Master Thesis

submitted within the UNIGIS Master's program
"Geographical Information Science & Systems – (UNIGIS MSc)"
to the Department of Geoinformatics - Z_GIS,
Paris-Lodron University of Salzburg

Testing the Cross-Regional Transferability of Mask R-CNN Building Extraction Models

by

Dipl. Geogr. Tomas Tintor
106789

Supervisor:
Assoz. Prof. Dr. Dirk Tiede

In partial fulfilment of the requirements for
the Degree of
"Master of Science", abbreviated "MSc"

Zülpich, 18. December 2024

Science Pledge

I hereby declare that this thesis is the result of my own independent work. I have cited all sources and resources used in this thesis and I have always indicated their origin. This thesis was not previously presented to another examination board and has not yet been published.

Zülpich, 18. December 2024

Tomas Tintor

Abstract

The rapid development of Artificial Intelligence (AI) and the increasing availability of very high resolution satellite and aerial imagery provide great new opportunities for the earth observation domain. One prominent research field is the AI-based automated detection and extraction of buildings on satellite imagery. The presented study focuses on the applicability of already existing models by testing the cross-regional transferability of pre-trained and fine-tuned Mask R-CNN building extraction models on new, previously unseen, target areas in Europe and the United States.

This research underlines the relevance of the target areas´ geographic characteristics for the building extraction results. A certain level of similarity between the training and target areas is a prerequisite for successful model transfer to new geographic locations. Model fine-tuning on two European training areas improved the results of the examined pre-trained models on some but not on all of the experiment target areas.

The application of simple post-processing methods improved the building extraction results of all tested models significantly. Particularly the higher extraction sensitivity of the fine-tuned models could be balanced through the dissolvement of boundaries, deletion of small objects and, the increase of the prediction confidence threshold. The level of improvement through post-processing depends not only on the model´s architecture but also on the specific geographical characteristics of the individual target areas. Satisfactory results after the application of post-processing were achieved on most but not all target areas.

This research examined in addition to the cross-regional performance differences also variances within the individual target areas. Residential areas characterized by single family houses are extracted remarkably well by all models throughout the different geographic locations. Industrial and commercial zones with large, complex buildings, as well as densely build up residential quarters with various apartment buildings, on the other hand, pose significant, cross-regional, challenges to the examined models.

This study demonstrated the complex interrelations between the target and training areas´ geographic characteristics and the cross-regional transferability of AI-based automated building extraction models. Further research on this topic is strongly recommended to further increase the cross-regional applicability of deep learning models.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Various newspaper reported on 08.11.24 about a historic event: NVIDA is the first company in history that surpassed a market capitalisation of 3.6 trillion USD. This is especially remarkable since the American tech company was not known to the larger public until recently. However, the current “Artificial Intelligence (AI) hype” and the fact that the graphic processing units (GPU) of NVIDIA are considered most suitable for the training of AI models, caused a 2000% rise of the companies share value since 2019. The fast development of NVIDIA reflects the exponential rise of the number of costumers: software developers, car manufacturer, the finance sector, military and public administration and many others.

In 2023, Bill Gates, the founder of Microsoft, stated that “the development of AI is as fundamental as the creation of the microprocessor, the personal computer, the Internet and the mobile phone. It will change the way people work, learn, travel, get health care, and communicate with each other. Entire industries will reorient around it. Businesses will distinguish themselves by how well they use it” (Gates, B., 2023). The two key components of this development are the soaring number of AI applications and, in parallel, the decline of technological access barriers resulting in a steadily growing number of AI-users.

In recent years, more and more (tech-)companies massively scaled up their investment in the development of AI models. Multi-billion dollar investments usually require beneficial applications for a large user community. The Generative Pre-Trained Transformer (ChatGPT) from OpenAI, released in Novmeber 2022 to the general public, is a great example for an AI technology with trend-setting applicability and usability. The current trend goes to the introduction of large AI foundation models like METAs “Segment Anything Model (SAM). According to the developers, it is capable of detecting and segmenting any object in an image with a single click, without lengthy and resource intensive additional training. However, by all optimism and achievements, a critical examination of the actual results of AI models is indispensable. Hereby, it is worth to test AI models on realistic applications in order to get a sound assessment whether it provides a significant additional value for users in existing workflows. In addition, does the model generalize well enough to deal with the variabilities of real world scenarios?

1.1 DEEP LEARNING IN EARTH OBSERVATION

Increased data availability facilitated numerous new opportunities within the earth observation research field. The provision of high resolution satellite imagery through

the Copernicus programme or through commercial satellite imagery providers like Maxar Inc. as well as the increasing utilization of Unmanned Aerial Vehicles (UAV) as remote sensing assets, allow an unprecedented perception of the earth (Hoeser et al., 2020; Khelifi et al., 2020). Due to the increased opening of earth observation data archives and the launch of new, very high-resolution (VHR) observation satellites, it can be expected that remote sensing data availability will increase exponentially in near feature (Hoeser & Kuenzer, 2020).

The ground-breaking results of Convolutional Neural Networks (CNN) in the computer vision domain, introduced by Krizhevsky et al. in 2012, provided new opportunities in the analysis and processing of remote sensing data (Hoeser & Kuenzer, 2020). Deep learning and CNN as a specific subtype, can be described as an AI technology, specialised in learning general patterns from large amounts of data as well as exploiting the learned knowledge to solve unknown problems. CNN models have been successfully applied since its introduction for image classification, object detection, semantic segmentation and instance segmentation tasks (Luo et al., 2021).

Object detection predicts locations of objects on an image as bounding boxes and provides a classification label such as buildings, cars or trees. Semantic segmentation segments the whole image into semantic meaningful classes. Instance segmentation extends the scope of the segmentation by detecting and delineating each object of interest in the image like individual buildings (Minae et al., 2022; Yuan et al., 2020).

For segmentation tasks, the contextual relationship of each single pixel is relevant. That information can be found in a range of long and small distances around each pixel. Contextual information depends on multiple factors. The size and continuity of each semantic segment, the relation to neighbouring segments and to the background. Image segmentation can be seen as a multi scale context problem (Hoeser & Kuenzer, 2020). By learning rich contextual information, and by extending the receptive field, CNNs automatically learn hierarchical semantically related representations from input data without any prior knowledge (Liu et al., 2019).

The ongoing increase of computational capacities and the increase of available training datasets, led to state-of-the-art (SOTA) results of CNN models on various object detection and image segmentation tasks (Liu et al., 2019; Minaee et al., 2022). A large variety of models, different setups and post-processing techniques were developed to further improve the performance of CNNs on those tasks (Abriha et al., 2023; El Asri et al.; 2023; Ghorbandazeh et al., 2020; Zhang, et al. 2020). With the unchallenged performance of the new CNN models on natural images, researcher started to replace traditional image segmentation methods with deep learning based approaches using various CNN models (Aleossaee et al., 2023; Shao et al., 2020; Song et al., 2023; Wurm et al., 2019). Additionally, to the processing of natural images, numerous applications were developed for remote sensing imagery. The automated

generation of land use – land cover maps, monitoring of glacier decline or wildfires, ship tracking, extraction and monitoring of road networks or the localization of military equipment in conflict zones are just a few possible applications (Hoeser et al., 2020; Khelifi et al., 2020).

However, within the field of AI, satellite images are considered as challenging data. They cover large areas and contain a great number of different object classes with varying characteristics (Aleossaee et al., 2023; Tahir et al., 2022). Significant objects on satellite imagery are often very small, consisting of only around 15 pixels in high-resolution images and are often densely clustered. Additionally, objects in satellite imagery may have different orientations and distortions depending on the camera axis during the image capturing (Tahir et al., 2022). Occlusion issues and partial coverage of relevant objects by other larger objects like buildings or trees, regularly complicates object detection and extraction on satellite imagery (Sublime et al., 2019).

1.2 DEEP LEARNING-BASED BUILDING EXTRACTION

The automated extraction of buildings from remote sensing data is a dynamic research field and a great example for the development and implementation of AI models in order to improve existing workflows and processes of a large and active user group (Abriha et al., 2023; Han et al., 2022). The extraction of buildings from aerial and satellite imagery is a semantic segmentation or instance segmentation task. It plays a key role in various applications such as 3D urban planning and modelling (Li et al., 2022; Luo et al., 2021; Sakeena et al., 2023), spatial simulation, map services (Blijecki et al., 2023; Li et al., 2021; Prakash et al., 2022), autonomous driving (Kang et al., 2022) or rapid population estimation (Li et al., 2019). Nurkarim et al. (2022) points to another interesting possible application of automated building detection; it can help governments to find illegally constructed buildings to adjust tax rates to the actual conditions on the properties. Lastly, automated building detection and extraction can significantly contribute to effective disaster management and damage assessment (Li et al., 2019; Nex et al., 2019; Valentijn et al., 2020).

The manual delineation of buildings by visual interpretation of human analysts, especially in case of large geographic areas, is laborious and expensive. Volunteered Geographic Information like the OpenStreetMap (OSM) project provides precise and up to date building data for certain geographic areas meanwhile other regions are only partially captured (El Asri et al., 2023). Thus, there is a large and growing interest in research, humanitarian assistance, public administration and the economy domain for an automated building footprint extraction which provides timely and reliable results, preferably worldwide and at low cost (Liu, 2019 et al.; Prakash et al., 2022).

1.3 CHALLENGES

The ascending computational capacities (e.g. stronger GPUs) and growing data availability render further improvement of the building detection and extraction results possible (Kang et al., 2022; Yu et al., 2023). On the other hand, the automated extraction of buildings on satellite and aerial imagery entails numerous technical challenges requiring further improvement of the entire process.

First of all, building extraction from remote sensing images remains a challenging semantic or instance segmentation task because of the large within-class and small between-class variance (Qiu et al., 2022). Buildings differ in interior tones and textures; shape and colour may vary from building to building. Even within the same type of building (e.g. residential), difference can be significant. This is particularly the case for different geographic areas (Han et al., 2022; Wang et al., 2023; Xie et al., 2019). Especially urban areas are considered as complex environments and particularly challenging for the automated building extraction task. Urban surfaces are highly heterogenous with many different materials in close proximity to each other. The contrast between buildings and other objects is often low (El Asri et al., 2023; Ayala et al., 2021). Increasing spatial resolution reveals even more objects for differentiation and delineation (Bakirman et al., 2022). Additionally, higher buildings tend to obscure each other visually and the casting of shadows can easily mislead segmentation algorithms (Abriha et al., 2023; Han et al., 2022).

