; learning rate, training and validation loss, accuracy and MIoU (Mean Intersection over Union). The road extraction models where trained with different image input datasets. One with Sentinel 2 data (on 10 metre resolution) and three models with super resolution data based on 5250, 10500* and 2100** images (all with 2,5 metre resolution). The final metrics of each trained model are summarised in a metrics report together with samples from the validation dataset, as shown by an example in appendix D. The raw data on all the evaluation metrics at each epoch can be found in appendix E.

Important comment to make beforehand is that all the road extraction models triggered the stopping criteria, meaning that continuation of training would lead to overfitting by the model. As that would making a qualitative comparison between models difficult due to the altering amount of epochs for each model a dashed line illustrates the curve for each metric as if all of them iterated for the duration of 20 epochs. A solid line illustrates the curve for the actual model that was used for testing the road extraction model and showing visual results, and a diamond representing the point where the stop clause was triggered and the model stopped improving. In regard to the metrics the stopping criteria as mentioned before is set by default to trigger when the validation loss did not improve for 5 epochs with a threshold value of 0,001.

6.1 Metrics road extraction

6.1.1 Learning rate

Just as in training a super resolution model an extraction model improves when iterating through the dataset and makes improvements based on the learning rate. The logic assumption would be that as the learning rate for the super resolution model has been auto-extracted by ArcGIS Pro this can be done for the extraction model as well. However this is not the case. Relying on the auto-extractor results in a relatively high learning rate (about 10 times higher in comparison to the super resolution model) and although the task at hand is different the visualization of metric curves for the training and validation loss indicate an unstable model where it cannot be assumed that after the last iteration a somewhat stable loss is achieved. The choice has therefore been made to reduce the learning rate to a set value of 0,0001, safeguarding a more stable modelling with small step improvements over time although this leads to a longer training time. The plots for the road extraction models can be found in appendix F to illustrate the instability the model in the case of relying on an auto-extracted learning rate, indicating a poor fit on the validation data and therefore giving more reason to try and reassess the model's initial parameter in order to create a stable and better performing model.

6.1.2 Loss

Figure 6.1: Training loss road extraction

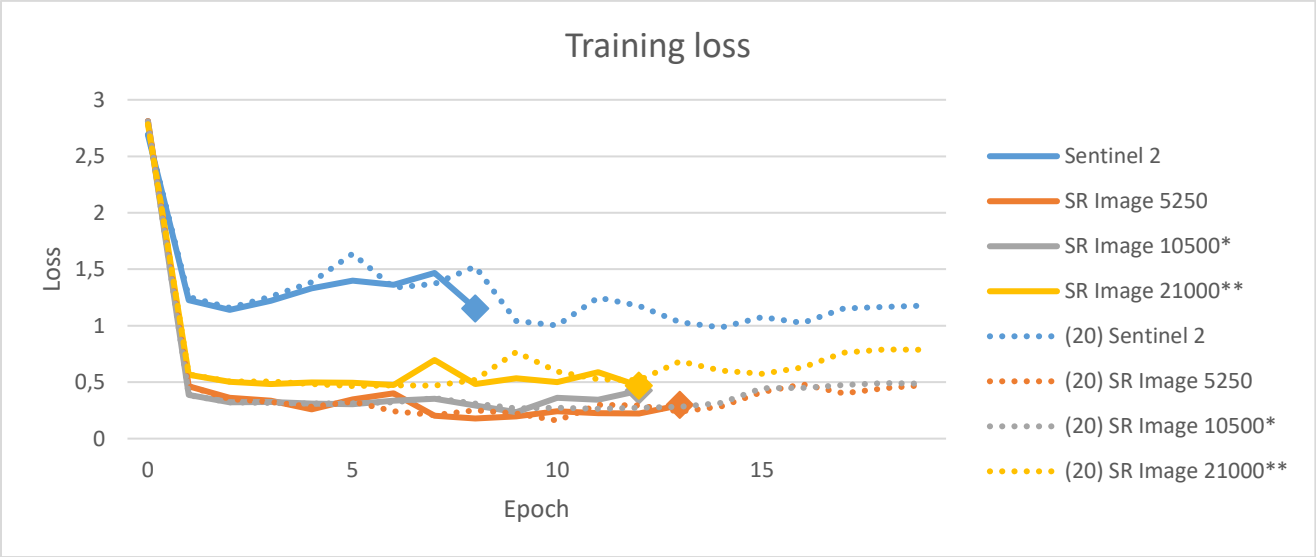
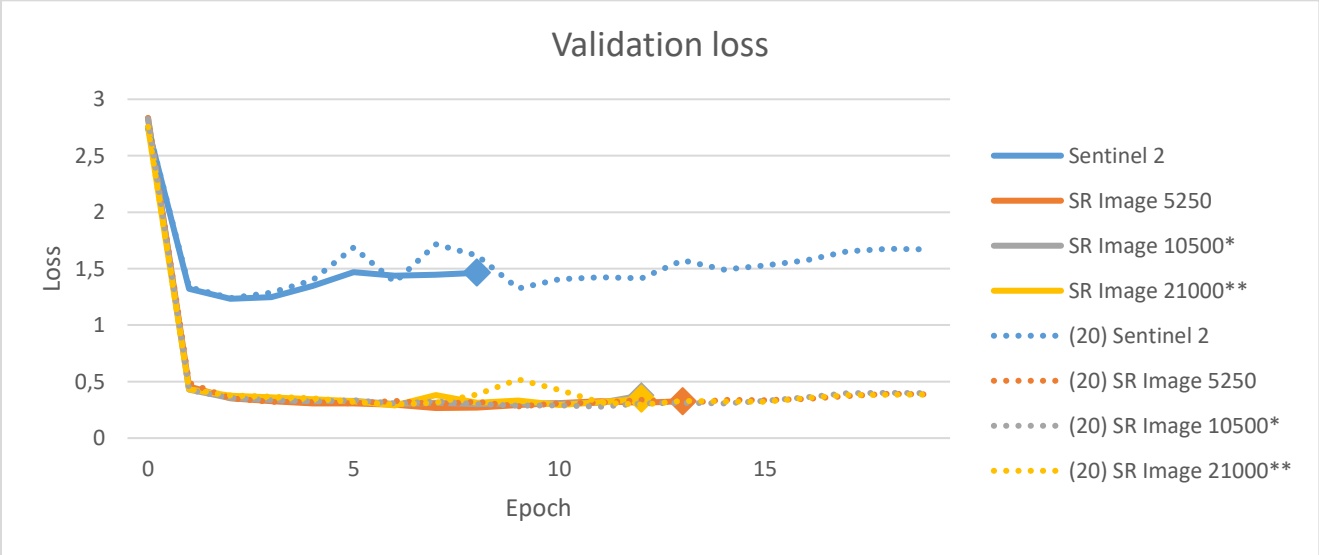


Figure 6.2: Validation loss road extraction



**: This model contains 5250 completely unique input images but the images have been rotated 180 degrees to double the amount of input images*

***: This model contains 5250 completely unique input images but the images have been rotated 90 degrees to quadruple the amount of input images*

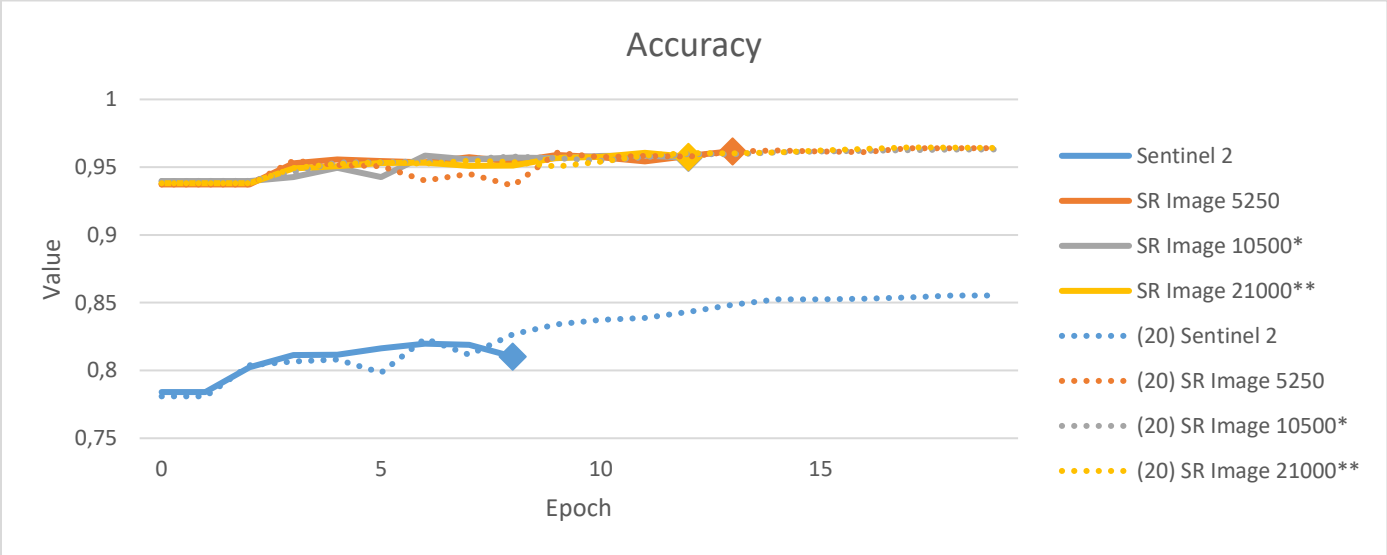
Both the training loss in figure 6.1 and the validation loss in figure 6.2 use the same calculation method as for super resolution. Meaning that loss should be seen as an MSE value indicating the mean-squared difference between predicted and actual values for in this case whether an area is a road or not. A difference with super resolution is that in this case the training of different SR models is also compared with Sentinel 2, which serves as a baseline for road extraction. What immediately stands out is definitely a result of the relatively higher learning rate compared to super resolution (although this was already altered as pointed out in paragraph 6.1.1) which results in the sharp decrease in loss during the first epoch. Comparing these loss curves with the previous auto-extracted rates as shown in appendix F

however shows that apart from the steep decrease the model is rather stable because of the low rate and only small changes are being made while iterating. This results in that after the initial decline in a somewhat stable loss value is already achieved, with some minor variation. The 20 epoch curves for the super resolution models even shows that increasing the amount of epochs increases loss in comparison to the point where the stop clause would otherwise be triggered, indicating possible overfitting. This is especially the case for the training loss in figure 6.1 and applies to a lesser extent to the validation loss in figure 6.2, although there the case for stopping the modelling early can be made because more iterations do not lead to model improvement and therefore saves processing time.

In general the use of super resolution based extraction models in this research leads to a significantly lower loss in comparison to the Sentinel 2 baseline, for both training and validation, where overall the difference in loss for validation is higher than for training. For validation the loss is more stable in value for the models while iterating and for the super resolution models also similar to each other while for the training data modelling is more unsteady and the differences between the super resolution models are more apparent.

6.1.3 Accuracy

Figure 6.3: Accuracy



**: This model contains 5250 completely unique input images but the images have been rotated 180 degrees to double the amount of input images*

***: This model contains 5250 completely unique input images but the images have been rotated 90 degrees to quadruple the amount of input images*

Accuracy as a metric tells something about the quality of performance, but requires an elaboration to be understood correctly. Annotated as a simple equation accuracy is calculated in the following way;

$$Accuracy = \frac{True_{Postive} + True_{Negative}}{True_{Postive} + True_{Negative} + False_{Positive} + False_{Negative}}$$

The deep learning model only classifies if a pixel is a road or not but to prevent ambiguity all the subsets that comprise this formula are summarised;

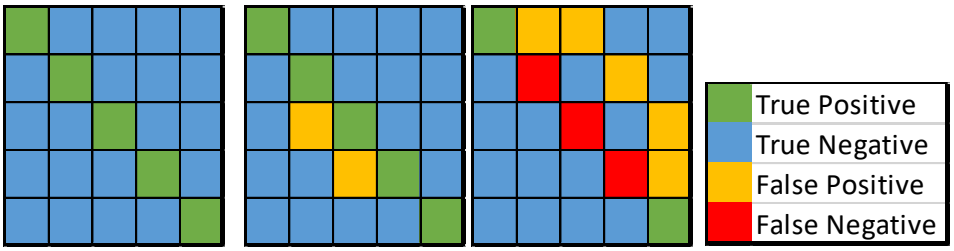
$True_{Positive}$ = *A ground-truth road pixel is classified as a road*

$True_{Negative}$ = *A ground-truth non-road pixel is classified as not being a road*

$False_{Postive}$ = *A ground-truth non-road pixel is classified as a road*

$False_{Negative}$ = *A ground-truth road pixel is classified as not being a road*

Figure 6.4: three road classifying scenarios (where left represents the ground-truth and mid and right are variant classifications)



In total there are 75 pixels to be predicted. As shown in green there are 12 true positive pixels. 54 blue pixels indicate a true negative. However 6 pixels indicated in orange indicate a false prediction, because the algorithm expected there to be a road in that pixel except the ground truth shows that there is no road. Also there are 3 pixels which should have been classified as a road but were not classified as such. For the three individual scenarios and the overall accuracy the results are;

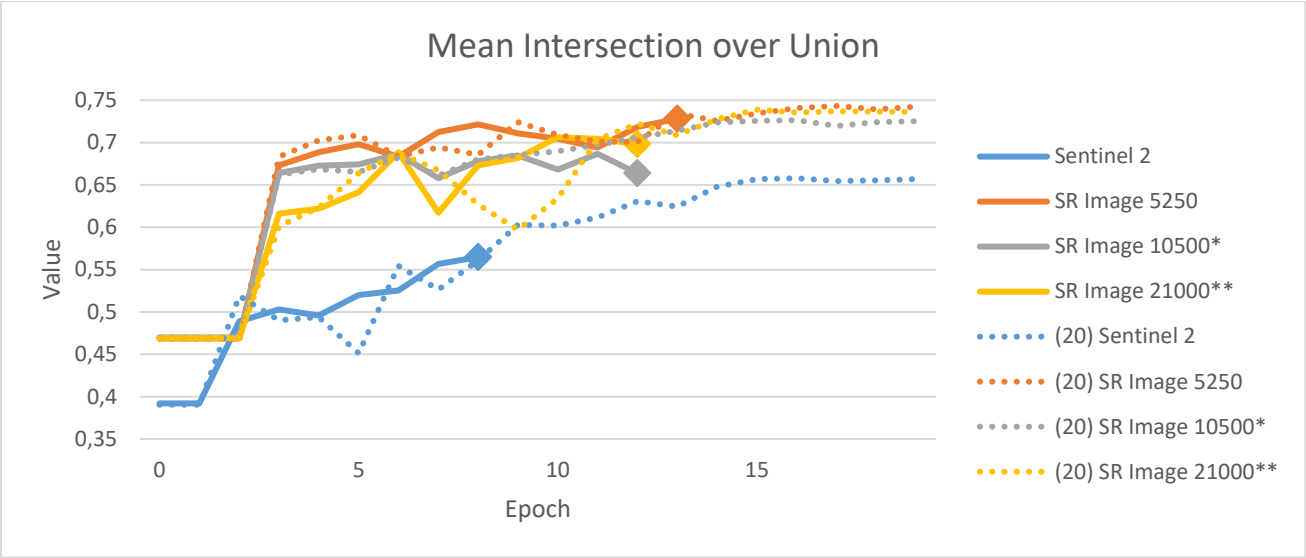
$$Accuracy\ 1 = \frac{5 + 20}{5 + 20 + 0 + 0} = 100\ \% \qquad Accuracy\ 2 = \frac{5 + 18}{5 + 18 + 2 + 0} = 92\ \%$$

$$Accuracy\ 3 = \frac{2 + 15}{2 + 15 + 5 + 3} = 68\ \% \qquad Mean\ Accuracy = \frac{12 + 53}{12 + 53 + 7 + 3} = 86{,}7\ \%$$

For the baseline Sentinel 2 model the initial accuracy when commencing training is already 78%, which can be considered satisfactory but can still be improved which the increase in value at each iteration also displays in figure 6.3. At epoch 20 the accuracy would have increase to 86% if it were not for the fact that the model triggered the early stop as if iterating beyond that point does not lead to significant improvements and could result in overfitting. The same can be said for the super resolution based models although their initial value was already 93% and only increase to 96% when the early stop was triggered and the variance in accuracy was already rather limited. However based on these models and the input data using super resolution imagery does lead to higher accuracy although the initial accuracy was already high in the case of super resolution and the improvement in accuracy as a result of model training was relatively small. The Sentinel 2 model made larger improvements in accuracy by training but had an initial lower accuracy and training did not increase it beyond the initial accuracy super resolution models achieved.

6.1.4 Mean intersection over union

Figure 6.5: Mean intersection over union



**: This model contains 5250 completely unique input images but the images have been rotated 180 degrees to double the amount of input images*

***: This model contains 5250 completely unique input images but the images have been rotated 90 degrees to quadruple the amount of input images*

MIoU (Mean Intersection over Union, also known as Jaccard Index) is a metric also assessing a form of accuracy, to be specific the ratio indicating how closely a prediction matches the ground truth of the model. The equation to calculate this is similar to accuracy but has one major difference and should be assessed on a different scale:

$$Mean\ Intersection\ over\ Union = \frac{True_{Postive}}{(True_{Postive} + False_{Positive} + False_{Negative})}$$

This equation leaves out the true negative aspect of assessing this accuracy like metric and evaluates the prediction on an object rather than global scale. To illustrate this take back the checkerboard example of a classification as shown in figure 6.5 with three different classification scenarios, which in total had an 87% accuracy.

$$IoU\ 1 = \frac{5}{5+0+0} = 100\ \% \qquad IoU\ 2 = \frac{5}{5+2+0} = 71{,}4\ \%$$

$$IoU\ 3 = \frac{2}{2+5+3} = 20\ \% \qquad MIoU = \frac{12}{12+7+3} = 54{,}5\ \%$$

The MIoU of the scenarios would result in a value of 54,5 %, which is significantly lower than the 86,7 % shown for accuracy in the previous subsection. For scenario 1 and 2 the values would not be that different if the choice is between displaying the global statistic of 86,7% compared to these zonal IoU's but for scenario 3 that is different as the model performed very bad in this zone while assessing based on the image as a whole (including true negatives) would not indicate. Accuracy displays how well the model was able to make a correct distinction between what is a road and what is not while (Mean) intersection over Union is a

metric displaying the fit between the prediction and the TruePositive predicted road, cancelling out the influence of TrueNegatives. This is needed because in the example there is a strong class imbalance, as the amount of TrueNegatives highly outnumbers TruePositives. And as illustrated by the accuracy calculation the TrueNegatives are overall correctly classified resulting in a high accuracy although the fit of the prediction masks with the ground truth positive pixels is relatively lower and when dividing it into segments can be even very low on specific objects.

Based on this example it should therefore be explainable that the curves shown for the MIoU of each model are lower in value than the accuracy. For Sentinel 2 the MIoU at the epoch the stop clause is triggered is 0,56 indicating there are on average almost just as many false positive and false negative predictions made in comparison to the actual amount of true positives indicating a relatively low prediction fit to the ground truth. The respective values of 0,73, 0,66 and 0,71 for the 5250, 10500* and 21000** images models reduce it to a ratio of about a third of false positive and false negatives in comparison to the amount of true positives, indicating an improvement from the Sentinel 2 baseline. Although somewhat comparable in value up to epoch 2 the SR models steadily improve in IoU while the Sentinel 2 model already starts to reduce in improvement of IoU. In general that would mean that the super resolution models on average make less False positive and False negative predictions in comparison to Sentinel 2, indicating a better fit of the predictions made in comparison to the actual ground truth. The MIoU for the Sentinel 2 data could still improve when iterating after the moment the stop clause would be triggered but for the super resolution models the gained MIoU is relatively limited and again this would be at the risk of causing overfitting.

6.1.5 Performance matrix road extraction

Summarising the different metrics discussed in the previous sections of this chapter, one can see the values for each model in table 6.1 at the epoch were the stop clause was triggered. For both training and validation loss, the Sentinel 2 model is less able to predict values correctly, resulting in large errors which eventually result in the high mean-square error as shown below. All the super resolution models return a training and validation loss in the same order of magnitude with numerical small differences but the 5250 model performance a bit better in those metrics. That could be caused by the fact that it ran for one more epoch than the other super resolution models until the stop clause was triggered. In accuracy all the super resolution models perform the same and provide a large improvement in comparison to the Sentinel 2 model. In MIoU the model with 5250 images shows the best fit between the ground truth data and the prediction being made for that data as the ratio is the highest. The 21000** images model is comparable and 10500* images displays the lowest MIoU of the super resolution models, still being a large improvement in comparison to the Sentinel 2 model. Although the 5250 images model also performs the best in the dice coefficient the calculation seems to be off or not consistent as explained in the previous paragraph.

Table 6.1: Performance metrics matrix road extraction

Model	Training	Validation	Accuracy	MIoU
Sentinel 2	1,15	1,46	0,81	0,56
5250	0,29	0,33	0,96	0,73
10500*	0,42	0,38	0,96	0,66
21000**	0,47	0,35	0,96	0,70

In all applying road extraction on super resolution instead of Sentinel 2 data can metric-wise lead to an improvement in results, up to a certain point as augmenting the data actually leads to a decrease in improvement. That being said performance is still greatly improved in

comparison to Sentinel 2 but using augmented data might therefore not be favourable as it does not lead to metrical improvements in results.

The next paragraph will further continue with evaluating road extraction but focusses on the visual evaluation. Important to note is that this will also include some statistical evaluation, as displaying the results visual allows for geometrical operations that can also provide additional data that can be evaluated.

6.2 Visual evaluation road extraction

A visual interpretation of the performance of road extraction will enrich the analysis already conducted in paragraph 6.1 but will also illustrate the added value super resolution in general. It shows what can be done with super resolution data and how it compares to both ground truth Sentinel 2 and Spot data but also in comparison to ground truth objects that are being analysed, in this case road infrastructure. It is therefore also a further in-depth analysis of the results already presented in chapter 5 about super resolution on itself.

Ground truth road infrastructure data is acquired open-source from OpenStreetMap. A distinction was made as discussed in the methodology between relevant infrastructure classes to include or exclude in the extraction, resulting in that only road features with labels matching these classes were included in the extraction task in the test area;

*Table 6.2: OSM road classes for road extraction**

Roads	**Link Roads**
Motorway	Motorway_link
Trunk	Trunk_link
Primary	Primary_link
Secondary	Secondary_link
Tertiary	Tertiary_link

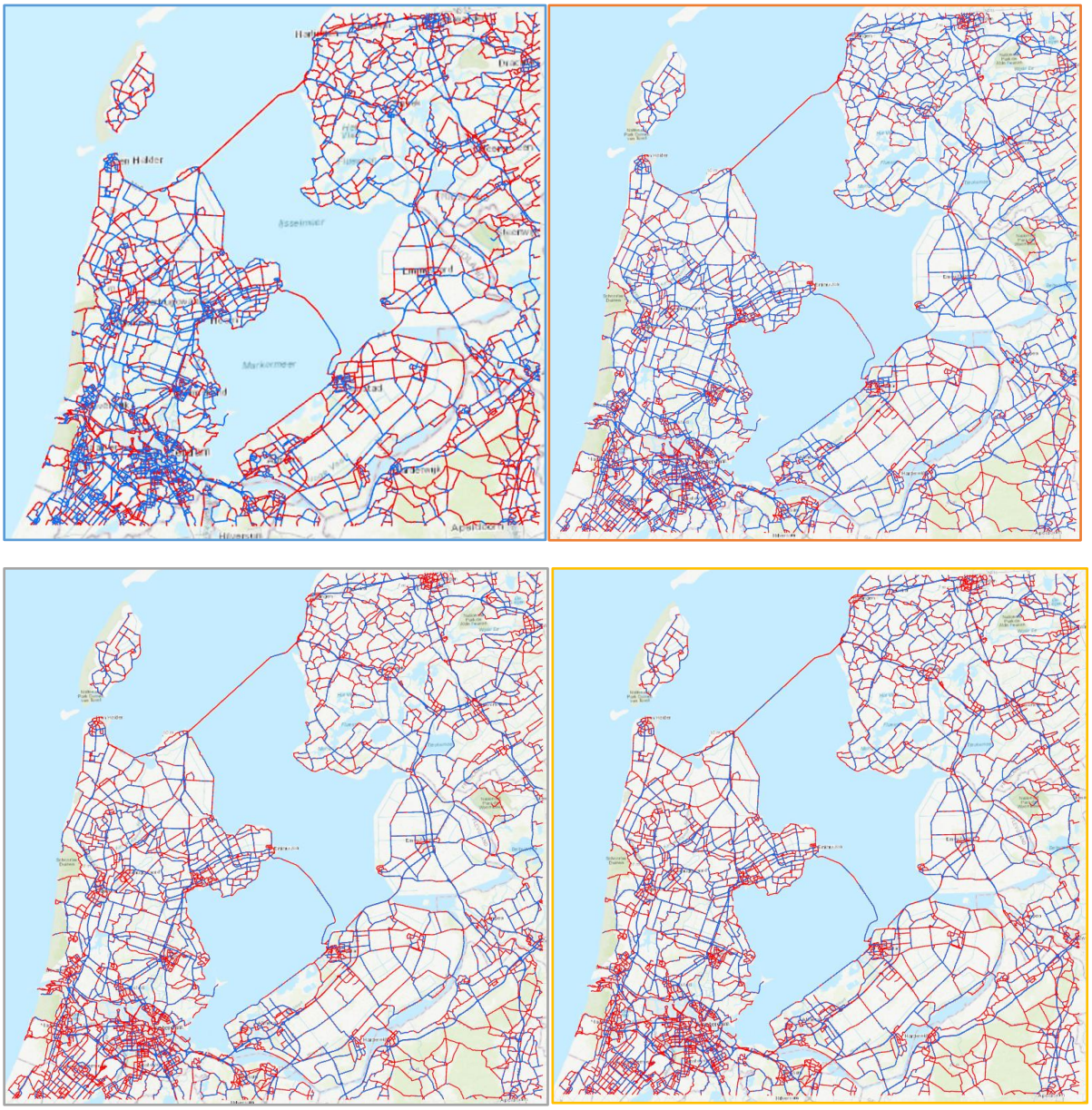
**This table is already included and explained in the methodology section but is repeated here for reference purposes*

Selecting only road features with these labels results in a visual network as shown in figure 6.1. The multi-task road extractor model in ArcGIS Pro only allows for classifying binary values and therefore the road network is displayed as a single value after the selection as described above has been made. The extractions on the test area are shown for each different model in figure 6.2 in blue, with the original OSM network on the background to show the quality of the extraction in comparison to the ground truth.