An accurate delineation of adjacent buildings or within close proximity to each other is a demanding task even for experienced human image analysts (Luo et al., 2021). Although numerous model variants and post processing workflows were proposed to receive sharp and precise building boundaries, incomplete building footprints and inaccurate boundaries are still a common problem (Shao et al., 2020; Qiu et al., 2022). Additionally, the heterogeneous sizes of buildings pose for all current state-of-the-art object extraction models a great challenge (Ji et al., 2018; Sakeena et al., 2023).

Another, more general, problem is the long-standing lack of sufficient training samples despite the large volume of remote sensing imagery. For computer vision task on natural images, crowdsourcing strategies were successful in labelling millions of training samples. However, experienced professionals are required to achieve a satisfactory accuracy in generating training samples on remote sensing imagery (Martin et al., 2021; Yuan et al., 2020).

1.4 CURRENT STATE OF THE RESEARCH

Traditional automated building footprint extraction methods like Random Forest (RF), Support Vector Machine (SVM) and Maximum Likelihood Classifier (MLC) rely

primarily on specific, manually by humans designed, features like the texture and geometric features of buildings. However, due to the relatively small model sizes, only a limited number of relevant features are processed. The deeper and more abstract features of building footprints and the spatial relationship between buildings and backgrounds cannot be fully represented in those shallow models (Han et al., 2022; Song et al., 2023). The mentioned enormous diversity in building appearance, the different environmental characteristics and the variety of sensors and scales puts the focus on the generalizability of building extraction models. Here, manually designed metrics of traditional methods, have a clear disadvantage (Ji et al., 2018; Yang et al., 2021). Shallow machine learning methods such as RF and SVM can achieve high accuracy on a specific dataset but the transferability to other geographic areas is generally a problem (Yang et al., 2021). Although CNN have generally increased the performance in the building extraction task compared to previous traditional methods, the generalization ability of deep learning models, when faced with diverse building structures and architectural styles from all over the world, is still an open research question (Ghorbanzadeh et al., 2019; Ghorbanzadeh et al., 2020; Luo et al., 2023; Tiede et al., 2021).

It is common practice in the CNN based building extraction domain, that model training and performance evaluation is executed on certain benchmark datasets. This is helpful to compare the performance of different models, although variances in the respective experimental setups and performance metrics pose a persistent challenge for a profound comparison (Minaee et al., 2022; Neupane et al., 2021). On the other hand, most of the relevant benchmark datasets have a narrow geographical scope, covering often just one or a few distinct cities. The vast majority of the recent building extraction research is conducted on a few, openly available data sets like the SpaceNet, INRIA, WHU, ISPRS and Massachusetts data sets (Li et al., 2024; Luo et al., 2021; Luo et al., 2023).

Building extraction results are reported by specific performance metrics like the F1 and the Intersection over Union (IoU) score. For both parameters, the minimum value is 0, and the maximum value is 1, which indicates, that the model extracted all relevant objects without any false proposals.

Liu et al. (2020) tested different building extraction models on the INRIA dataset achieving an IoU between 0.69 and 0.77 and IoU scores between 0.64 and 0.91 on the WHU dataset. Liu et al. (2019) repots likewise for various models an IoU between 0.74 and 0.81 on the Massachusetts dataset and IoU values between 0.73 and 0.8 on the INRIA dataset. The models reported by Xie et al. (2020) achieved an IoU between 0.69 and 0.74 on the Massachusetts dataset and an IoU between 0.7 and 0.79 on the INRIA dataset. An IoU between 0.88 and 0.94 is reported by Li et al. (2021) on the

ISPRS dataset. El-Asri et al. (2023) reported for the INRIA dataset a maximum IoU score of 0.67 and an F1 score of 0.8 using various model evaluation scenarios.

1.5 GEOGRAPHIC TRANSFERABILITY

It is generally acknowledged that the training and the subsequent testing of a deep learning segmentation model on the same dataset achieves the highest performance (Calton & Wei, 2022; Maggiori et al., 2017; Nex et al., 2019). However, numerous studies reveal that the application of a model on new, previously unseen datasets often leads to a significant performance degradation (Bouchard et al., 2022; Neupane et al., 2021).

Transferability in the remote sensing domain deals with the challenge of applying a model that was trained on one geographic region on other geographic locations. Differences between the two locations can occur due to differences in sensor parameters, illumination and atmospheric conditions or environmental characteristics (Kucharczyk et al., 2020; Li et al., 2019). Gella et al. (2023) and Gao et al. (2022) also report from strong performance degradation when a model is transferred to a completely different geographic region. The disappointing results are attributed to variations in the background environment, site-specific differences in the characteristics of the settlements and spectral variations by changing seasons. Nex et al. (2019) examined the performance of multiple models in different test scenarios, concluding that best results were achieved when the model testing occurred on the same geographic area as the training of the models. Shao et al. (2020) tested the performance of 4 different building extraction models on a new, unseen dataset. The models achieved F-1 scores between 0.31 and 0.43. The authors explain the rather poor performance of all four models with the fact, that the training and test datasets cover different geographic regions and consist of images with different spatial resolutions.

The study of Li et al. (2019) on the DeepGlobe Challenge dataset revealed significant differences in the results for the four test areas: an F1 score of 0.89 for the Las Vegas test area, 0.74 for Paris, 0.62 for Shanghai and a relatively low F1 score of 0.54 for the Khartoum test area. The authors explain the high variation in performance of the same model on different test datasets with the diverse geographic characteristics of the four cities. The buildings in Las Vegas and Paris test areas have a more unified architectural style and consist predominantly of residential buildings. The Shanghai test area covers more high-rise buildings, an extensive construction site and more industrial areas. The low result on the Khartoum test area meanwhile is caused by the tremendously dense and often irregular composition of the residential buildings.

Similarly, Sakeena et al. (2023) report for a Mask R-CNN based building extraction model, F1 scores of 0.88 for Ls Vegas, 0.76 for Paris, 0.64 for Shanghai and only 0.58 for Khartoum. Li et al. (2024) also stated that a building extraction model, which is trained on samples from European cities, performs badly on a test dataset consisting of African cities. Likewise, Luo et al. (2023) reported very low scores on test areas located in Khartoum and Potsdam for a model which was trained and tested successfully on a New Zeeland dataset (i.e. WHU dataset).

Numerous other studies also report a negative impact of certain geographic characteristics like particularly small or large buildings or buildings with irregular shapes on model performance (Chen et al., 2021; Yi et al., 2019). Also, Luo et al. (2023) state that models performing well on the Massachusetts and WHU datasets can suffer a severe performance decrease on test areas with different characteristics. Moreover, according to the authors assessment, the above mentioned standard datasets lack the ability to serve as benchmarks to evaluate the generalization ability and performance of a model for more practical scenarios where the target areas cover other cities than the few selected areas covered by those datasets. According to Xie et al. (2019), the problem of model generalization capability is currently the biggest challenge in the deep learning-based building extraction research domain.

The terminology “Geographic Transferability” is often used in the building extraction research domain (Nex et al., 2019). In the context of performance evaluation, a high geographic transferability indicates high generalizability of a building extraction model. The ultimate test of the model accuracy in the building extraction context comes when it is applied on geographic regions that the models have not been trained on (Xu et al., 2019). Unfortunately, the generalization performance of proposed models is rarely tested (Duarte et al., 2018; Ji et al., 2018; Qiu et al., 2022).

From the user perspective, the generalization capability of a model is indispensable for large-scale applications (Li et al., 2024). For example, in the disaster management domain, deep learning models are often trained and tested on building samples captures from a specific study area. If an earthquake occurs in a geographic region where the building characteristics significantly differ from that location, the model performance on the new, previously unseen target area, is unknown (Yang et al., 2021). A high level of geographic transferability is generally desirable for building extraction applications. The alternative, namely the training of individual models for each and every geographic region, is not practicable. That approach would not only require enormous resources, it would also fail to provide timely results. In a disaster relief context, for example, time is extraordinarily critical as every hour after the occurrence of an earthquake counts (Gella et al., 2023; Wiguna, et al. 2024).

In the literature, two methods are discussed to enhance the generalizability and geographic transferability of CNN building extraction models. Firstly, a high diversity of

the training dataset allows the model to better learn invariant and robust features. Nex et al. (2019) and Li et al. (2024) for example, amplify the need for a more diverse training dataset with a balanced number of samples from different environments and building characteristics. Additional samples from regions which are not well represented in the current training datasets, would increase the generalizability of the models leading to more reliable results on new, unseen locations. Yang et al. (2021) also recommends to consider the generalization capability of a model already during its training. CNN models for damaged building identification for example, should learn from sufficient buildings samples covering different geographic locations. Results show that the performance of the models is affected by the composition of training samples used for the training. Models trained with samples covering different locations, performed best. Luo et al. (2023) emphasizes, that not only the quantity of the training data, but also its diversity has a great impact on model generalizability.

Gupta et al. (2019) likewise recognized the necessity for generating a comprehensive building damage database of accurately labelled remote sensing images for model training. In 2019, the xBD dataset was published, covering 8 disaster types in 15 different countries with around 700.000 building annotations from rural and urban areas. Since its introduction, xBD is regularly used for the training and testing of destroyed building detection models. According to Martin et al. (2021), the availability of the xBD dataset with its global dimensions could be considered as a first step towards achieving a general transferable model for building detection and extraction. However, Yang et al. (2021) tested eight pretrained models with the diverse xBD dataset on new, unseen geographic regions not included in the training dataset. In those test cases, the building detection accuracy was still error-prone. The authors consider the differences in size, structure and surrounding environment of the buildings between the training and test areas responsible for the weaker results.

The other promising approach to increase the geographic transferability of building detection and extraction models, is the fine-tuning of existing models (Nex et al., 2019). To bypass the data-intensive training of CNN models, they can be trained on available data from one geographic region and later fine-tuned with a relatively small amount of data to a new target region (Gella et al., 2023; Xu et al., 2019). Although most of the studies on model fine-tuning report an improvement of the model´s geographic transferability, the exact mechanisms of that process are still not fully encompassed (Zhang et al., 2021).