Figure 6.1: Ground truth OSM road network

Visually it can be seen that the road infrastructure with the chosen labels as shown in table 6.1 mainly include infrastructure linking settlements, with an increase in density around larger urban areas but the actual small scale infrastructure is left out in this analysis. By making this selection for relevant infrastructure beforehand the amount of feature segments in this area is reduced from 540.000 to 64.000 meaning that the amount of features with which an extraction can be matched is much lower, but that the ones it can match with are reduced to a relevant portion of the entire OSM network.

Figure 6.2: Sentinel 2 (top left), 5250 images (top right), 10500 images (bottom left) and 21000** images (bottom right) versus the ground truth (red) OSM network*



Evaluating the results of the extraction on this scale (the whole test area) shows at first that although the model training seemed to indicate metric-wise that the super resolution models outperformed the Sentinel 2 one greatly, that does not seem to be significant when testing the models as shown in figure 6.2. The network classified based on Sentinel 2 data in figure 6.2 shows that it does not classify a large segment of the ground truth OSM network as being a road, but that where it does it is mainly concentrated around the more dense urban areas. Especially in the southwest part of the test area on the axis Haarlem-Amsterdam the model seems to perform better in comparison to the overall test area. For the 5250 images super resolution based extraction shown in figure 6.2 this seems to be opposite and it is actually the roads linking settlements and urban areas with each other that are dominant in the extraction result and the urban areas to a lesser extent, as also shown in the close up of the networks on Amsterdam in figure 6.4. What is also clearly visible is that this super resolution based model classifies more road segments as being a road in comparison to the Sentinel 2 model as the

ground truth network (which in this sequence shows the parts of the network that are missed by the extraction) is less apparent.

Table 6.3: Length and percentage of extraction networks versus ground truth

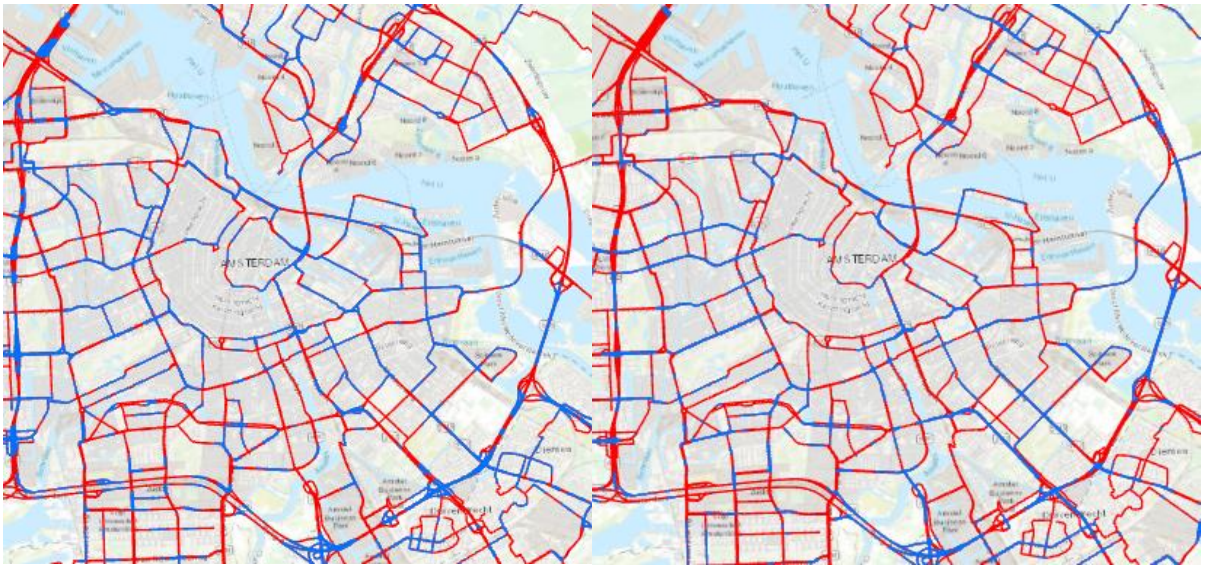
	OSM	Sentinel 2	5250 Images	10500* Images	21000** Images	SR Merged
Network Length	10.268.866 metre	3.720.866 metre	4.482.509 metre	3.889.930 metre	3.520.527 metre	5.889.880 metre
Percentage of Ground truth	100 %	36,23 %	43,65 %	37,88 %	34,28 %	57,36 %

An interesting observation is that the other super resolution models with 10500* and 21000** images do not show a further increase in the amount of ground truth segments that are correctly classified as roads. Outstandingly is that although the metrics seemed to perform relatively similar based on the statistical evaluating the models might still do so when tested for this area, but they all classify different segments of the ground truth OSM network.

Table 6.3 shows the total length of both the OSM ground truth network in the test area and the length of the classified network, which shows which portion of the network is actually classified by each model. This supports the visual evaluation which shows that indeed the 5250 images model outperforms the Sentinel 2 model, although comparing the quality of extraction in comparison to the overall ground truth the results might be considered modest. The percentage of the ground truth that is being classified by the super resolution models actually decreased after the 5250 images model and the 21000** images model actually performs worse than the Sentinel 2 model. Most interesting is that combining the super resolution model results with each other actually results in a net increase in the part of the ground truth network that is being classified indicating they do not classify exactly the same segments. Merging them geographically as shown in figure 6.3 also indicates the variance in segments classified.

Figure 6.3: Merged super resolution extraction networks

Figure 6.4: Close up of the Sentinel 2 (left) and 5250 images (right) extraction networks

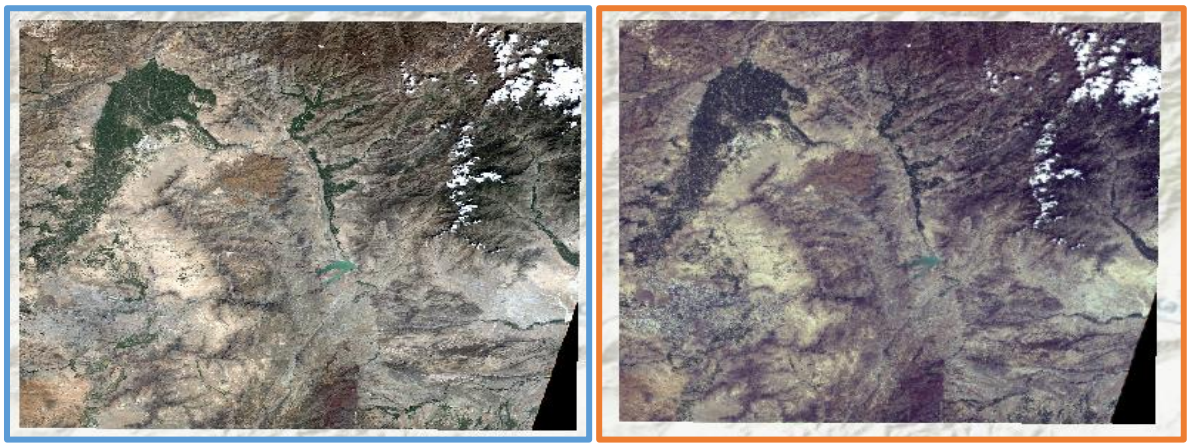


6.3 Generalizability

Paragraphs 6.1 and 6.2 elaborated on the performance of an analysis model, to be specific a road extraction model which used super resolution satellite data to analyse performance in comparison to a Sentinel 2 model. The similarities in physical environment between the training and test area (as they are located in the same country) made for somewhat ideal circumstances to evaluate model performance, which is not a problem as it is about the initial assessment and proof of concept.

However it is also relevant to test the models in an entirely new and possibly even unfavourable environment then where it was initially trained for. Figure 6.5 shows Sentinel 2 data of such kind of environment, containing the Afghan capital of Kabul and the area to its east. As can be seen it is a dominantly arid environment with some sections of vegetation near the mountain ranges north of Kabul which also dominate the physical environment. Also there is some cloud cover to the northeast and a small corner of the image is left out.

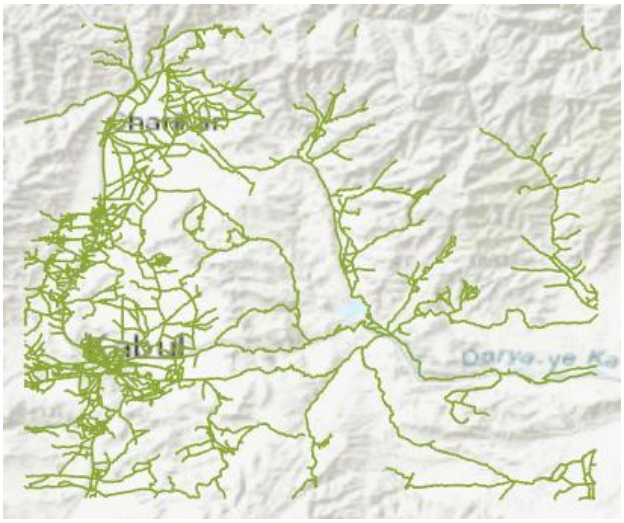
Figure 6.5: Sentinel 2 versus 5250 images super resolution of Eastern-Afghanistan



Comparing the original Sentinel 2 image with the 5250 images super resolution the initial result can be deemed as good, as it preserves structural integrity and shapes (even of the cloud cover). The illumination however is just as in the original test lower for the predicted super resolution image in comparison to the original Sentinel 2 image. This reduces the contrast in colour in the image making significant features less visually apparent. The water reservoir in the middle of the image area is a good example of that.

The subsequent question is how using a super resolution image of an in this case new and less favourable environment influences a possible analysis task, in this case road extraction. Figure 6.6 shows the ground truth OSM network of this test area.

Figure 6.6: Ground truth OSM network of the image area



Just as for the super resolution model one should consider that the model which is being used for executing the road extraction is trained on an entirely different area in regard to the physical environment. The ground truth network is also different in this area in comparison to the earlier test area, although it is still similar in the sense that it is high density in the urban environment but in the rural areas it is of very low density.

Figure 6.7: Sentinel 2 (left) versus 5250 images super resolution extraction network

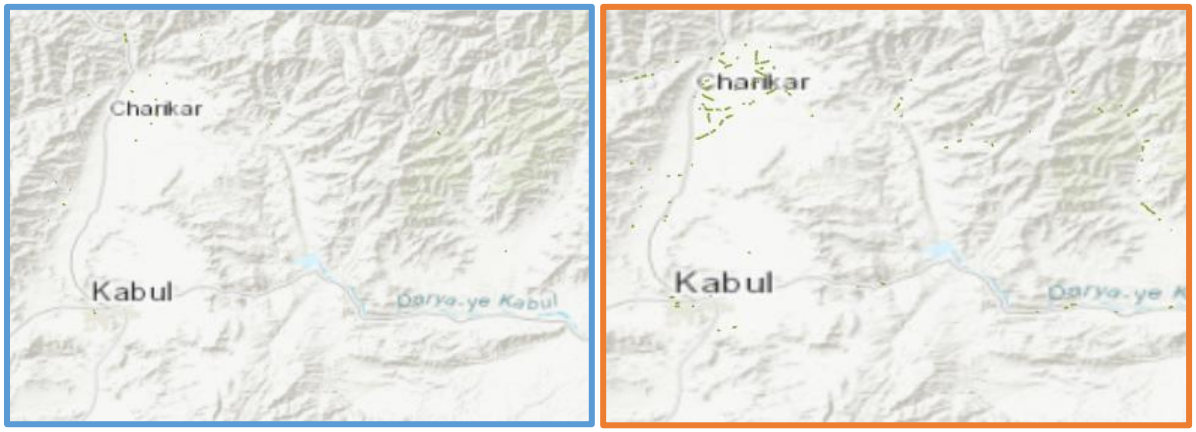


Table 6.2: Length and percentage of extraction networks versus ground truth Afghanistan

	OSM	Sentinel 2	5250 Images
Network Length	4.080.106 metre	4276 metre	42.908 metre
Percentage of Ground truth	100 %	0,1 %	1,1 %

What immediately stands out in both figure 6.7 and table 6.2 is that the extraction models have a lot of trouble operating in this different type of environment. The percentages of classified ground truth segments are very low for both Sentinel 2 and the SR 5250 images model. Outstanding in that sense is the area where it performed best is a forest /vegetated area around in the north-western part of the area (see figure 6.5 for reference), which might be explainable as it is somewhat representable to the area on which the model was originally trained.