2. RESEARCH MOTIVATION

As the automated semantic and instance segmentation on remote sensing data improves, the focus is shifting from academic proof-of-concept studies towards the applicability of AI models on realistic applications (Hoeser et al., 2020).

A perfect model for a worldwide, immediate and highly accurate building extraction, does not exist yet (Bai et al., 2020; Bakirman et al., 2022). In addition, numerous (potential) user do not have the capacities and capabilities to develop for each and every new case and target area specific and perfectly matching models from the scratch. The user´s challenge is to choose from the growing number of available models those that suit their specific purposes. An appropriate assessment of the model´s capabilities and limits are inevitable for reasonable decisions. For automated building extraction models, the assessment of performance variations across different geographic locations is one major concern in order to estimate whether the model can be reliably used on different target areas (Valentijn et al., 2020). Thus, the motivation for this research is to place the (future) users of AI into the centre by turning the focus on the assessment of already existing automated building extraction models.

2.1 RESEARCH OBJECTIVES

This research aims to examine the transferability of building extraction models on new, previously unseen target areas. The focus lies on the interrelation between model performance variations and differences in the geographic characteristics of the training and target locations. Two existing building extraction models and three fine-tuned variants of them, are tested on multiple target areas. All examined models share the same architecture and parameters. The distinguishing feature is the geographic location of the individual training areas. This setting enables to examine the impact of the selected training areas on the model´s geographic transferability to new, previously unseen target locations. Numerous previous studies conducted cross-regional model performance assessment of pre-trained and fine-tuned building extraction models (Li et al., 2019; Luo et al., 2023; Sakeena et al., 2023) This research goes one step further by analysing model performance variations also within close geographic distances. Some of the selected target areas are located within the same geographic region, the same country or even the same city. In an additional research step, model performance variations are examined within the very same target areas.

Another aim of this research is to examine the impact of post-processing methods on the building extraction results across the various target areas. Here, the interrelation

between the geographic characteristics of the target areas and the positive effects of post-processing are analysed in detail.

2.2 RESEARCH QUESTIONS AND HYPOTHESES

The first research question addresses how geographical distance between training and target locations impact the building extraction models´ performance. The underlying hypothesis is based on Tobler´s First Law of Geography which states that everything is related to everything else, but near things are more related than distant things. The *geographical proximity* between the model´s training and target area is unambiguously measurable in kilometres, miles etc. This distinct parameter could be a great support for user to estimate the expected performance of available models on new target areas. Thus, the first research question and the corresponding research hypothesis are:

- **R 1.1**: What is the impact of geographical proximity between the locations of the training and target areas on model performance?
- **H 1.1**: Model performance is high on new, previously unseen target areas which have a close geographical proximity to the model´s training area. Model performance decreases with declining geographical proximity between the training and target areas.

The *geographical closeness* of locations focuses on geographical similarity instead of the spatial distance. Moscow and Vladivostok, for example, are considered geographically closer than Moscow and Munich, although the spatial distance is six times higher. Historical, political, cultural, architectural and environmental aspects have to be considered by the determination of geographical closeness. The drawback of such an approach lies in its arbitrariness. This study defines the geographical closeness of the utilized target and training areas; however, this is not a unanimous consensus but an individual opinion. The first research question is extended to examine the impact of geographical closeness on the model´s cross regional transferability:

- **R 1.2**: What is the impact of geographical closeness between the locations of the training and target areas on the model performance?
- **H 1.2**: Model performance is high on new, previously unseen target areas which exhibit a large geographical closeness to the model´s training area. Model performance decreases with declining geographical closeness between the training and target areas.

The second research question focuses on the building extraction process itself. It is not unusual in the remote sensing segmentation domain to conduct post-processing to improve the final extraction result. In this research, three different post-processing

methods are tested on the five examined models and ten target areas. The following three post-processing methods are executed:

1. Utilization of a “Dissolve Boundaries” tool to counter over-segmentation of the identified buildings.
2. Exclusion of all predicted objects smaller than 25 m².
3. Increase of the prediction confidence threshold from 50% to 80%.

Other studies utilized similar post-processing methods to clean the results from undesirable artefacts and errors (Bakirman, 2022; Li, 2019).

This research question examines possible cross-regional improvement variances of post-processing.

- **R 2**: Do specific post-processing workflows achieve the same result improvement throughout different geographic regions?
- **H 2**: Building extraction post-processing will improve the results in dependence of the local geographic characteristics. The exclusion of predicted objects smaller than 25 m² for example, is considered particularly effective on suburban areas with numerous small sheds and garages but less effective on locations with large buildings like apartment blocks or central business districts.

Finally, the third research question examines the impact of certain geographic characteristics on the model´s performance. For the building extraction task, the building type (e.g. single family house, apartment buildings, apartment blocks, industrial buildings), urban structure (e.g. building density, homogenous urban quartets, separate Central Business District, historical town centre) or local environment characteristics, are of particular interest. Here, the assumption is, that certain specific feature attributes are more challenging for building extraction models than others.

- **R 3.1** : What is the impact of certain geographic feature attributes like building types and urban structures on the performance of building extraction models?
- **H 3.1**: Automated building extraction works generally better on target areas with certain geographic characteristics like residential areas with mostly single family houses.
- **R 3.2**: Are the impacts of certain geographic features consistent throughout different geographical regions?
- **H 3.2**: Model performance on target areas with certain specific geographic characteristics is consistent throughout different geographic regions.

3. MATERIALS AND METHODS

3.1 RESEARCH METHODOLOGY

The following chapter is subdivided in three blocks. The first part deals with data acquisition and processing. The second part describes the model fine-tuning. Finally, the model evaluation methods of this research are presented.

Throughout the entire research, the Arc GIS Pro 3.1 software of ESRI Inc. was utilized. Sawa et al. (2024) consider it as a standard software, used for GIS applications by more than 350.000 organisations. With the integration of pretrained CNN models for object detection and extraction, ESRI provides an interface which enables the execution and examination of deep learning applications without advanced coding skills. ArcGIS software is used in this research likewise for data acquisition and the post-processing of the building extraction results.

3.1.1 GENERAL DEEP LEARNING BUILDING EXTRACTION WORKFLOW

The foundation for the entire automated building extraction process is the acquisition of sufficient training and test data. CNN models require generally three sets of data for model training, validation and testing. The training data is used to determine the parameters of the network, its weights and connections. In the model training, one iteration over the entire training dataset is called an epoch. After each epoch, the model performance is assessed against the validation data which is not used for the training. The validation is used to determine, if additional iterations would be useful to further improve the model´s performance. The number of epochs depends on many factors but it can reach hundreds. Once the model training is finalised, the test dataset is used for the accuracy assessment (Maxwell et al., 2021). Data acquisition can be significantly reduced through the utilization of pre-trained models.

Having the required quantity and quality of training, validation and test data available, the training of the CNN model can start. Model training does not only require expert knowledge but it is a lengthy and resource intensive process. It consumes enormous processing capacities and necessitates appropriate high-performance hardware, such as advanced GPUs. With model size and complexity, the hardware requirements for successful model training can quickly exceed the capacities of individual researchers, universities and public institutions. Once the trained model achieves satisfactory results on the validation dataset, the model performance evaluation is conducted on the prepared test dataset.

3.2 DATA

The exact effect of training data on the model performance, is not known in its entirety. However, it is generally agreed upon that appropriate input data is essential for the development of competitive CNN models (Huang et al., 2019). If the size of the data is too small, too noisy or not representative enough, the model will probably perform relatively well on that specific dataset but will likely generalize poor to new, unseen test data, resulting in low model performance (Abriha et al., 2023, Hoeser & Kreuzner, 2020). Thus, for a successful building extraction, high-resolution imagery and the corresponding reference building footprint data in sufficient quantity and quality, is indispensable.

3.2.1 SATELLITE IMAGERY

For this research, all of the utilized satellite imagery, is downloaded from governmental institutions free of charge. The imagery is published as an online web service. The customer can outline the requested area with a bounding box and download the imagery as a high-resolution image file. The standard image processing for spatial analysis like geo-referencing and orthorectification are already realised by the data provider institutions. The provided documentation on the website reveal the most important image parameters like resolution, acquisition date and sensor type. Appendix 1 provides the details to the obtained imagery. and lists the source web portals.

Optical satellite and aerial imagery consist of multiple bands, each capturing different parts of the electromagnetic spectrum. Most of the deep learning based remote sensing segmentation and object extraction studies are carried out on natural colour images with three bands (i.e. red, green and blue bands). (Hoeser et al., 2020). The images utilized for this research consist of the standard three bands (RGB) following the research mainstream. This is particularly important as the utilized pretrained Mask R-CNN models are also trained on RGB images. In one case, an existing fourth band (i.e. IR) was manually deleted to comply with the model´s requirements.

The camera angle of the sensor during the image recording, has an important impact on the building extraction process. The Off-Nadir angle refers to the angle between the sensors nadir, which is directly below the sensor, and the point on the earth being observed. A value of zero off-nadir is reached when the sensor is directly above the area of interest. As the sensor moves farther away, the off-nadir value increases correspondingly. For the building extraction task on remote sensing imagery, a low off-nadir value is highly recommended. Figure 1 demonstrates successful

building extraction on a low-off-nadir image and, on the other hand, it depicts the negative extraction effect caused by a high off-nadir value. The model extracted not only the building footprint (i.e. the roof geometry) but also the entire building façade. Low off-nadir imagery eases the matching between the imagery and the corresponding building footprint reference data tremendously. Another benefit of low off-nadir imagery is the avoidance of building occlusion through other, nearby buildings.

Figure 1: Off-nadir impact on building extraction. Successful building extraction on low off-nadir image is shown on the top. The negative effect of a large off-nadir is shown on the bottom.

The exact off-nadir values of the obtained images for this research cannot be determined. That is because the governmental institutions used as image source for this research, do not publish single images but a compilation of selected and pre-processed imagery. Visual inspection indicates however that the image collections are co-registered and have in general low off-nadir values.

For the building extraction task on remote sensing imagery, the spatial resolution of the images is of high importance. It is commonly measured in meters or centimetres per pixel. A 30 cm spatial resolution means that each image pixel corresponds to an area on the ground with a size of 30 x 30 cm. Consequently, the smallest distinguishable detail captured by the sensor is approximately 30 cm. Obviously, buildings are much larger than 30 cm. However, the registration of fine details in shapes, edges and textures of buildings, is inevitable for successful delineation and extraction of buildings. Most of the obtained imagery for this research has a spatial resolution between 20 and 30 cm. Only the images from the USA have a lower

resolution of 50 cm, which is still considered as good enough for accurate building extraction.

Lastly, the most important image characteristic is its content. Although trivial at first view, this point is of utmost importance for the training and testing of building extraction models. In the literature, it is generally accepted that the training imagery should cover a high variety of different building types and urban scenes to increase the generalization capacity of the segmentation models. Rastogi et al. (2020) for example use for their research training data covering a heterogeneous set of building types within urban areas including densely built-up areas, isolated detached houses and urban slums with buildings spatially very close to each other making even a visual delineation difficult. The imagery obtained for this research cover likewise various urban scenes with a high variety of building architectures and compositions. Different building types like single family houses, apartment blocks, historical buildings, Central Business Districts and large industrial zones are present in the training and test imagery (Figure 2).

Figure 2: Daugavpils target area. It contains a large variety of different building types at close distance.

3.2.2 REFERENCE LABELS

Many of the best performing remote sensing deep learning object extraction models, including the Mask R-CNN model used for this research, are based on a supervised learning approach Supervised learning means in this context a method where the model learns from labelled training data. During the training process, the model is

provided with input data (i.e. imagery) and the corresponding correct output (i.e. reference labels). The model adjusts during the training independently, without human experts' contribution, its internal parameters (i.e. weights), based on the observed errors, aiming to minimize the difference between its predictions and the reference. This procedure allows the model to learn to generalize in order to make accurate predictions also on new, unseen data (Hoeser & Kuenzer, 2020). The drawback of supervised learning is the acquisition of the reference labels in sufficient quantity and quality. Only a handful (benchmark) datasets provide satellite and aerial imagery with the corresponding reference labels for model training and testing. Accurate reference labels are not only required for model training but also for the subsequent model performance evaluation. The predictions of the model are compared against the reference dataset to find the features not accurately predicted by the model or to identify areas falsely depicted as relevant features. The reference labels must fit to the imagery, and not vice versa. This implies that each satellite and aerial image requires its own individual reference dataset (Abriha et al., 2023; Huang et al., 2019; Li et al., 2022).

Manual labelling of the imagery is one, very laborious, way to obtain the reference data. For the building extraction task, this means the visual inspection of the entire image and manual digitalization of each building with a labelling software. This process does not only take, depending on the image size, a huge amount of time but it requires also experience and concentration to exactly detect and delineate all relevant features (El Asri et al., 2023; Touzani et al., 2021). Additionally, the manual labelling of building footprints is a rather subjective task. Mistakes can occur easily during the labelling process and the model has to deal with those ambiguities (Valentijn et al., 2020). Another possible approach is to use already existing reference labels and align them to the utilized imagery. This is usually less time intensive than the manual labelling of the entire image from the scratch. However, a 1:1 matching and perfect alignment is not possible and manual correction is still required. The amount of the adjustments depends on the quality of the existing reference dataset and the difference between the actual image and the image on which the reference labels were created. Newly constructed buildings, missing on the reference dataset, have to be added, and spatial distortions caused by different off-nadir angles, corrected through manual shifting of building footprint polygons (Li et al., 2024).

Numerous studies use the feature labels of the OpenStreetMap (OSM) project as reference data instead of manually label all relevant objects on the imagery (Ayala et al., 2021; Biljecki et al., 2023; Ghaffarian et al., 2020). OSM is a voluntarily contributed, worldwide, geospatial database. Since its start in 2004, millions of contributors have mapped billions of features like roads and buildings. The mapping is based on field surveys, aerial and satellite images and local knowledge. The database is released on

open license, and freely downloadable (Biljecki et al., 2023). However, a well-known persistent disadvantage of the OSM database is its varying level of quality (Parkash et al., 2022). Li et al. (2022) and Spasov & Petrova Antonova (2021) among others examined the quality of OSM in terms of completeness and accuracy and found that it varies significantly among different geographic regions. OSM data coverage can be considered high, especially for urban areas, in the USA and Europe, reaching in some cases almost 100% completeness, meanwhile in other parts of the world, data only rudimentary exists.

This research utilized OSM feature labels for eight, and official governmental building labels for two target areas. In general, the acquired governmental dataset and the OSM building labels are accurate and up-to-date. In some cases, the imagery depicts large construction sites where the more recent OSM data already labels the newly constructed buildings (Figure 3). Here, the reference data is accordingly adjusted to the imagery and the building labels are deleted in the reference dataset.

Figure 3: Discrepancy between the ground reference data and the satellite imagery. Vienna target area.

A more detailed examination reveals further discrepancies between the reference labels and the imagery. Hence, it is a necessary step to zoom in and go through the entire study area to adjust the reference labels accordingly to the imagery. Maggioiri et al. (2017) for example also used inaccurate OSM labels as a starting point and manually refined it to derive a building footprint reference dataset. The manual amendment of an existing dataset is, in general, significantly less time consuming than

to generate the reference building labels from the scratch (Li et al., 2022; Usmani et al., 2023).

ESRI ArcGIS software was applied in this research to edit and correct the acquired OSM building datasets. Having the imagery as background and the reference labels on the top, each building polygon can be moved to its correct place to match the corresponding building on the imagery. Significant spatial noise of several meters can occur because the satellite images used by the OSM contributors for digitalizing the building footprints were not correctly registered or the off-nadir angle differs between the imagery upon which the reference labels were edited and the actual imagery that was selected for this study (Usmani et al., 2023, Rastogi et al., 2020). Figure 4 depicts the spatial discrepancy between actual imagery and the OSM labels used as reference data.

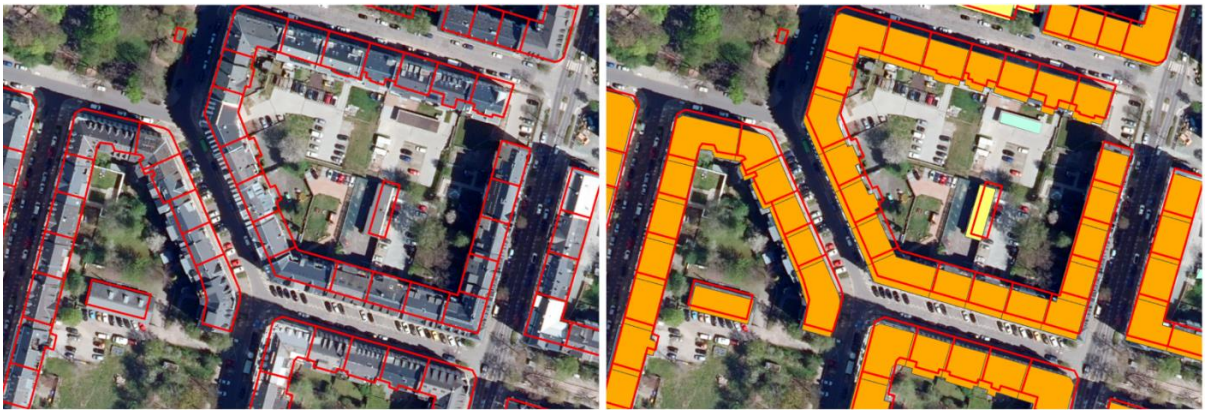

Figure 4: Manual correction of the discrepancy between the ground reference data and the satellite image. The corrected and utilized ground reference is depicted on the right in orange.

The spatial adjustment of the reference building labels is relatively fast in comparison to the completion of the reference dataset with missing building polygons. Larger buildings like apartment blocks, administration, commercial and industrial buildings are in all utilised reference datasets almost complete. However, numerous smaller buildings like detached houses and especially garages are missing in the OSM building data as well as in the Estonian governmental data. Thus, missing buildings in the reference data were manually edited with the ArcGIS software in this research.

At this stage, it has to be exactly defined, what objects are actually considered as buildings. For large commercial buildings, factories, apartment blocks and single family houses, the answer is obvious. However, for smaller objects it is more complicated. Are small garden sheds or bus stop shelters buildings? Are awnings part of the house? In the OSM building label datasets, significant differences exist. This is not surprising as the OSM database is edited by different contributors. Although the OSM project gives some guidance for the editing process, the decision if a garden shed counts as a building or not, is often an individual decision of the editors (Biljecki et al., 2023).

It is in general agreed upon in the deep learning literature, that false or missing building reference labels can have a significant negative effect on model training and model performance (Huang et al., 2019). If features with typical building characteristics on the training imagery are not labelled as buildings, the chance is high that the model gets confused and learns false representations. The same applies for missing reference labels for small objects like sheds or garages (Luo et al., 2023). Missing reference labels can not only negatively impact the model training but are also a problem for the performance evaluation of the models. When a model correctly predicts a building in the imagery but the corresponding reference label is missing, the statistical measurement counts an additional error. If a building extraction model recognizes small objects as buildings because of high similarity in features like texture, colour, edges etc. but the reference data doesn't contain objects smaller in size than a certain threshold, a huge number of false positives is produced.

The delineation of adjacent buildings in the reference data poses an additional challenge. In the OSM reference data, terraced houses and garage blocks are in some cases separated individually, whereas in other cases labelled as one large building block. To a lesser degree, it occurs also that apartment blocks are separated into multiple buildings. Most likely, the editors had local knowledge and used the multiple entrance doors of the building as delineation parameter. A special case is the typical superblock architecture in Barcelona where delineation of individual buildings is nearly impossible. Figure 5 depicts some examples of such buildings with the according reference data.

Figure 5: Ground reference datasets. Separate labelling of every individual garage in the Kosice ground reference data (in pink). [Top]. Large number of adjacent apartment buildings within the Catalonian block apartment structures in Barcelona (in orange). [Bottom].