7. Conclusion

The goal of this research was to provide a proof of concept and a better understanding of the possible added value of super resolution to remote sensing analysis and the broader geo-information domain. This was primarily done by training a deep learning model with high resolution satellite data which is then applied on test images (with a lower resolution) to predict and create super resolution images. With super resolution imagery a deep learning model was trained to extract road features from images. Simultaneously this was done for unmodified lower resolution imagery in order to make a comparison in extraction results. Although super resolution images can be evaluated on themselves and the models used to create them provide statistical data the use of images in a deep learning extraction model provided an extra dimension to how the quality of super resolution could be assessed. Training data was extracted from the north-eastern region of the Netherlands, test data from the central/western region of the Netherlands and to evaluate possible environmental bias the models were also tested on a case in Central-Afghanistan. All images are taken in the summer and under cloud-free conditions.

This empirical analysis was aimed to contribute to answering the following leading research question;

To what extent can super resolution be of significant added value to feature extraction on Sentinel 2 data?

The conclusion (and therefore answer to the research question) is that super resolution results in a better performance for feature extraction in comparison to Sentinel 2 but only relatively within the used data and constraints as part of this research. On an absolute scale there is opportunity for improvement.

The sub-questions focused on the data requirements, how to define SR and how it is already being used, how SR compares to original Sentinel 2 data and how these two would compare (in both quality and statistics) when used in an actual analysis task.

Super resolution is the technical operation of improving the resolution of imagery to a higher resolution standard. Super resolution predicts imagery at a higher resolution to provide more detail than the original image but one should consider the technical parameters that evaluate the quality of a super resolution operation. Super resolution is well-applied as part of computer vision in general (so also on non-geographical data) but research is mainly focused on the initial data and not per se its use in further analysis.

Data should be preferably cloud-free and also be well illuminated and provide colour contrast of the area, which can be different depending on the weather, time of year and time of sensing. The moment (both season and time of day) at which data to train the model and the image on which super resolution is applied on should be similar to reduce differences in conditions. Data itself should be in three bands (RGB) and in an 8 bit data format to be operable in ArcGIS Pro. If initial data does not meet these conditions modifications can be done to make the data operable. Satellite imagery can also be augmented to provide more training samples when the area for which high resolution data is available is limited.

Dependent on the research requirements the factor can be different but in this research the super resolution model provided data predictions on a 2,5 metre resolution in comparison to the original 10 metre resolution. In performance metrics the super resolution model with augmented data outperformed the initial super resolution model but further augmentation indicated a qualitatively lower performance so making sure input data is unique is still of

importance and augmentation is not an infinite solution to low data availability. Visually augmentation or not made no difference and on the scale of image specific performance metrics super resolution proved to be underperforming on an absolute scale.

Evaluating super resolution by applying geographical analysis, in this case road extraction, showed that super resolution models (regardless of augmentation) greatly outperformed an original Sentinel 2 based model in all performance metrics. Visually and in comparison to the ground truth of the extracted features all models performed poorly and extracted only a minority of the total amount of road features in the ground truth dataset. The initial super resolution model outperformed the Sentinel 2 data based model with 20% but with an extraction score of 43,7% left more features out than it actually classified. Data augmentation on the training imagery actually negatively influences performance and in one case even performs worse than Sentinel 2. In a desert environment the initial super resolution model performed not significantly different than for the other test area. In the extraction task the Sentinel 2 based model performed negligible with only 0,1% and super resolution performed relatively better with 1,1% but on an absolute scale that remains a negligible result.

The following and final content chapter of this thesis will evaluate this research and provide some hands-on analysis about the choices and constraints that were part of it and to also provide where these opportunities could be for future research.

8. Discussion and recommendations for future research

This research discussed the possible added value super resolution could have in remote sensing analysis, within certain bounds defined by the research design, research institute and also resources in personnel, financial, material and time aspect. After the conclusion that were made in the previous chapter this chapter will try to take a step outside those boundaries to evaluate the research itself and to what extent the lessons learned could be useful or provide stepping stones for future research. As an intuitive and sequential structure this chapter will touch upon the different aspects of the research in the same way as it was empirically designed, meaning it will start with the initial input data and will end at the super resolution and feature extraction data. Along the way all the other relevant aspects will be further explained.

8.1 Data

To start with the initial satellite imagery. As mentioned the used high resolution satellite imagery of the Spot 6 satellite was accessible via the NSO, which was fairly straightforward and above all practical. High resolution imagery is hard to come by as it is either expensive (from commercial sources) are not meant to be openly accessible (military). In this case that did not prove to be a problem as the imagery was free and of such a resolution that it fitted the research goal but if for example sub-metre resolution was needed to classify small scale objects that would have been difficult. Just as the open source satellite imagery from NASA's Landsat and ESA's Sentinel missions can be openly accessed it proved that a space agency was also for free high resolution imagery a good source and is therefore recommended as it is a crucial part of this type of research and is also a suitable solution of financial resources are limited. The amount of data needed was also not a problem in this case as the NSO provided much more images than were actually needed to train the model but was limited to both time and area from which the NSO made data available via their server, where in this case the NSO only provided relatively complete datasets from 2014 and 2015. This was however suitable for this research and scope because several datasets were unfit because of weather conditions but the catalogue was extensive enough to use other datasets.

Outside the scope of this research but also relevant from a data and intelligence standpoint is looking towards other types of remote sensing data (besides optical), although dependent on the area and object of interest. For example radar or infrared remote sensing data might be useable or favourable for certain analysis tasks. Radar because of its capabilities to penetrate surfaces and microwave to sense heat can provide information not captured in optical imagery which especially in a military context might be relevant to do so. However it applies to the analysis not the super resolution aspect as these types have other magnitudes of resolution and also for road infrastructure as this is difficult to capture in these sensor types but it is relevant to mention that dependent on the scope other data sources might come in more useful or advantageous than optical remote sensing.

8.2 Methodology

Another interesting point of discussion and a source for recommendations is the software used to handle, predict and analyse the satellite data as part of both the super resolution and road extraction deep learning tasks. As American-based Esri is the owner and developer of the used ArcGIS Pro software and is a commercial provider (although free-handed in providing universities with licences) which means using their software professionally

as a government or company can be costly. That also automatically means that using deep learning based analysis like the super resolution and road extraction models as executed in this research are dependent on the licensing (and resources) of future research. Although it served in this research as a solution to enable the initial proof of concept one could argue that one is limited to the use of Esri's API for enabling these types of analysis or that alternatives exist. As stated in the methodology there are several open-source environments available for deep learning like SR4RS and Open CV and the ArcGIS.learn API is inherently linked with already existing open-source python packages as these are commonly used and already well developed. Esri's own tutorials on deep learning illustrate that (as it also can be coded in Jupyter Notebook and imports packages like Torch and Fastai) but only the actual package with the deep learning functions as used in this research are Esri-based and are the tools that are being accessed via ArcGIS Pro. Esri prohibits independent use of these packages outside the ArcGIS Online environment meaning that the API's are inaccessible with proper licensing. An open-source alternative would therefore only need to replace the actual modelling which is now done by Esri's API as the pre-processing tools are already openly available. But that requires a level of technicality and time which may not always weigh op against the financial cost of licensing via Esri but would have the great advantage of independence in both development and use of the required deep learning packages. ArcGIS Pro's deep learning toolbox proved to be, possibly because it used to be separated and required scripting in Jupyter Notebook, susceptible to bugs and errors when trying to execute analysis and therefore alternative software solutions might be favourable. That would not mean no bugs or errors could arise but provides the possibility to debug the scrips used in comparison to try and navigate through the ArcGIS Pro environment front-end where it might not always be directly reproducible which tools and scripts are being used.

In regard to the chosen methodological aspects of this research there are also recommendations that can be made. The backbone of both ArcGIS deep learning models (Super Resolution and Multi-Task Road Extraction) were ResNet-34 models which were trained on the ImageNet database for object detection. Although the establishment of both Resnet as a neural network architecture and ImageNet as a pillar in advancing computer vision points of recommendation could be the layer depth and insight on the topic of ImageNet. For Resnet 34 layers is relatively shallow, not being the bare minimum of 18 layers as used by ArcGIS Pro but also not that extensive as the possible 152 layers model. To an extent deep-layered models provide a better performance for the task at hand but one should also consider its practicality in use especially when time to complete a task is not indefinite. To illustrate the training times in this research could increase up to 2 days for the maximum amount of input data used and testing it in images several hours so using a deeper-layered architecture further extends those timeframes. Discussing this here aims to point out the chosen architecture and how the amount of layers set the timeframes as experienced in this research but whether or not a different layered architecture would be more favourable is more research dependent. ImageNet is a well-established benchmark in computer vision and provides a versatile and properly indexed dataset which is also free to use for research. Although the complete dataset can be used to train a model to learn visual patterns as they exist in all kinds of images of objects one could argue that to make training as purposeful as possible the used subset should be filtered for the test images one wants to use it for. ArcGIS Pro uses a subset of about one million images but does not provide metadata on this and it is therefore something to consider when using the pre-selected ImageNet set as handled by ArcGIS or when designing an own research methodology.

Important to stipulate is the fact that the analysis focussed on extracting road infrastructure, which are features and not objects. This is important to consider as it proved difficult to accurately assess the extraction quality, even apart from the fact that the user interface

provided by Esri made insight into the actual used python tools difficult as extracting a network of features makes that a model can result in only partially extracting a feature. The results showed that as different extraction models extracted different segments of the network although it was assumed that it would be similar in results. It is possible that it is due to the fact that this research applied only a binary classification, e.g. is a pixel belonging to a road or not. This was due to the classifier's constraint of only being able to handle binary data but a more specific task (like only extracting highways) might make the classifier more effective and would also improve the level of information by providing ordinal data if multiple classifiers are applied for different road classes. Although in this research a segmentation in relevant classes was already made future research could definitely benefit from further segmenting the data as it will enable a higher level of information provided by the analysis.

8.3 Generalizability and bias

To conclude this discussion the results, which do not only comprehend the direct model output and the test that were executed but also the broader implication, usability and possible bias accompanied with the models and how they were used. The super resolution model was trained and tested within the administrative boundaries of the Netherlands, meaning that the structures and physical environment were typical for this area of Europe. The Netherlands is also data-wise a developed country from which a lot of data and information is made available by all kinds of institutions. That made the model inherently biased to this type of environment but not per se unusable outside of it as a primary test in Afghanistan which also greatly differed in physical environment did not show an inherent bad performance when it comes to super resolution. Road extraction did perform considerably worse in this alternate environment which does prove that there are some boundaries when it comes to generalisability of the model and the bias caused by the choice of training data. A recommendation could therefore be to further experiment with how to comprise the training data (although the research objective should still be leading) as a further development of this initial proof of concept to try and explore how a model could be more generalizable and therefore also more practical in its use for predicting super resolution and road infrastructure and to what extent that would influence the level of detail in predictions.

This research proved the versatility, accessibility and also workability of deep learning models as part of geographical research and remote sensing analysis, which depending on the research needs and also technical knowledge can be as practical or theoretical and technical as possible. This study was a proof of concept in the practical sense, with a clear analysis objective in mind to also show the feasibility of further improving remote sensing analysis by the Ministry of Defence. The discussion did not prove major issues or challenges that would put the result and validity of this thesis in a different light but rather laid out several points for which a different approach by a new researcher could lead to a difference in insight on these points and also a different methodological design. Just as this research stood on the shoulders of predecessors the hope is that this research can contribute to the field of knowledge on deep learning and analysis in the geo- and remote-sensing domain and provide future researchers into this theme with some hands-on experiences and insights.