The research literature does not provide recommendations and best practice on the appropriate utilization of reference label datasets. (Stiller et al., 2019). This research handles the above mentioned challenges with the utilized reference datasets as follows: Garages, larger garden sheds, bus stop shelters and similar are regarded as buildings and complemented in the existing reference dataset as long as they are visually identifiable. Smaller objects like doghouses, tiny sheds and storage boxes are not considered. However, uncertainties remain. Awnings and sun blinds are not considered as building parts and are not added to the reference data but when those objects are already incorporated in the reference building labels as part of the buildings, a subsequent manual deletion is not conducted. Particularly problematic are greenhouses, constructed from plastic sheets. Some study areas encompass also suburban areas where such objects are very common. This study does not count them as buildings and in the respective OSM reference data such greenhouses are missing, too. The separation of terraced house rows and apartment blocks in the reference data, remains unchanged. In the Kosice ground reference dataset, each garage is labelled separately as a building within large garage blocks.

Lastly, a classification of the buildings is conducted to differentiate between different building types. This is a prerequisite step to answer research question R3. The buildings are categorized as “single family house”, “terraced houses”, “apartment

buildings” , “apartment blocks”. Large industrial commercial and public institution buildings are summarized into the “ICP buildings” category. The Figures 6-8 provide visual examples of each building-type category. Building type classification was conducted through visual inspection of the utilized imagery as the building reference label datasets does not provide a constant classification.

Figure 6: Building type categorization. Single house family buildings (dark blue), terraced houses (light blue), garage blocks (pink). [From left to right].

Figure 7: Building type categorization. Apartment buildings (orange) have various roof types, apartment blocks (red) are usually separated and have a flat rooftop. Historical districts (brown) can be found in the Tallinn and Kosice target areas. [From left to right].

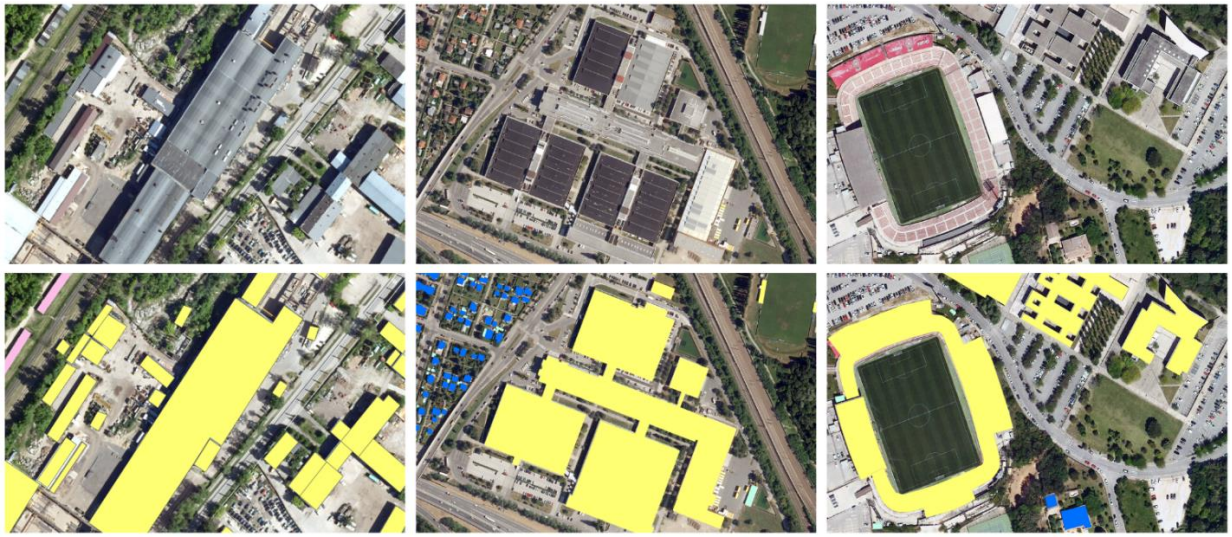

Figure 8: Building type categorization. Industrial buildings (left), commercial buildings (middle) and other public buildings like stadiums or sport halls (right).

3.2.3 TRAINING DATA

Training data is utilized in this research to fine-tune already existing, pre-trained building extraction models. A common approach in the deep learning building extraction domain is to divide the acquired imagery and the respective reference data randomly between training and test samples. Usually, 70% or 80% are used for model training, and the remaining samples to test the model performance (Li et al., 2019; Nepuane et al., 2021; Wang et al., 2022). Other researchers consider this approach problematic, especially in case of testing model generalizability and geographic transferability (Maxwell et al., 2021). This research therefore ensures the spatial separation between training and target areas to examine model performance on new, previously unseen, target areas.

Numerous studies on building extraction model fine-tuning point out the relevance of the training data characteristics to achieve satisfactory results on new target areas. Of particular importance is a high diversity of the training area in regards of building types and urban structures, and a certain degree of resemblance between the training and new target area (Kucharczyk et al., 2020; Nex et al., 2019).

The first training area of this study is located in eastern Tallinn (Estonia). It is characterized by a large variation of building types and urban structures, including a vast industrial and commercial zone, a part of the historical oldtown, port facilities, high-rise apartment blocks and multiple districts with predominately detached houses. The second training area covers a part of southern Vienna (Austria). The area features a large variance of building types like commercial zones, various types of low- and high-rise apartment blocks and areas with detached houses. Table 1 provides the number of building footprints and image batches for both training areas.

Training Area	Size km²	Image Chips	Buildings	Building Type					
				SF_House	Apt. Bldg.	Apt. Block	ICP Bldg.	Dwelling	Other
Tallinn	76.9	4384	12091	3589	1107	1387	2579	2389	1040
Vienna	10.7	8153	10023	3583	627	950	760	3148	955

Table 1: Training areas for model finetuning. Single family house ("SF_House"), apartment buildings ("Apt. Bldg."), apartment blocks ("Apt. Block"), industrial, commercial and public buildings ("ICP Bldg."), dwellings. Garages, historical buildings, terraced houses are summarized as "Other".

Data augmentation is an additional, frequently utilized technique, to further increase the training data size and diversity. It can potentially increase model generalizability and can train the model to better deal with object variety, facilitating better results on new, unseen target areas (Liu et al., 2020; Rastogi et al., 2020). Although the exact positive impact of various data argumentation techniques on model performance is still an open research question, flipping, vertical and horizontal shifting, rotation of 90 degrees, blurring, increasing noise are utilized very often to increase the training data. (Abriha et al., 2023; Ghorbanzadeh et al., 2019; Han et al., 2022; Ye et al., 2019). For this research, ArcGIS Pro 3.1 "Export training data for deep learning" tool was used to generate training data for model fine-tuning. Stride and rotation angle techniques were utilized to augment the training datasets.

3.3. MASK R-CNN-BASED BUILDING EXTRACTION

3.3.1 INSTANCE SEGMENTATION

For the extraction of buildings, a model for semantic segmentation or instance segmentation is appropriate. The main difference in those two processes is that the former differentiates between all buildings on one side and the background on the other side meanwhile the latter extract each building individually (Liu et al., 2021). This study considers instance segmentation as well suited for the building extraction task.

Recently, two very promising instance segmentation models were released: Segment Anything Model (SAM) from Meta in 2023 and YOLOv9seg in 2024 (Kirilow et al., 2023, Zhou et al., 2024). However, an extensive examination of those models on the remote sensing building extraction task is yet missing. Thus, this study explores the well-established and explored Mask R-CNN instance segmentation model architecture, which was introduced in 2017 and since them frequently used for building footprint extraction tasks (Sakeena et al., 2023; Tiede et al., 2021).

3.3.2 MASK R-CNN MODEL

Mask R-CNN is an extension of Faster R-CNN, a region based network for the detection and classification of multiple objects of interest in images (Tiede et al., 2019). First, Girschick et al. introduced region based object detection networks in 2013 (R-CNN) and Fast R-CNN in 2015, achieving a significant improvement in the visual object detection domain (Hoeser & Kuenzer, 2020; Wu et al., 2020). Ren et al. increased with Faster R-CNN the speed and accuracy of the classification and object detection task showing state-of-the-art results in 2015 (Ma et al., 2020; Zhan et al., 2022). He et al. proposed Mask R-CNN in 2017, a network for object detection and instance segmentation, outperforming previous state-of-the-art models on many instance segmentation challenges (Minaee et al., 2022). In fact, Mask R-CNN is a Faster R-CNN with three output branches. The first generates bounding boxes localizing objects of interests, the second proposes the respective classes of those objects and the third provides a segmentation mask for each detected object of interest. Thus, Mask R-CNN is able to accurately predict the location information of the target object and, additionally, extract the target objects shape as a mask, providing fine-grained pixel level boundaries for a detailed instance segmentation (Han et al., 2022; Liu et al., 2021).

A major enhancement of the Mask R-CNN network is the utilization of a Feature Pyramid Network (FPN) that provides a multi-scale feature representation incorporating features from different scales which allows the model to gain a more comprehensive understanding of object context and a better object segmentation across a wide range of different object sizes. High-resolution features provide rich semantic information, meanwhile low-resolution features add more precise spatial details. The use of a Region of Interest Align (ROI Align) is another significant improvement of the Mask R-CNN in comparison to previous region based networks. It ensures accurate spatial information capture from the input feature map for each region of interest proposal achieving an improved pixel-wise segmentation accuracy, particularly for small objects. During the training, the model is optimized using classification loss, bounding box regression loss and mask segmentation loss. That allows the model to learn simultaneously to detect objects of interest, refine the bounding boxes and generate precise segmentation masks.

Several limitations of Mask R-CNN models have to be considered as well. Such models are complex and the training and application on high resolution images is computationally intensive. Additionally, several studies report of performance shortcomings especially by the localization and segmentation of large buildings. Buildings with complex roof structures like industrial areas or densely populated build up regions, where buildings are close to each other, pose significant challenges for

Mask R-CNN models (Sakeena et al., 2023; Zhan et al., 2022). Furthermore, instance segmentation often fails on adjacent buildings (Spasov & Petrova Antonova, 2021).