9. References

Ayala, C., Sesma, R., Aranda, C., & Galar, M. (2021). A deep learning approach to an enhanced building footprint and road detection in high-resolution satellite imagery. *Remote Sensing*, *13*(16), 1–21. https://doi.org/10.3390/rs13163135

Bhagavathy, S., & Manjunath, B.S. (2006). Modelling and detection of geospatial objects using texture motifs. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(12), 3706–3715. https://doi.org/10.1109/TGRS.2006.881741

Bioucas-Dias, J. M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N.M. & Chanussot, J. (2013). Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and Remote Sensing Magazine*, *1*(2), 6–36. https://doi.org/10.1109/MGRS.2013.2244672

Castelluccio, M., Poggi, G., Sansone, C., & Verdoliva, L. (2015). *Land Use Classification in Remote Sensing Images by Convolutional Neural Networks*. 1–11. http://arxiv.org/abs/1508.00092

Chen, X.L., Zhao, H.M., Li, P.X., & Yin, Z.Y. (2006). Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. *Remote Sensing of Environment*, *104*(2), 133–146. https://doi.org/10.1016/j.rse.2005.11.016

Ehlersa, M., Klonusa, S., Åstrandb, P. J., & Rossoa, P. (2010). Multi-sensor image fusion for pan-sharpening in remote sensing. *International Journal of Image and Data Fusion*, *1*(1), 25–45. https://doi.org/10.1080/19479830903561985

Esri (n.d.). Understanding segmentation and classification in ArcGIS Pro 2.8. Consulted on the 9th of November 2022 from https://pro.arcgis.com/en/proapp/2.8/tool-reference/image-analyst/understanding-segmentation-andclassification.htm

Fernandez-Beltran, R., Latorre-Carmona, P., & Pla, F. (2017). Single-frame super resolution in remote sensing: a practical overview. *International Journal of Remote Sensing*, *38*(1), 314–354. https://doi.org/10.1080/01431161.2016.1264027

Galar, M., Sesma, R., Ayala, C., Albizua, L., & Aranda, C. (2020). Super-resolution of

Sentinel-2 images using convolutional neural networks and real ground truth data. *Remote Sensing*, *12*(18). https://doi.org/10.3390/RS12182941

Gargiulo, M., Mazza, A., Gaetano, R., Ruello, G., & Scarpa, G. (2019). Fast super resolution of 20 m Sentinel-2 bands using convolutional neural networks. *Remote Sensing*, *11*(22), 1–18. https://doi.org/10.3390/rs11222635

Goodfellow, I., Bengio, Y., & Courville, A. (2015). Deep learning. *The MIT Press*.

He, K., & Sun, J. (2015). Convolutional neural networks at constrained time cost. In CVPR.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In CVPR.

Hu, F., Xia, G. S., Hu, J., & Zhang, L. (2015). Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, *7*(11), 14680–14707. https://doi.org/10.3390/rs71114680

Huang, B., He, B., Wu, L., & Guo, Z. (2021). Deep residual dual-attention network for super-resolution reconstruction of remote sensing images. *Remote Sensing*, *13*(14). https://doi.org/10.3390/rs13142784

Krizheysky, A., Sutskever, I., & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Handbook of Approximation Algorithms and Metaheuristics*, 1097–1105. https://doi.org/10.1201/9781420010749

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., … Shi, W. (2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *Computer Vision Foundation*, *2*(3), 4. http://openaccess.thecvf.com/content_cvpr_2017/papers/Ledig_PhotoRealistic_Single_Image_CVPR_2017_paper.pdf

Li, K., Yang, S., Dong, R., Wang, X., & Huang, J. (2020a). Survey of single image super resolution reconstruction. *IET Image Processing*, *14*(11), 2273–2290. https://doi.org/10.1049/iet-ipr.2019.1438

Li, Z., Li, Q., Wu, W., Yang, J., Li, Z., & Yang, X. (2020b). Deep recursive up-down sampling networks for single image super-resolution. *Neuro computing*, *398*, 377–388. https://doi.org/10.1016/j.neucom.2019.04.004

Liebel, L., & Körner, M. (2016). Single-image super resolution for multispectral remote sensing data using convolutional neural networks. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *41*(July), 883–890. https://doi.org/10.5194/isprsarchives-XLI-B3-8832016

Luus, F.P.S., Salmon, B.P., Van Den Bergh, F., & Maharaj, B.T.J. (2015). Multi view Deep learning for Land-Use Classification. *IEEE Geoscience and Remote Sensing Letters*, *12*(12), 2448–2452. https://doi.org/10.1109/LGRS.2015.2483680

Marmanis, D., Datcu, M., Esch, T., & Stilla, U. (2016). Using ImageNet Pre-trained Networks. *IEEE Transactions on Geoscience and Remote Sensing Letters*, *13*(1), 105–109.

Ministerie van Defensie (2019). Nederlandse Defensie Doctrine (Dutch).

Ministerie van Defensie (n.d.). DP-1142: Wegendetectie en –classificatie (Dutch).

MIT Technology Review (2013). Acquired from https://www.technologyreview.com/10breakthrough-technologies/2013/

Müller, M.U., Ekhtiari, N., Almeida, R.M., & Rieke, C. (2020). Super-Resolution of Multispectral Satellite Images Using Convolutional Neural Networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*(1), 33–40. https://doi.org/10.5194/isprs-annals-V-1-2020-33-2020

Nogueira, K., Penatti, O.A.B., & Dos Santos, J.A. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, *61*, 539–556. https://doi.org/10.1016/j.patcog.2016.07.001

Penatti, O.A.B., Nogueira, K., & Dos Santos, J. A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2015-October*, 44–51. https://doi.org/10.1109/CVPRW.2015.7301382

Pouliot, D., Latifovic, R., Pasher, J., & Duffe, J. (2018). Landsat super-resolution enhancement using convolution neural networks and Sentinel-2 for training. *Remote Sensing*, *10*(3). https://doi.org/10.3390/rs10030394

Romero, L. S., Marcello, J., & Vilaplana, V. (2020). Super-resolution of Sentinel-2 imagery using generative adversarial networks. *Remote Sensing*, *12*(15), 1–25. https://doi.org/10.3390/RS12152424

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In ICLR.

Tian, J., & Ma, K.K. (2011). A survey on super-resolution imaging. *Signal, Image and Video Processing*, *5*(3), 329–342. https://doi.org/10.1007/s11760-010-0204-6

Volpi, M., & Tuia, D. (2017). Dense semantic labelling of sub-decimetre resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(2), 881–893. https://doi.org/10.1109/TGRS.2016.2616585

Wang, X., Gao, X., Zhang, Y., Fei, X., Chen, Z., Wang, J., … Zhao, H. (2019). Land-cover classification of coastal wetlands using the RF algorithm for Worldview-2 and Landsat 8 images. *Remote Sensing*, *11*(16), 1–22. https://doi.org/10.3390/rs11161927

Xia, G., Yang, W., Delon, J., Gousseau, Y., & Sun, H. (2010). Structural High-resolution Satellite Image Indexing to cite this version. *ISPRS TC VII Symposium - 100 Years ISPRS*, *XXXVIII*, 298–303.

Yang, Y., & Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land use classification. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 270–279. https://doi.org/10.1145/1869790.1869829

Yang, Y., & Newsam, S. (2013). *Geographic Image Retrieval Using Local Invariant Features*. *51*(2), 818–832.

Yue, X., Chen, X., Zhang, W., Ma, H., Wang, L., Zhang, J., … Jiang, B. (2022). Super-Resolution Network for Remote Sensing Images via Pre classification and Deep–Shallow Features Fusion. *Remote Sensing*, *14*(4), 925. https://doi.org/10.3390/rs14040925

Zhu, X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). *Deep learning in remote sensing: a review*. *41501462*, 1–60. https://doi.org/10.1109/MGRS.2017.2762307

Zou, Q., Ni, L., Zhang, T., & Wang, Q. (2015). Remote Sensing Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing Letters*, *12*(11), 2321– 2325.

10. Appendices

Appendix A: ArcGIS Pro tooling for data preparation and modelling

Super resolution training samples

The creation of training samples is the preparation of imagery in order to be suitable for conducting the training of a deep learning model. What it does is split up the overall image in smaller samples based on the input parameters to train the model image-by-image instead of immediately training it on a larger scale. The details of this pre-processing and specific tooling that is used to make data suitable for training is discussed step-by-step for each part of the workflow.

1. **Downloading data and inspecting metadata**

In this research high resolution imagery that will be used for training is acquired from the Netherlands Space Office (NSO), which provides free-to-use data via a server in several formats. For training the model data should be in an 8 bits format and preferably cloud-free. The NSO provides data in 8 bits, RGB type and 1,5 metre resolution format and is downloaded via an ftp-client. Inspection of the metadata learns that data is stored in four bands instead of three, which requires an extra-preprocessing step because training cannot be done in a four band format. Dependent on the data source pixel depths, resolution or number of bands can be different then shown here but it is important to be aware of the required 8 bit depth and three-band format in order to be able to train a super resolution model and the required pre-processing steps.

2. **Area extent with Raster Calculator**

Creating an extent of the area for which data is available is necessary in order to create a physical processing extent for further analysis but also to make modifications to the data. Modifications include cutting out areas that are not relevant for the eventual training process but also areas that are for whatever reason unfit for analysis. In the case of satellite imagery that would be for example cloud cover.

The “Raster Calculator” tool in ArcGIS Pro provides the ability to write an expression to create an area extent. The expression needed is;

Con(“Name_Of_Dataset” >= 0,1)

The output is a raster with the size and extent identical to the input dataset. With the “Raster to Polygon” tool the datatype can be changed from raster to feature class and this is needed because in the next step a feature class will be made which will alter this extent.

3. **Feature Mask and extraction**

In the file geodatabase (the default database which is created and linked to each individual project) a new empty feature class can be added. Subsequently this feature class can be modified in the edit section, which in this case means creation as it does not contain any geometry yet. For this research it is needed to draw features around areas that are not relevant or suitable for the purpose for which the model will be trained. The segments where cloud cover makes seeing the actual surface impossible could be cut out because the model will have difficulties with extracting information from it. Because this is a coastal area the part of the imagery which displays the sea and part of the islands should be left out as well. Learning to predict super resolution for the sea is not relevant for eventually extracting and classifying roads from the images. The islands do not contain any relevant or prominent road features and

can therefore be left out as well. A major advantage of this is that it greatly reduces the file size and therefore also improves processing capability.

Drawing features for the areas that are not relevant to include or are unsuitable for analysis is a manual operation but for relatively cloud-free images should not be too cumbersome to do. When the manual drawing of features is finished it can be saved to the feature class /it should be updated when exiting the edit mode.

By using the tool “Erase” the drawn features in this feature class can be used as an overlay on the original area extent and the drawn areas will be erased from this extent. The area extent now can be used as a mask on the actual imagery to leave out the areas with cloud cover or that are not of relevance to be used in training.

The tool “Extract by Mask” extracts the right areas from the satellite image and leaves out the ones that cannot be used. What remains in this case is an image without (except for a small portion of coastal area) large segments of sea, the nearby islands and the areas where cloud cover makes visual observation of the surface too cumbersome.

4. Resolution and composite bands

When the extract by mask is successfully done there remain two parameters that should be altered in order to suit the requirements of training a deep learning model. The spatial resolution should be changed to 2,5 metres instead of 1,5 metres and the amount of bands should be altered to 3 instead of 4.

The tool “Make Raster Layer” allows for the creation of a raster layer but more specifically also which bands will be exported with it. In this case band “4” needs to be left out as band 1-2-3 represent Red-Green-Blue (RGB) and are the required and also maximum amount of bands a deep learning model can process at once.

Exporting the data will cause the output layer to be presented as a rectangle polygon, but can again be reduced to the area of interest by performing an “Extract by Mask” operation.

The tool “Resample” offers the opportunity to resample the imagery to a different cell size. Resampling to 2,5 metres offers two benefits; it again reduces file size and a more suitable cell size in the actual training process. In training the training data is downsampled to a lower resolution to create synthetic data for which the model will be trained to “predict” the same area but at a higher resolution (the actual super resolution). For this research the goal is to predict super resolution for sentinel 2 imagery, which has a 10 metre resolution. Downsampling 2,5 metre resolution to 10 metre resolution would need a downsample factor of 4, which is an easier number to work with instead of a decimal number when the resolution would remain 1,5 metre.

Inspecting the metadata of the output raster data should confirm that both the amount of bands and the resolution are altered correctly.

5. Export Training Data for Deep Learning

Exporting Training Data is an image analyst tool that coverts raster data (optionally combined with a feature class, classified raster or table) to image chips that can be used to train a Deep Learning Model. If the previous steps were followed correctly raster data should be in an 8-bits format and composed out of three bands. For Super Resolution an additional feature class is not relevant. However adding a feature class polygon for the purpose of masking can be useful as it then makes sure image chips fall completely within that designated area (it prevents that images might be partially black). Chosen tile size is 512 with a stride equal to that, meaning

that tiles will not overlap with each other to maintain their uniqueness to each other. If there is a need for more input image chips the choice can be made to make the stride smaller than the tile size. This results in more images but because of the overlap their similarity might be higher than anticipated. In this case the choice was made to work with a rotation angle of 90 degrees to get more image chips from the same area but with a relatively low similarity because of that rotation.

For Super Resolution the required metadata format needs to be export tiles, as then the image chips will not have any labels. Labels will be made when training the model which in this case will be the downsampled version of the same image to train super resolution.

In the environments tab the extent can be set but if a mask feature polygon is in use setting the extent will do just the same as that but it can be done to make sure the processing is limited to the required area. Setting a parallel processing factor (between 0 and 100%) ensures optimal use of CPU cores for the operation. Important parameter to set is the cell size which should in this case be 2,5 metres as that is the resolution on which the model will be trained.

Train a Super Resolution Model

For this the tool “Train Deep Learning Model” from the GeoAI toolbox is used. The folder containing training data can be instantly imported in this tool under “Input Training Data”. The Deep Learning Package (.dlpk filetype) and Esri Model Definition File (.emd filetype) this tool will produce can be put in a folder of choice at “Output Model” but should be empty to be able to store files there after training. The maximum amount of epochs can be set at “Max Epochs” and is by default set to 20.