3.4 TRANSFER LEARNING

Transferability is described by Cheng (2020) as a foundational ability of human learning. Humans can gain relevant knowledge from other related problems and apply it to solve new problems with only a few samples. The generalization theory of transfer postulates that learning to transfer is the result of the generalization of experience. The prerequisite of successful transfer is a connection between two learning activities (Zhuang et al., 2021). Inspired by the capability of humans to transfer knowledge across different domains, transfer learning in the deep learning domain aims to leverage knowledge from a related domain (i.e. source domain) to improve the learning performance or reduce the number of required training samples in a target domain (Nepuane et al., 2021; Zhuang et al., 2021). Lin et al. (2022) describe transfer learning as a method of transferring the knowledge trained or learned from deep learning models to a new pending task. The overall aim of transfer learning is to improve the performance on the target domain by transferring the knowledge from different but related source domains (Abriha et al., 2023, El Asri et al., 2023).

Transfer Learning has lately become increasingly popular in different deep learning based earth observation applications (Abdi & Jabari, 2021). Training a deep learning model from the scratch is challenging. It requires sufficient data – and sufficient in this context often means millions of correctly labelled training samples – and it takes a long time and enormous computational resources to reach convergence (Luo et al., 2021; Panboonyuen et al.; 2019; Shao et al., 2020; Yang et al., 2021). For the building extraction task, transfer learning can be used to improve model accuracy on target locations where labelled data is scare with the utilization of training on geographic regions where enough training data is available (Li et al., 2022).

For successful transfer learning, two critical points have to be considered. First, the difference between the source- and the target domain characteristics. If there is little in common between the two domains, the chance is high that the transfer of knowledge will fail (Wurm et al., 2019; Zhuang et al., 2021). The other important decision in the transfer learning process is about the application of a pre-trained model. Such a model can be used as a feature extraction tool in which the weights of the convolutional basis of the network are frozen (Yang et al., 2021). Here, the aim is to train a model as generic as possible for a given task like building extraction. If the model is trained with a large enough dataset covering various building types and different geographic areas, the model will show, theoretically, good performance even on unseen target areas

(Bouchard et al., 2022; Prakash et al., 2022). Another approach is to fine-tune the pretrained model with a limited number of training data from the target source (Yang et al., 2021). Obviously, the number of training samples should be as low as possible to keep the effort small but high enough to ensure a performance improvement of the model. Valentijn et al. (2020) reported from two studies where 10% and 15 – 25 % of the available building labels of the target domain were used for the fine-tuning of the pre-trained model. In both studies, the finetuning improved the results significantly. Regarding the characteristic of the training samples, there is a general consensus that it should be diverse and representative for the target domain (Gao et al., 2022; Gella et al., 2023).

The fine-tuning of a pre-trained model is a critical process. In psychology, the phenomenon, when new information causes someone to forget old information, is referred as retroactive interference. This often occurs during language learning, especially when the new language is similar to the already known (Zhuang et al., 2021). In transfer learning, catastrophic forgetting can occur when a model is trained on multiple tasks and the weights that are important for task A are changed to meet the objectives of task B (Kirkpatrick et al., 2017). For the building extraction task, the performance and generalizability of a pre-trained model can increase or decrease, depending on the fine-tuning of its weights (Gao et al., 2022; Valentijn et al., 2020). Qiu et al. (2022) for example, trained different building extraction models on the DeepGlobe Las Vegas dataset. In one experiment, the model was directly applied on the Shanghai and Paris datasets. In the next experiment, the pre-trained model was fine-tuned on the Shanghai and Paris datasets before testing them on those target areas. The authors report an increase in performance after fine-tuning the pre-trained model. The IoU score on the Shanghai dataset for example increased from 0.38 without fine-tuning to 0.65 after the fine-tuning. Other studies found that fine-tuning of a pre-trained model did not generate a significant difference in the model´s performance at all (Bouchard et al., 2022). Li et al. (2022) even reports from a decrease in the accuracy metrics after fine-tuning a building extraction model.

Although model fine-tuning is considered as a great possibility to increase model performance on new, previously unseen, target areas, two limitations have to be considered. First, there is no perfect recipe how much additional training data is required to achieve a significant improvement. It seems, that the impact of fine-tuning depends on the respective pre-trained model, the characteristic (e.g. diversity) of the new training data and the new target area. Qiu et al. (2022) claim that fine-tuning on even 10% of the new target dataset results in performance improvement.

It is in general recommended, that the higher layers get fine-tuned while the underlying layers focusing on general features, remain unchanged (Luo et al., 2021; Shao et al., 2020; Yang et al., 2021).

To test the impact of fine-tuning on model performance, this study performs fine-tuning with two training datasets on two different pre-trained building extraction Mask R-CNN models. Those fine-tuned models are then tested on different target areas and the performance is compared with the results of the pre-trained models without fine-tuning.

3.5 MODEL PERFORMANCE EVALUATION METHODS

Research on deep learning based object detection and extraction models often focus on the testing of different models on one common task in order to find the best performing model. From the user´s perspective, it is more important to estimate how well a model will perform on future tasks putting realistic use cases into the focus. Here, the generalizability and transferability of the model plays a crucial role. Hence, for the building extraction task, model performance should be elaborately tested on different target areas and test scenarios (Li et al., 2021). To assess the performance and interpret the results, it is advisable to use standardized and expressive measurement metrics (Bai et al., 2022).

This research uses the confusion matrix which is regularly applied in the deep learning domain, as foundation for further performance analysis. It has four elements: True positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) (Bai et al., 2022). For the building extraction task, true positives are correctly predicted buildings, false positives are objects falsely predicted as buildings and false negatives are buildings not recognized by the model (Ma et al., 2020). The category of true negatives, that means the correct recognition of non-buildings, is not recorded in this research in accordance with other studies on object oriented building extraction (Maxwell et al., 2021). To determine if a predicted building footprint is counted as a true positive, the intersection over union framework is applied. It calculates the area of overlap between the predicted building footprint (i.e. polygon) and the corresponding building polygons in the reference dataset. The higher the ratio, the better the model´s accuracy. However, a 100% overlap is not realistic. Han et al. (2022) for example consider an intersection ratio of 0.6 (i.e. 60%) as true positives. Numerous other studies use an overlap threshold of 0.5 (i.e. 50%) to accept the prediction as a true positive (Chen et al., 2022; Sawa et al., 2024). The 0.5 threshold is also applied for the benchmarking building extraction DeepGlobe challenge and the PASCAL VOC object detection competition (Arya et al., 2021, Pi et al., 2019). This study follows the mainstream and defines a spatial overlap of at least 50% between predicted and reference buildings footprints, as a true positive.

For an adequate assessment of the model performance, it is important to note that the presented research executes an object-based evaluation method. This approach

emphasises not only the importance of accurate detection of building areas, but also the complete identification of building footprints (Li et al., 2019). Mask R-CNN models tend to split up building predictions in multiple polygons, especially in case of large or complex buildings (Li et al., 2021). For the model evaluation, those predictions are, however, all false positives (i.e. errors) because none of the predictions covers more than 50% of the reference building itself. In practice, this could lead not only to one false negative but also a large number of false positives, namely the predictions covering small parts of the target building (Figure 9).

Figure 9: Building extraction on large objects and complex roof structures. Mask R-CNN models often generate multiple predictions for the very same object (in orange).

After determining the exact number of TP, FP and FN for every model and on each target area, multiple classification accuracy assessment metrics are utilized to interpret the results. In the deep learning domain, the precision, recall, F1 score and Intersection over Union (IoU) ratio metrics are regularly applied (Abdollahi et al., 2020; Li et al., 2024). This research applies those well-established measurement metrics to examine the model performance.

Precision

Precision measures the percentage of correct positive predictions among all predictions. It answers the question how many of the predicted buildings were actually buildings. A high precision score indicates that the model produces few false positives (El Asri et al., 2023, Panboonyuen et al., 2019).

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall

Recall is the ratio of true positives to the total number of positive samples. It answers the question how many of the reference buildings were actually predicted as buildings. A high recall score means that the model can detect most of the actual buildings. (Chen et al., 2021, Tiede et al., 2021)

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

F1 score

The F1 score is a harmonious average of recall and precision providing a balance between precision and recall. It is often used to measure the overall model performance with values ranging between 0 on the lower end and 1 representing a faultless model performance (Li et al., 2019; Shao et al., 2020).

$$F1\ score = \frac{2\ x\ Precision\ x\ Recall}{\mathrm{Precision} + \mathrm{Recall}}$$

IoU:

In addition to the F1 score, the Intersection over Union (IoU) ratio is also a widely used performance measurement metric in the object detection and semantic segmentation domain (Bai et al., 2022; Prakash et al., 2022). The IoU ratio is the intersection of the reference and predicted samples with the union of the two groups. The scores range, likewise to the F1 score, between 0 and 1. Numerous studies on building detection and extraction utilise both, the F1 score and the IoU ratio to measure model performance (Abdollahi et al., 2020; Li et al. 2021). This study follows that approach by calculating the F1 score and the IoU ratio to receive a robust model performance examination and to ease the comparison with other studies on building extraction.

$$IoU = \frac{\mathrm{TP}}{TP + FP + FN}$$

3.5 RESEARCH SETUP

In order to examine and validate the research questions and hypothesises, a multi-stage research setup is implemented. In a first step, two available pre-trained Mask R-CNN building extraction models are evaluated on ten, previously unseen, target areas. In a following experiment, model fine-tuning is conducted on those two pre-trained models. Building extraction is then executed on the established target areas with three different fine-tuned models to assess the geographical transferability of those models and to compare the results with the performance of the pre-trained, unmodified models.

The outlined research setup follows in general numerous other studies on geographic transferability and cross-regional segmentation model performance evaluation (Nex et al., 2019; Stiller et al., 2019).

3.5.1 PRETRAINED MASK R-CNN MODEL

The ESRI analytics team provides on the ArcGIS Living Atlas website (http://livingatlas.arcgis.com) numerous pre-trained deep learning models for various tasks such as car detection, pool detection, land cover classification and building footprint extraction. For the building footprint extraction task, in total six Mask R-CNN models are available as of the time of this study, each focusing on different geographic areas such as the USA, Africa, Australia, New Zeeland, China and Saudi Arabia (KSA). All models can be downloaded in the ESRI deep learning package format. The download is free of charge and open for usage with the ArcGIS software. ESRI provides for each model a short description, containing the date of publication, recommended input imagery characteristics, model accuracy and the recommended geographic region for model application. Two pre-trained models were selected for this research: a model trained on training areas located in the United States and another model with the same architecture, trained on training areas located in Africa. The two examined models are called in this study “PT USA” and “PT AFR”.

One drawback of the utilization of the ESRI building extraction models is that an elaborate model description is missing. The size and characteristic of the training datasets are not provided. A specific question addressed to the ESRI analytics team in December 2022 regarding the geographic characteristics of PT USA´s training data resulted in a short answer that the model was trained on all type of geographies including suburban, urban, rural and industrial areas in the United States. The current ESRI policy of not providing details on the training of the released models was also experienced by Sawa et al. (2024). Furthermore, ESRI does not publish an elaborate model performance description except a short information about the alleged model performance. The US model is expected to work well on target areas covering the United States reaching an average precision of 0.71. ESRI expects for the Africa building extraction model the best results in Uganda and Tanzania with an average precision score of 0.78.

3.5.2 MODEL FINE-TUNING

Model fine-tuning is often conducted on a small part (e.g. 10 – 25%) of the new target dataset (Nex et al., 2019: Qiu et al., 2022). For this research, however, the geographical limits of model fine-tuning are of particular interest. Hence, fine-tuning is

not conducted on a portion of the target areas but on spatially separated training areas. In accordance with research hypotheses H 1.1. and H 1.2, the positive impact of model fine-tuning should be still significant on target areas within high geographical proximity and closeness to the selected training areas. The performance of the fine-tuned models should decrease with decreasing geographical proximity and closeness between the fine-tuning training areas and the target areas.

Fine-tuning of PT USA is executed on two different training areas in Tallinn and Vienna, receiving two different fine-tuned versions of the pre-trained model: “FT USA TLN” and “FT USA VNA”. PT AFR is fine-tuned only on the Vienna training dataset to save processing time, resulting in the fine-tuned model version “FT AFR VNA”.

The utilized training area images comply with the input requirements of the two pre-trained models: RGB, 8 band high resolution 10-40 cm images. The Tallinn training area image has a resolution of 25 cm, the Vienna image 20 cm. The ground reference labels of the two training areas were double-checked and manually amended to avoid confusion during the model training phase. The training samples (i.e. image chips) were created with the ArcGIS Pro “Export Training Data for Deep Learning” tool. Data augmentation (i.e. stride and rotation) were executed for both training areas. A larger stride and rotation (i.e. 90° instead of 180°) was conducted on the Vienna training samples to outbalance the larger geographical size and higher building diversity of the Tallinn target area. Table 1 provides additional details for the two training areas.

ArcGIS Pro “Train Deep Learning Model” was used to fine-tune PT USA and PT AFR. 10% of the training samples were used for validation during the training. The “Freeze Model” option ensured that only the final layer of the model is impacted by the new training samples, while its core layers remain unchanged. This setting avoids the risk that the model “unlearns” its core knowledge (i.e. catastrophic forgetting). Model training was stopped when the model did not improve further. The fine-tuning of PT USA on the Vienna training area (i.e. FT USA VNA) stopped after ten epochs with an average precision score of 0.57. The fine-tuning on the Tallinn training area reached an average precision score of 0.70 after 18 epochs. The fine-tuning of PT AFR on the Vienna target area stopped likewise after 18 epochs with an average precision score of 0.62.

The ESRI analytics team reports for the two Mask R-CNN building extraction models, that are utilized in this research as PT USA and PT AFR, an average precision score of 0.71 and 0.78. Thus, the obtained fine-tuning results, particularly on the Vienna training area, have to be considered as relatively low. However, the research goal is not to achieve peak results with the fine-tuned models but to examine the model´s cross-regional performance variability.

3.5.3 TARGET AREAS

In order to examine the performance of the two pre-trained and three fine-tuned building extraction models on different geographic locations, and to answer the research questions 1.1 and 1.2, the target areas are distributed among different cites, countries, geographic regions and continents (see Appendix 2). The target areas cover predominantly urban areas and partially some suburban or even rural structures. All of them entail various building types like single family houses, apartment blocks, industrial facilities, commercial complexes and large public buildings.

The spatial extent, as well as the number and type of buildings are depicted in Table 2. In the following section, a short description of the test areas is provided. Appendix 3 provides image samples of each target area.

This study uses in total 83.510 reference building footprints within 10 target areas for the examination of building extraction models´ cross-regional transferability. In comparison, Chen et al. (2022) used 93.000 building footprints and Bakiman et al. (2022) trained and tested their model on around 40.000 annotated building footprints in Istanbul.

Target Area	Size in km²	Nr. Bldg.	Building Type							
			SF_House	Apt. Bldg.	Apt. Block	ICP Bldg.	Ter.House	Dwelling	Garage	Other
Tallinn	69.5	13.993	4172	2126	1513	1978	320	2782	952	150
Narva	13.5	3094	584	332	251	461	0	1195	267	4
Daugavpils	16.7	9572	2270	578	480	1495	1	4466	257	25
Kosice	11.4	11.532	1842	228	1335	788	762	1912	4322	343
Chemnitz	11.7	7396	214	3118	632	1372	25	1604	421	10
Vienna	10.2	9739	3385	577	732	718	860	3403	50	14
Girona	8.1	5663	1116	2128	211	491	914	777	19	7
Barcelona	5.6	7371	0	6493	82	444	19	332	0	1
New York I	8.3	3604	1541	7	671	398	505	478	2	2
New York II	6.9	11.546	4082	0	341	631	4001	2487	2	2

Table 2: Target areas to test cross-regional model performance.

Two test areas cover parts of the Estonian cities Tallinn and Narva. Tallinn, the capital of Estonia, is generally considered as a modern European city with a population of around 450.000 inhabitants. It has a considerably large medieval old town, numerous districts with predominantly single family houses and large, soviet-era, apartment blocks. Several industrial complexes and port facilities contribute to the heterogeneity of the Tallinn target area. Narva, less than 200 kilometres east of Tallinn, can be described as a typical Eastern European small city with slightly above 50.000 inhabitants. The town consists mainly of apartment blocks and several areas with a

rather rural characteristic. The third test area covers the Latvian town of Daugavpils. It is a former Soviet garrison town with around 80.000 inhabitants. Daugavpils is characterized, similarly to Narva, by soviet-era apartment blocks and rather rural areas in close proximity to each other. Several, large garage blocks, especially in the north of the test area, are a distinctive feature of this test area.

The following three target areas are located in different Central-European states. Kosice, a town of 230.000 citizens in eastern Slovakia, shows also a mixture of communist-era building blocks, long rows of terraced houses, districts with predominantly single family houses and a small historical oldtown. Chemnitz, located in eastern Germany, has around 250.000 citizens. The city suffered great damage during the Second World War and large areas were reconstructed after the war with predominantly low-rise apartment buildings. The third Central-European target area covers a section of northern Vienna, the capital of Austria. The city has a population of over 2 million and is considered a major metropolis in Central Europe. The selected area contains a great variance of different urban structures like multiple types of high-rise apartment blocks, areas with exclusively detached houses, large industrial and public buildings.

Two target areas are located in Spain. They cover parts of Barcelona and Girona. The spatial distance between the two target areas is less than 100 kilometres. Girona, a medium-sized town with around 100.000 inhabitants, is located in north-eastern Spain. The target area covers mainly apartment buildings, a large commercial complex in the south and some detached houses with associated pools and yards. Barcelona, on the other hand, is a metropolis with more than 1.5 million citizens. The city is characterized by a unique architectural style composed of tessellated apartment blocks. The test area is covered largely by those apartment blocks and a few commercial buildings.

Lastly, two target areas cover two separate parts of the Bronx, New York City. Both areas have a similar urban structure with densely built-up district blocks largely consisting of low-rise apartment houses and closely built up detached houses. Some high-rise apartment buildings and large public and commercial buildings complement the test areas.

For the examination of hypothesis 1.2, an unambiguous demarcation of geographical regions is required to analyse the impact of geographical closeness between target and training areas on model performance. The highest geographical closeness is assumed for training and target areas that are located within the same city. This is the case for the Tallinn and Vienna target and training areas. The second highest geographical closeness exists for locations within the same country. This category applies for the Tallinn training and Narva target area. The third level of geographic closeness applies for locations within the same geographic region, like the

Daugavpils target area and the Tallinn training area are both located in the Baltic Region. The target areas in Kosice and Chemnitz are considered geographically closer to the Tallinn than the Vienna training area. More than 40 years of communist city planning and development in Slovakia and Eastern Germany with the corresponding results in urban structures and characteristic building features like high-rise and low-rise panel construction apartment blocks are considered more relevant for the automated building extraction process than the spatial distance between those cities and Vienna. The Mediterranean region represented by the test areas in Girona and Barcelona, is considered to have low geographical closeness to the Central- and Eastern European training areas. Finally, the selected European urban areas are considered geographically closer to the United States than to Africa. Consequently, it is expected in accordance with H 1.2, that PT USA outperforms PT AFR on the selected target areas.

3.5.4 MODEL PERFORMANCE SUB-REGION TEST AREAS

To examine model performance variations within a target area and to examine the impact of specific building types and urban structures on model performance, the target areas are further subdivided into multiple, more homogeneous sub-regions (i.e. micro regions). Sakeena et. al. (2023) likewise conducted model testing on subsections of the test areas. In their research, every image tile represents a sub-region. Those sub-regions are classified as predominantly single family buildings, multiple family buildings and industrial buildings, manually by a domain expert judging the predominant type of building in each tile.

This study subdivides with a hexagon tessellation split each target area into 0.5 km² large grid cells. After that, each of the hexagon grid cells is classified according to its geographical characteristic. The class attribution of each micro-region follows the largest number of a certain building type as well as the total size in square meters occupied by the specific building types. This approach enhances the reproducibility of the research findings on other target areas and avoids the potential arbitrariness of class determination through individual domain experts.

Obviously, the size of the tessellation grid cells has an influence on its characteristic. The chosen size of 0.5 km² is a compromise. A smaller ground area would lead to more homogeneity but the processing time would increase and at a certain level the explanatory value for a general assessment would diminish. On the other hand, if the grid cells are too large, significant heterogeneity of the geographical characteristics of the examined micro-regions will likely occur. In total, 23 sub-regions are utilized to examine sub-regional building extraction performance. The Figures 9 and 10 provide some samples of the micro-regions within the Tallinn target area.

Figure 9: Subdivision of the Tallinn target area with the tessellation grid (green). The reference building labels of the selected micro-regions are displayed.