Under “Model Parameters” a specific model type can be set (which are at the moment 25 different ones that are usable in ArcGIS Pro) and super resolution is one of those categorized under Image Translation. Setting the model type to super resolution will also directly set default model arguments and the backbone model. For the model parameters the Batch size can be set to define the quantity of training images that can be processed simultaneously. Dependent on the types of training samples and also the modelling task more or less memory is needed to perform the task and if the batch size is too high for the available memory the process will fail, for which the simple solution is reducing the batch size. But this is also dependent on the specifications of the operating system so there is no ideal answer to this as it will differ for each type of data, operation and system that is being used. The standard model arguments for super resolution are a downsample factor of 4 and a monitor set to valid_loss. This means that the images will be downsampled from 2,5 metre to 10 metre resolution (which as discussed before is desirable for this research) and the model will look at validation loss as a reference for a possible model stop, more on this later.

In the “Advanced” tab the learning rate can be defined that the model will use, which again can be case-dependent. If it is not clear beforehand which rate is desirable the rate can be set to 0 and the model will determine itself what the optimal rate is by plotting the learning rate versus loss and auto-extract the most optimal rate from this plot.

For super resolution the backbone model is a standard ResNet model with 34 layers. Dependent on the requirements this can be scaled up or down to let the data pass through a more or less complex model. Changes in this and also the amount of epochs will influence the required processing time to complete training. If applicable the .emd files or deep learning packages from an already trained model can be added here to optimize it further.

The training-validation split can be made here, indicating which percentage of the initial training dataset will be saved to validate the model on. By default it is 10% but can be altered based on the research requirements.

By default the Stop when model stops improving and Freeze Model will be checked. The first indicates that if no significant improvements are made when modelling (for which validation loss will be monitored as defined earlier) the model will stop the training. This will be if there are no changes larger than 0,001 for the duration of 5 epochs. A different monitor value can be set as well at the model arguments section. If the stop criteria is met the model will stop training at the 5th epoch after no significant improvements were made and this can therefore be earlier than the maximum amount of epochs. Freeze Model indicates that the weights as defined for training of the backbone model will be used instead of altering those during training on this specific dataset as set by the user. If checked it will improve processing but if a decision is made to leave it unchecked processing time can increase but the results can be more optimized towards the specific training dataset.

For the environments tab setting a parallel processing factor works the same as in the previous step of exporting data. What is different that here the choice can be made to run the model on either the CPU or GPU. For modelling the GPU can be desirable as it will have increase processing power but is also hardware dependent (memory of CPUs and GPUs can differ greatly) so it is up to the specific user. If set to GPU the parallel processing factor will be discarded.

Classify Pixels Using Deep Learning

After the training of the model this tool from the Image Analyst toolbox can be used to do actual classification and application of the trained model.

The Input Raster is the datafile on which the operation will be applied, which for this research will be a 10 metre resolution Sentinel 2 file. The Output Classified Raster will then generate a name and storage location based on the default geodatabase that the opened ArcGIS Pro project uses currently. For Model Definition either the deep learning package (.dlpk) or .emd file generated by the training tool can be inserted. Dependent on the type of operation additional arguments will be listed here and can be altered based on the research requirements or in the case of batch size the available hardware.

The Environments tab is similar in layout as encountered in the other tools. Important is to set the cell size to the required output size (in this case 2,5 metre) to get the required result. All the others like the coordinate system, extent, parallel processing factor and processor type are optional but can be defined as well.

Super resolution vs Road Extraction

The guide focused on the super resolution aspect of the different tooling that can be used to do all the processing correctly, but should require a small addendum about road extraction.

For road extraction most procedures will be the same. For the Export Training Data for Deep Learning the Input Feature Class field (which is empty for super resolution) should be used to add the feature class which contains the road network data. For the Train a Deep Learning Model the model type should also be changed to Multi-Task Road Extraction instead of Super Resolution. For the Classify Pixels Using Deep Learning operation different model arguments will be imported for python as those will not be the same as for Super Resolution. For the other settings and parameters as used by the deep learning applications in ArcGIS Pro they will not

differ between the types of models but as said often in this guide one can alter them based on the available hardware or based on the research requirements.

Road extraction analysis

To re-evaluate the quality of road extraction analysis in comparison to the ground truth road network, as for example the data from OpenStreetMap that was used in this research, a small post-processing step is needed to perform certain geographical analysis operations.

The raster file resulting from the “Classify Pixels using Deep Learning” tool can be converted to a feature class using the “Raster to Polygon” tool. This converts the pixels to polygons where the pixels classified as roads have a value of 1 and all the other pixels have a value of 0. By applying a “Select by Attributes” the feature class can be filtered to only include the features that represent what was classified as a road.

In this research the analysis was done to calculate the percentage of the ground truth network that each classification model actually classified as being a road. This was done by clipping the original ground truth network (the OSM road infrastructure network) using the feature class which resulted from converting the classification raster to features. The geometry of the clipped segments can be recalculated as they now display the value from the original ground truth network to display the total length of the classified network in comparison to the ground truth as was done in paragraph 6.2 of this research.

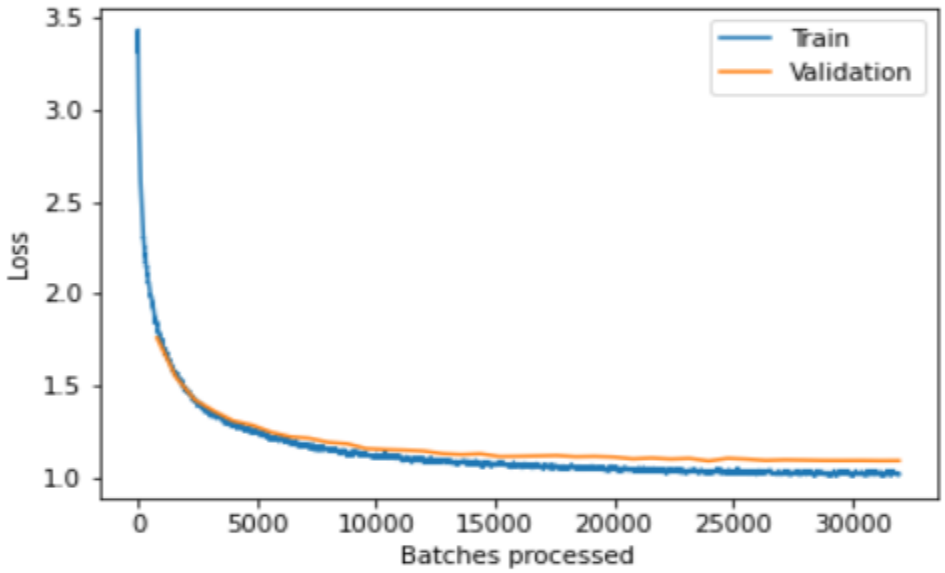
Appendix B: Super resolution model report

SuperResolution

Backbone: resnet34

Learning Rate: slice('1.3183e-05', '1.3183e-04', None)

Training and Validation loss



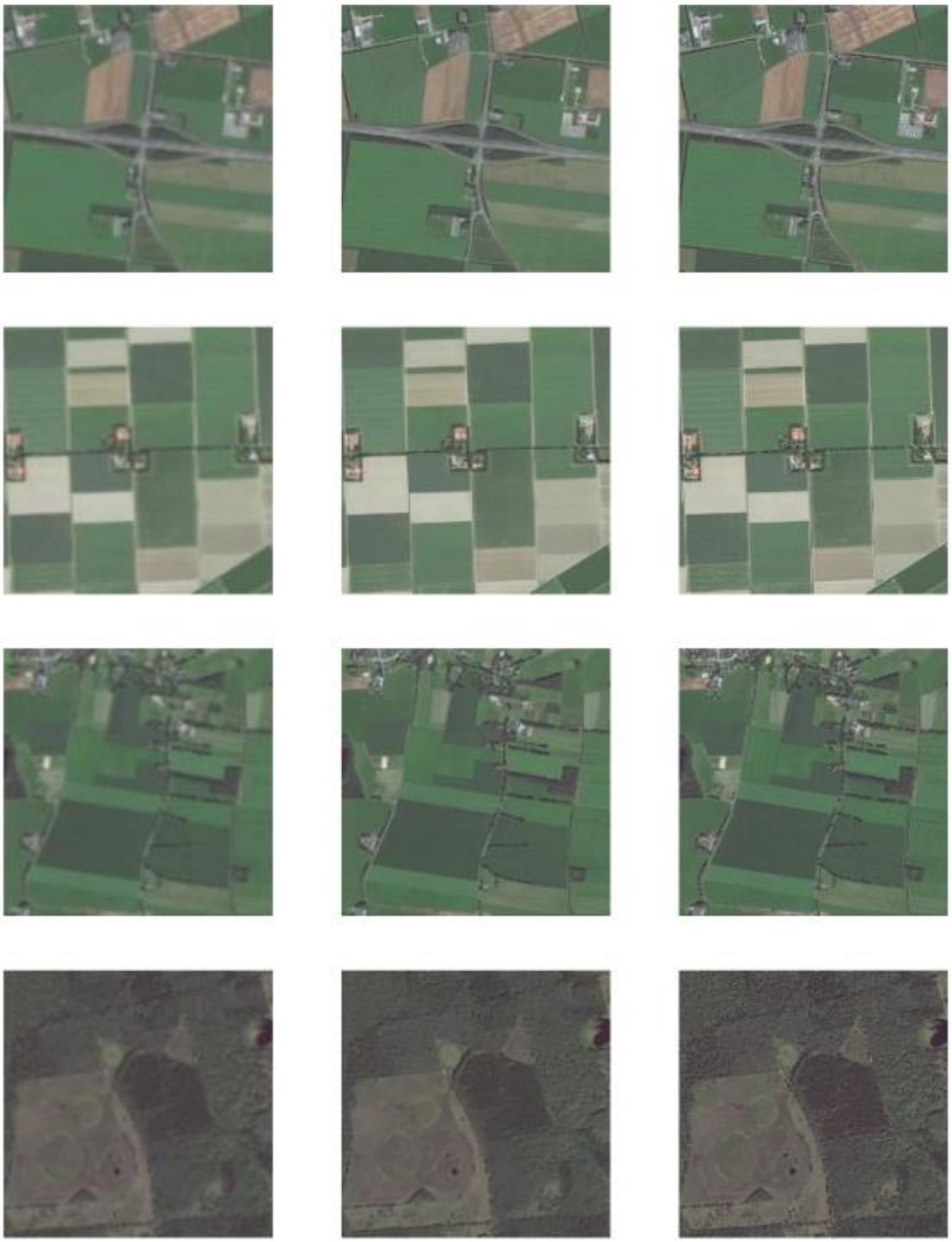
Analysis of the model

PSNR Metric: 6.9145e+00

SSIM Metric: 1.2818e-01

Sample Results

Input / Prediction / Target



Appendix C: Super resolution model metrics data

5250 Synthetic Images	40 Epochs	
LR: 0.00013	PSNR: 6.9484e+00	SSIM: 1.2879e-01
Training:	Validation:	Pixel:
2.033170700073240	1.760602235794060	0.281048864126205
1.811545968055720	1.544457197189330	0.243235886096954
1.684386491775510	1.422185182571410	0.232446596026420
1.617874383926390	1.361915588378900	0.224576190114021
1.542024731636040	1.308670759201040	0.219460293650627
1.537880063056940	1.285739421844480	0.217523589730262
1.497437119483940	1.246843576431270	0.216863811016082
1.472835183143610	1.222208857536310	0.213269665837287
1.458054065704340	1.214183449745170	0.209605574607849
1.433493971824640	1.191991448402400	0.213661044836044
1.422772407531730	1.184286952018730	0.208739534020423
1.395799160003660	1.157637596130370	0.210009962320327
1.391966700553890	1.152575612068170	0.209976166486740
1.381188988685600	1.149610996246330	0.208949461579322
1.369956493377680	1.144802212715140	0.209448620676994
1.366252660751340	1.130436420440670	0.210101038217544
1.359735369682310	1.125568985939020	0.206945896148681
1.359045743942260	1.129621863365170	0.205912783741951
1.345919489860530	1.114618659019470	0.207250446081161
1.345406532287590	1.116010069847100	0.205937117338180
1.354629874229430	1.117824316024780	0.204725772142410
1.366237521171560	1.120125412940970	0.205686658620834
1.346509099006650	1.113852620124810	0.207008719444274
1.346198201179500	1.115591526031490	0.206042245030403
1.352154016494750	1.112213373184200	0.204357549548149
1.335335731506340	1.102210521697990	0.204558610916137
1.344991207122800	1.106792569160460	0.204282283782958
1.333145618438720	1.101207375526420	0.204699620604515
1.334546327590940	1.104724049568170	0.203177571296691
1.320653915405270	1.090755820274350	0.204670384526252
1.342234492301940	1.105195283889770	0.203121542930603
1.333109855651850	1.100565195083610	0.203697621822357
1.325608372688290	1.093747854232780	0.202280834317207
1.324204683303830	1.096756339073180	0.203086927533149
1.326298594474790	1.095253586769100	0.203230232000350
1.325513839721670	1.093355178833000	0.203088015317916
1.324034690856930	1.092905998229980	0.203180164098739
1.322400212287900	1.093108892440790	0.202965140342712
1.322160840034480	1.092016100883480	0.203162372112274
1.322031021118160	1.092350006103510	0.203017473220825

10500* Synthetic Images	40 Epochs	
LR: 0.00012	PSNR: 7.1230e+00	SSIM: 1.5463e-01
Training:	Validation:	Pixel:
1.334762930870050	1.343694329261770	0.216828867793083
1.296176314353940	1.270813822746270	0.212193906307220
1.243793487548820	1.212439656257620	0.213029921054840
1.218218684196470	1.178690552711480	0.210722491145133
1.211939096450800	1.167146444320670	0.206730172038078
1.221743702888480	1.164338707923880	0.211432635784149
1.181287169456480	1.142772793769830	0.208052337169647
1.184316992759700	1.133651018142700	0.204803436994552
1.179010391235350	1.131767988204950	0.205034136772155
1.174649953842160	1.122003078460690	0.204689905047416
1.158336997032160	1.109095215797420	0.206729844212532
1.162215232849120	1.118044734001150	0.203978568315505
1.161971449851980	1.109128117561340	0.205323293805122
1.146059036254880	1.097421884536740	0.202902480959892
1.142165064811700	1.092767357826230	0.204029276967048
1.144057393074030	1.094247817993160	0.201380103826522
1.140143156051630	1.087175965309140	0.200207948684692
1.131739854812620	1.077967882156370	0.199572920799255
1.138974905014030	1.088581800460810	0.197855308651924
1.118444442749020	1.070598363876340	0.198217973113060
1.141301870346060	1.080607891082760	0.198640123009681
1.125033259391780	1.066473245620720	0.196512192487716
1.126885414123530	1.068861365318290	0.198526278138160
1.115805268287650	1.064465880393980	0.196618378162384
1.104901552200310	1.060004353523250	0.195684373378753
1.117203354835510	1.064801812171930	0.197050720453262
1.118956208229060	1.066478610038750	0.196505084633827
1.110285282135000	1.060131192207330	0.196195319294929
1.113738894462580	1.061389803886410	0.195583105087280
1.099035620689390	1.047207951545710	0.196160748600959
1.101677060127250	1.049500584602350	0.196093127131462
1.104009747505180	1.048262357711790	0.196037277579307
1.100862503051750	1.050201654434200	0.195393636822700
1.100501537322990	1.047721266746520	0.195061221718788
1.096801400184630	1.045173764228820	0.195256233215332
1.099804520606990	1.046950101852410	0.194938302040100
1.094547867774960	1.042984247207640	0.195149093866348
1.093515634536740	1.042530655860900	0.195292249321937
1.095183372497550	1.043597936630240	0.194867983460426
1.093645930290220	1.043203115463250	0.195021599531173

21000** Synthetic Images	40 Epochs	
LR: 0.000006	PSNR: 7.0145e+00	SSIM: 1.4050e-01
Training:	Validation:	Pixel:
1.649221897125240	1.583451390266410	0.244664341211318
1.488003849983210	1.412504196166990	0.229087799787521
1.448211193084710	1.356918811798090	0.219970524311065
1.391785860061640	1.297874331474300	0.219702035188674
1.363462924957270	1.271413922309870	0.217985108494758
1.330869674682610	1.241864442825310	0.215699672698974
1.292516946792600	1.214172124862670	0.218532457947731
1.281428217887870	1.208523988723750	0.214546933770179
1.271834492683410	1.191818237304680	0.213484779000282
1.234594821929930	1.165501594543450	0.209986716508865
1.232379198074340	1.164305329322810	0.208730310201644
1.206934571266170	1.147516608238220	0.208479970693588
1.197310209274290	1.134421706199640	0.207552284002304
1.197564363479610	1.135317802429190	0.208318129181861
1.194246411323540	1.130628705024710	0.205346733331680
1.181972861289970	1.123579502105710	0.204405412077903
1.194256305694580	1.126165986061090	0.203412353992462
1.187659025192260	1.127141714096060	0.202882856130599
1.175362110137930	1.119860529899590	0.203212201595306
1.176252245903010	1.119630098342890	0.201539129018783
1.188709020614620	1.110815286636350	0.202269867062568
1.173126101493830	1.104804992675780	0.201040118932724
1.169271826744070	1.103946805000300	0.201677247881889
1.165689468383780	1.102670073509210	0.202177330851554
1.160314679145810	1.092529654502860	0.202385991811752
1.163564801216120	1.102436661720270	0.200283437967300
1.158297896385190	1.092085599899290	0.199929118156433
1.161686897277830	1.096046209335320	0.200684458017349
1.155257225036620	1.088894367218010	0.200298443436622
1.153947591781610	1.087780117988580	0.200645595788955
1.159106731414790	1.090214729309080	0.200294807553291
1.156036734580990	1.087246298789970	0.200282827019691
1.151178956031790	1.085277557373040	0.200265720486640
1.153108596801750	1.083400249481200	0.200228452682495
1.153496146202080	1.082845807075500	0.200407385826110
1.152282238006590	1.084599018096920	0.199835315346717
1.151509761810300	1.084717512130730	0.199717909097671
1.149540662765500	1.083735823631280	0.200126454234123
1.150632143020620	1.084389925003050	0.199496895074844
1.152593374252310	1.084751725196830	0.199956819415092

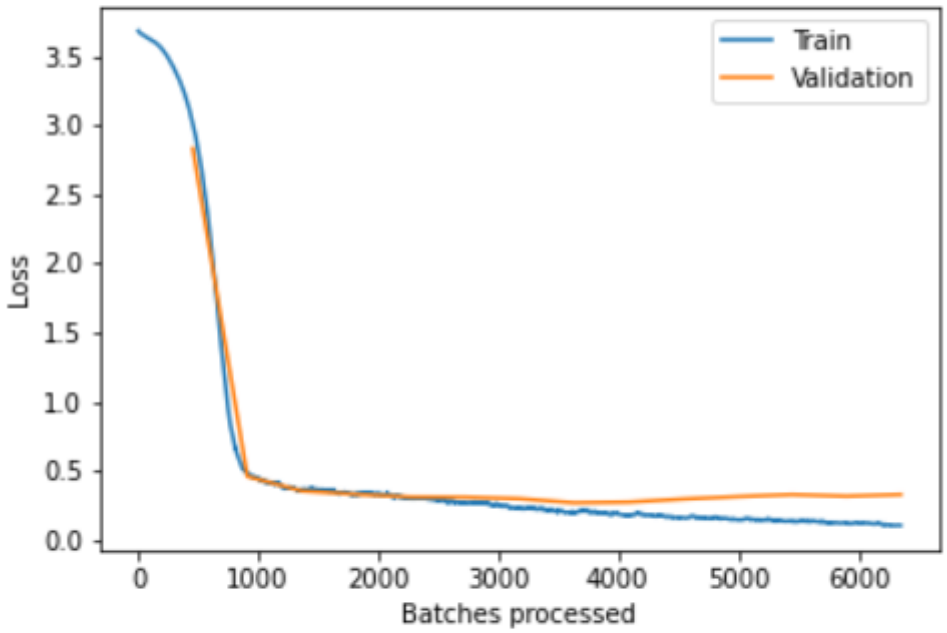
Appendix D: Road extraction model report

MultiTaskRoadExtractor

Backbone: resnet34

Learning Rate: 1.0000e-04

Training and Validation loss

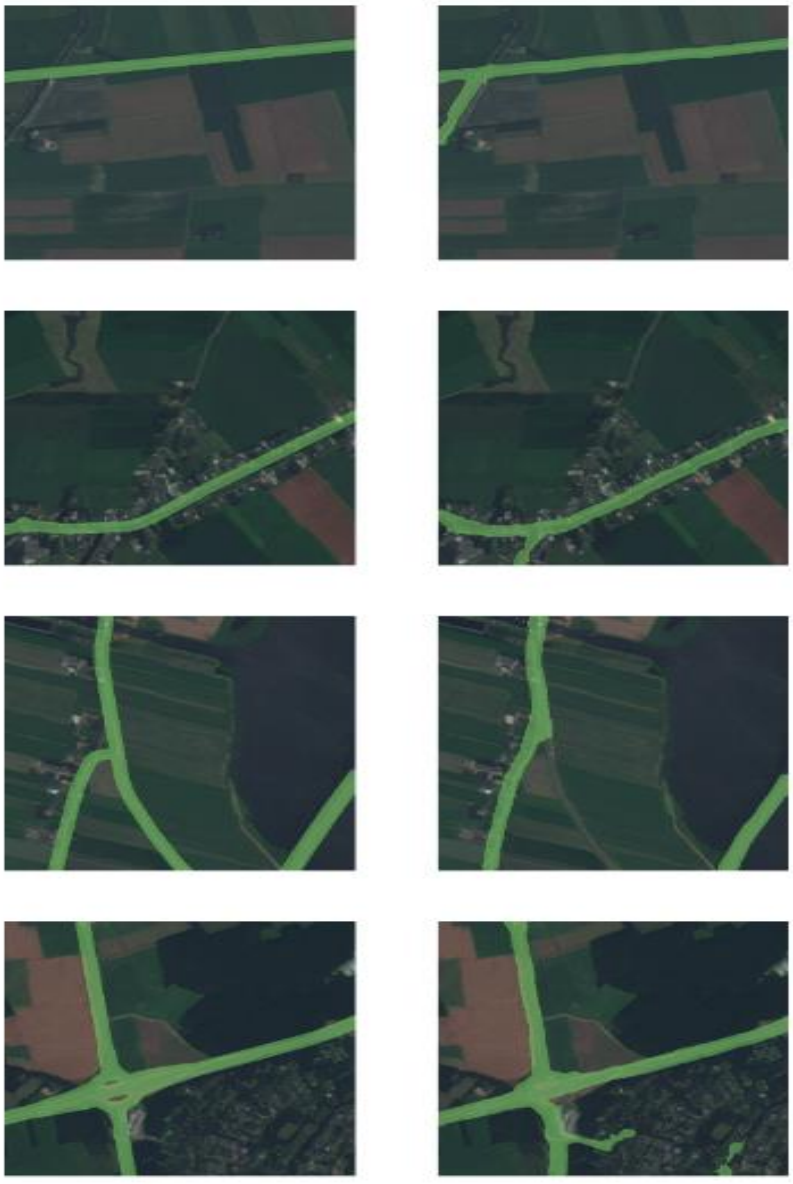


Analysis of the model

mIoU: 0.7280166392942059

Sample Results

Ground Truth / Predictions



Appendix E: Road extraction model metrics data

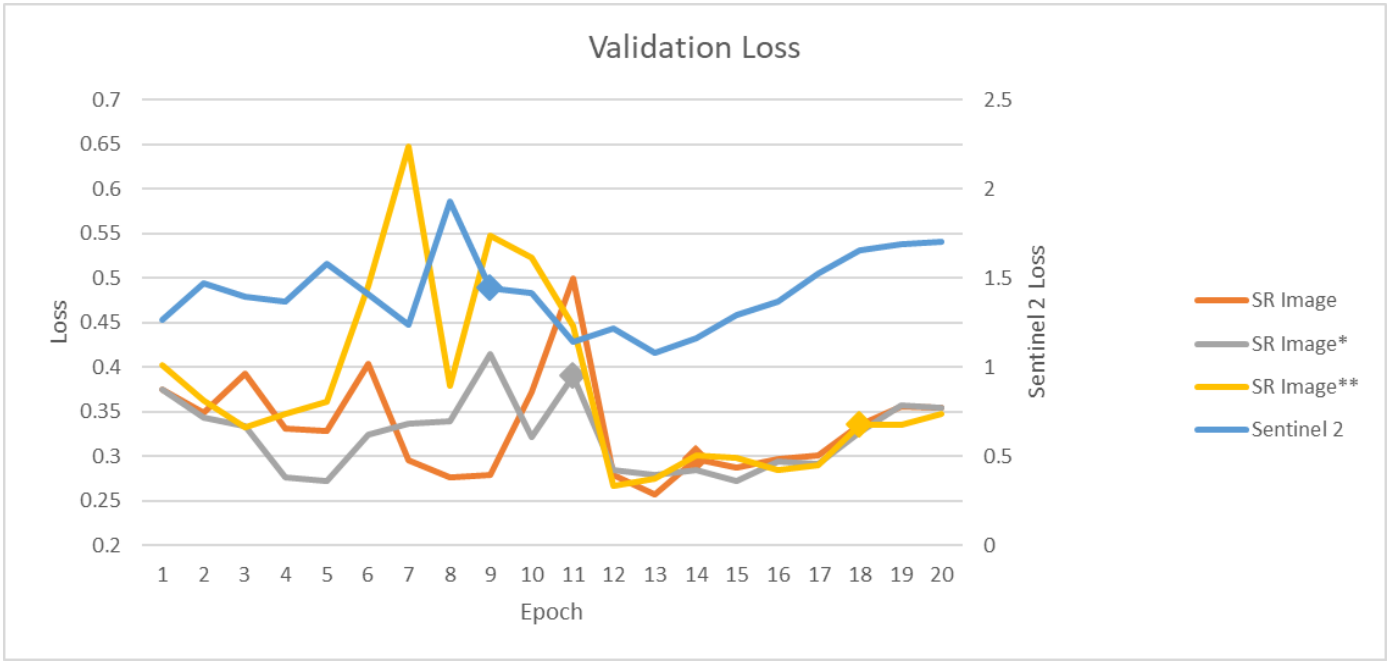
Sentinel 2	LR: 0.0001	MIoU:	Accuracy:
Training:	Validation:	Accuracy:	MIoU:
2.68884635	2.745195389	0.784067392	0.392033682
1.22590363	1.321766973	0.784067392	0.392033682
1.140180826	1.233328819	0.802391112	0.489113722
1.220234156	1.248624921	0.811260879	0.50282802
1.330352068	1.347181439	0.811513543	0.495950869
1.397547722	1.468655825	0.816224635	0.520002495
1.363068104	1.437577128	0.819777727	0.525418727
1.466689706	1.447768927	0.818845391	0.556679478
1.151776195	1.463991761	0.810102999	0.564721203

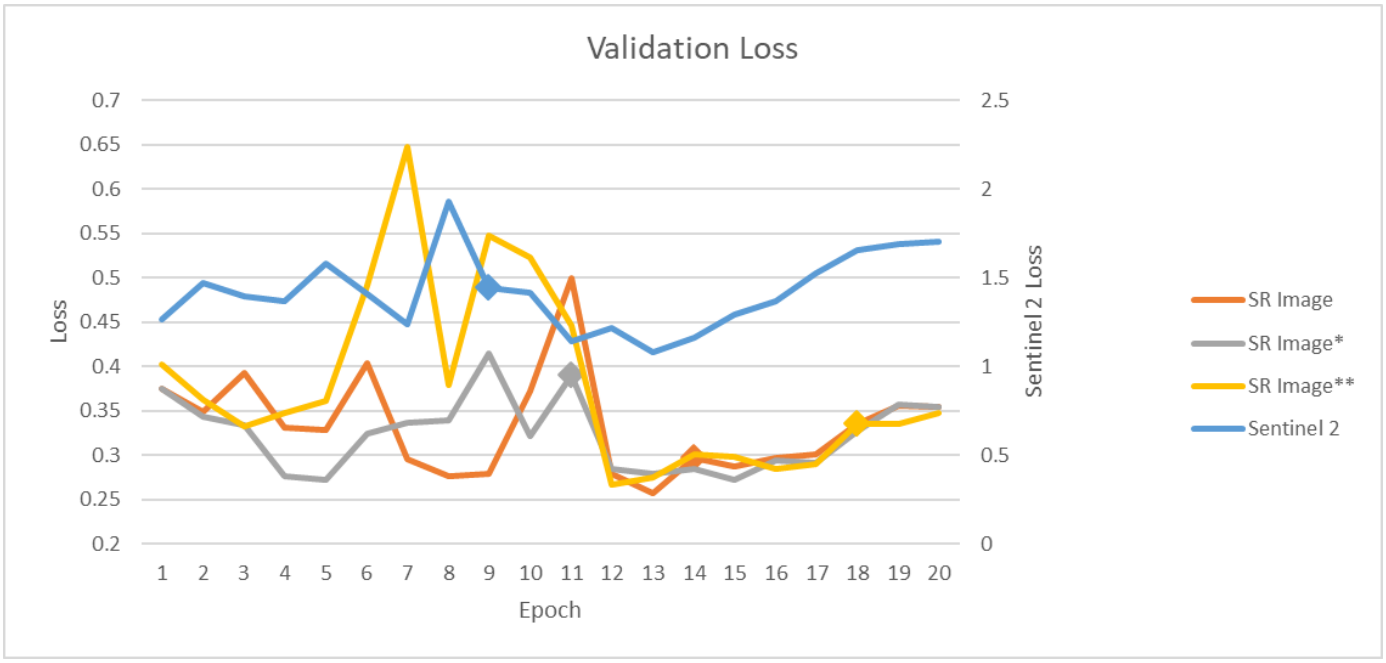
5250 images	LR: 0.0001	MIoU:	Accuracy:
Training:	Validation:	Accuracy:	MIoU:
2.812415838	2.834621668	0.937259674	0.468629828
0.462077826	0.457108945	0.937259734	0.468629832
0.360241383	0.351799548	0.937259734	0.468629832
0.335254252	0.327545315	0.952899814	0.673249019
0.260354787	0.308036327	0.955638289	0.68869152
0.347914875	0.307193518	0.954393208	0.697860604
0.39989391	0.295631319	0.953472137	0.683215105
0.202102616	0.266003102	0.957422674	0.712370652
0.178400993	0.27100718	0.953949451	0.721488879
0.198200896	0.293307334	0.959093034	0.710718252
0.241535932	0.310929149	0.957229137	0.704440933
0.225892097	0.325407445	0.954218745	0.694074046
0.222035825	0.313841134	0.958231628	0.718352456
0.296525538	0.325761646	0.961895168	0.728016639

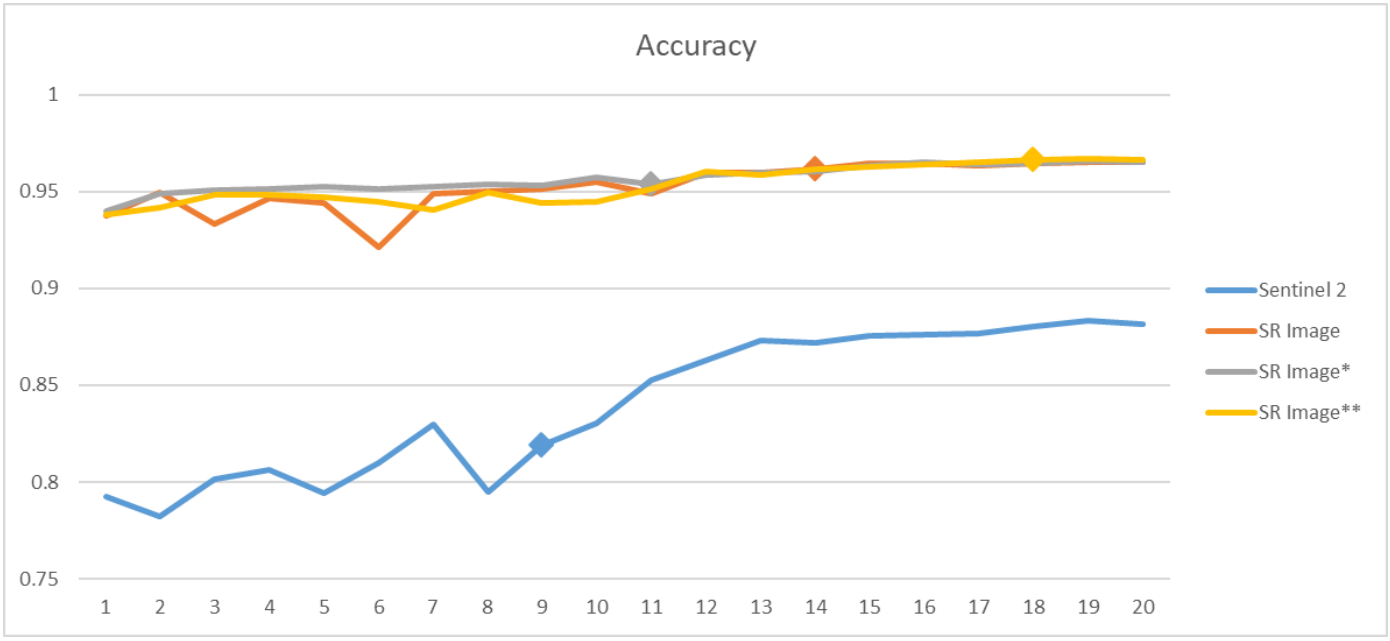
10500* images	LR: 0.0001	MIoU:	Accuracy:
Training:	Validation:	Accuracy:	MIoU:
2.811663151	2.827604294	0.939765036	0.469882482
0.387070149	0.427649677	0.939765036	0.469882501
0.322607577	0.355817646	0.939765036	0.469882501
0.323824376	0.338678032	0.942698479	0.663924338
0.310285062	0.325599879	0.949516177	0.672768938
0.303729177	0.329685956	0.942638218	0.674290006
0.336024225	0.30097577	0.958446503	0.686353203
0.353702724	0.313380539	0.955684006	0.657925461
0.292784363	0.300514758	0.957197189	0.677996486
0.234049439	0.305585265	0.956612051	0.684703801
0.361341715	0.307916969	0.958376169	0.668021335
0.344477057	0.311792374	0.95860672	0.686701736
0.425162464	0.378018528	0.956861913	0.663650633

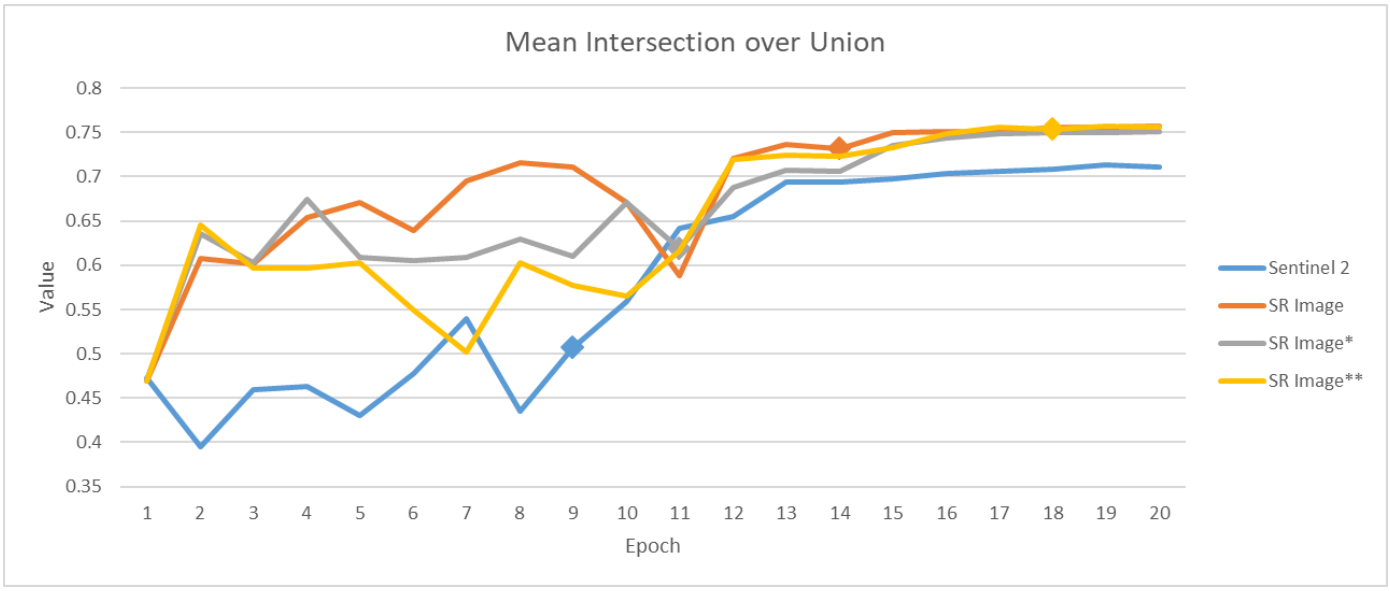
21000** images	LR: 0.0001	MIoU:	Accuracy:
Training:	Validation:	Accuracy:	MIoU:
2.785525322	2.754606962	0.938174009	0.469087017
0.566177368	0.43026644	0.938174009	0.469087026
0.502864778	0.376036823	0.938174009	0.469087026
0.483599782	0.361158669	0.94901818	0.615922925
0.49620986	0.343892455	0.950618625	0.622211701
0.495473206	0.328092247	0.952977121	0.64119481
0.473933339	0.289308459	0.953347564	0.688630748
0.696563005	0.381380677	0.951048434	0.617179201
0.482156277	0.311297119	0.951084256	0.672993719
0.535131037	0.332208931	0.95684135	0.681614779
0.500074148	0.292734921	0.957543194	0.706326388
0.58722043	0.316978514	0.960593343	0.704604461
0.466335058	0.347463399	0.957525074	0.698142502

Appendix F: Road extraction metrics with auto-extracted learning rate






Accuracy
1
0.95
0.9
0.85
0.8
0.75
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Sentinel 2
SR Image
SR Image*
SR Image**


Mean Intersection over Union
Value
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Epoch
Sentinel 2
SR Image
SR Image*
SR Image**