Figure 10: Micro-regions within the Tallinn target area. Large building blocks (left), a commercial and industrial area (middle) and a residential area with predominantly single family houses (right)

3.5.5 MASK R-CNN MODEL INFERENCE

ESRI Inc. provides in its software environment a tremendously simple access to deep learning models. However, certain technical preconditions have to be ensured to utilize the provided AI resources. Minimal recommended computational requirements for pre-trained model inference is a GPU with 4 GB dedicated memory. For model training (e.g. fine-tuning), 8 GB are strongly recommended. This research was conducted with a NVIDIA RTX 2060 MaxQ GPU with 6 GB dedicated memory. The fine-tuning of the pretrained models took more than 5 hours each. The inference of the pre-trained and fine-tuned models was, depending on the size of the target area, significantly shorter with around one to two hours.

Other studies on deep learning building extraction models rely usually on higher hardware resources. Deep learning model training and processing is in general conducted on various NVIDIA GPUs. Li et al. (2021) and Bakirman et al. (2022) use similarly to this research the NVIDIA RTX 2080 series, however with 11 GB memory. Other research was conducted on the more advanced RTX 3090 or NVIDIA Quatro series with 24 GB (Abriha et al., 2023) or even on highly professional RTX A6000 GPU with 48 GB memory (Luo et al., 2023). A few studies utilized significantly weaker GPUs with only 6 GB. To compensate for the relatively low processing power, the batch size during model inference was reduced to 4 (Chen et al., 2021, Shao et al., 2020). According to Bakiman et al. (2022), the batch size does not impact the model performance. However, it increases the processing time significantly. Hence, the execution of deep learning models is possible with semi-professional hardware but it reaches relatively quickly its limits due to the increase in processing time.

In regards of the hyperparameter settings for model inference and model fine-tuning, the recommendations of the model provider (i.e. ESRI analytics team) are largely followed. For the inference of the two pre-trained and three fine-tuned modes, the tile size was set to 512, batch size to 4 and the prediction confidence threshold to 0.5. Non-max suppression was used to reduce overlapping predicted polygons in accordance with the research of Chen et al. (2022) and Sakeena et al. (2023)

3.6 RESEARCH EXPERIMENTS

The presented research experiments facilitate the examination of the model´s geographic transferability and performance on new, previously unseen target areas throughout various geographic locations. In this section, the experimental setup, the experimental goals and the expected outcome are described in detail. The section is subdivided in accordance with the three research questions.

3.6.1 RESEARCH QUESTION R 1

Research question R 1 examines the impact of geographical proximity and closeness between the training and target areas on the performance of Mask R-CNN building extraction models. In a first iteration, PT USA and PT AFR are executed on the selected ten target areas. The spatial distance between the European target areas and the training areas located in the United States and Africa, is equally large. Thus, low performance is expected for both models in accordance with H 1.1 It is expected in accordance with H 1.1 and H 1.2 that PT USA performs well on the two New York target areas. Acceptable results of PT USA are expected also for the Central- and Eastern European locations due to a relatively high geographical closeness between those target areas and the US model´s training sites. Poor results are expected for the two Mediterranean target areas due to a lower geographical closeness. It is expected that the PT AFR underperforms PT USA on all target areas.

In the next experiment, the fine-tuned model variants are compared with the performance of the two pre-trained models. For FT USA TLN, FT USA VNA and FT AFR VNA, a significant performance increase is expected in accordance with H 1.1 on the Tallinn, respectively Vienna target areas. Good performance is expected for the nearby target areas like Narva and Daugavpils for US-FT TLN and on Bratislava and Chemnitz for US-FT VNA

According to H 1.1, the positive effect of model finetuning on the performance diminishes with growing spatial distance from the training area. However, model performance improvement is expected on target locations further away from the Tallinn and Vienna training areas where the geographical closeness between the target and training areas is high. For FT USA TLN, relevant performance improvement is expected on the Narva and Daugavpils target areas, and some improvement on the Chemnitz and Bratislava target areas. For the Vienna target area, a lower performance increase is anticipated due to the lower geographical closeness. No improvements are expected on the Girona and Barcelona target areas. Similarly, for FT USA VNA, the highest performance increase is expected for the Vienna target area followed by the Chemnitz and Bratislava target areas. Minor performance improvement is expected for the Baltic Region target areas and no improvement for the target areas in Spain and the USA.

The fine-tuning of PT AFR on the Vienna training area should result in significant performance improvement, especially on the Central European target areas. Relevant performance improvement on the remaining European and US target areas is expected in accordance with H 1.2. No improvement is anticipated for the two Mediterranean target areas.

3.6.2 RESEARCH QUESTION R 2

In this research, three different post-processing techniques are executed on the predictions of the five examined models. The first iteration is without any post-processing. These results are regarded as standard upon which possible improvements through post-processing methods are measured. The first post-processing technique is the utilization of the ArcGIS “Dissolve Boundaries” tool in order to merge adjacent polygons. Mask R-CNN building extraction tends to produce multiple polygons for the same building, especially for large buildings or in case of complex roof structures. The “Dissolve Boundaries” tool can also improve the extraction results through the altering of the reference dataset building footprint polygons. The manual digitalization of the reference datasets often results in over-segmented building footprints. Capturing adjacent awnings, sun blinds or oriels as separate (building) objects results in a significant distortion of the model performance measurement. The same applies for apartment blocks that are subdivided in multiple parts, correspondingly to the respective entrances. Another example are large garage blocks where each individual garage is separately labelled as an individual building. Significant differences exist not only between the reference datasets of different cities but also within the very same settlements. The usage of the “Dissolve Boundaries” function ensures equal standards throughout the reference datasets. It is expected, in accordance with H 2, that the execution of the Dissolve Boundaries tool is particularly beneficial on areas with large industrial and commercial buildings and apartment blocks as well as on locations with large garage blocks.

The next examined post-processing round focusses on the building size. This research defines a minimum building size of 25 m² to exclude non-building in the prediction and the reference datasets. Small single family houses and garages remain meanwhile numerous false positives like small garden sheds, parasols, and cars are deleted. The 25 m² threshold post-processing is conducted after the “Dissolve Boundaries” step. This sequence avoids perforated building footprints by merging adjacent prediction polygons first and deleting afterwards the remaining ones below the defined minimal size. In general, it is expected that the erasure of small objects leads to a general improvement of the extraction results. However, target areas covered largely by single family houses and backyards containing numerous small objects like sheds as well as large industrial zones with complex building structures, will likely reveal the most significant improvements.

The increase of the model prediction confidence threshold is the third examined post-processing method. Deep learning models work with variable certainties. The Mask R-CNN model defines for each recognized object of interest (i.e. building) a prediction confidence. Generally, deep learning models never reach a 100% certainty

in classifying objects. Thus, a probability of around 90% is considered highly likely that the model prediction is right (i.e. TP). Values below 50% imply a high risk of incorrect predictions. In general, the choice of a low probability threshold tends to generate more false positives, that means objects are wrongly classified as buildings. A high threshold on the other hand risks to let out relevant objects of interest resulting in an increase of false negatives. Therefore, the choice of the right threshold is always a balance. This research determines 50% as the default confidence threshold. In the second post-processing step, the prediction confidence threshold is elevated to 80%. It is expected, in accordance with hypothesis H 2, that a significant positive impact on the results will occur particularly on areas where numerous building-like objects like plastic sheet greenhouses exist. The extraction results on industrial and commercial areas with freight trains and parking trucks should likewise improve. The prediction confidence of such building-like objects is expected to be relatively low. An increase of the prediction confidence threshold to 80% should result in the elimination of numerous false positives.

3.6.3 RESEARCH QUESTION R 3

The research questions R 1 and R 2 examine model performance on various European and US cities. Research question R 3 goes one deeper and focuses on the sub-region level. Differences in model performance within cities and the impact of certain urban structures and building types on the building extraction results, are of particular interest. In total, 23 sub-regions within 7 target areas, were selected to examine the research questions R 3.1 and R 3.2. It is expected, in accordance with hypothesis H 3.1 that model performance is highest on areas with predominately single family houses. Poor results are expected for areas covered by large and complex industrial and commercial buildings.

In regards of hypothesis H 3.2, it is expected that model performance is consistent on sub-regions with similar characteristics throughout different geographic regions. That means, that a model which achieves good results on single family houses in Chemnitz should also perform well on areas with predominantly single family houses in Daugavpils.

At this point, it has to be emphasized, that hypothesis H 3.2 is opposing the hypotheses H 1.1 and H 1.2. If model performance on certain urban characteristics is consistent throughout different geographic regions (i.e. different target areas), then geographical proximity and closeness between the training and target areas, are of less importance.

4. RESULTS

In this chapter the experimental results are presented, subdivided along the three research questions. Visual evidence is provided in addition to the standard evaluation metric scores.

4.1 RESEARCH QUESTION R 1

The first research question examines two pre-trained Mask R-CNN building extraction models on ten different, previously unseen, target areas. Table 3 reveals that PT USA performs best on the New York target areas with F1 scores of 0.66 and 0.74. This is in accordance with the research hypotheses H 1.1 and H 1.2. PT USA scores lowest on the Barcelona and Girona target areas with F1 scores of 0.11 and 0.34. This result was expected in accordance with H 1.2 due to the lower geographical closeness between those Mediterranean cities and the model training areas in the United States. However, the hypotheses H 1.1 and H 1.2 does not explain the significantly lower performance on the Kosice and Chemnitz target areas in comparison to the Tallinn, Narva and Vienna target areas. Furthermore, the small performance difference of 0.02 between the F1 scores on New York I and the Tallinn target areas is quite surprising, too.

PT AFR achieves consistently significantly lower results than PT USA. Similarly, to PT USA, PT AFR scores lowest on the Barcelona and Girona target areas. Low performance is also observed on the Kosice and Chemnitz target areas. PT AFR achieves the highest F1 score on the Narva and Daugavpils target areas.

Region/Model	Precision	Recall	F1	IoU	Region/Model	Precision	Recall	F1	IoU
EST TLN					**AUS VNA**				
PT USA	**0.64**	**0.65**	**0.64**	**0.48**	PT USA	**0.61**	**0.43**	**0.5**	**0.34**
PT AFR	0.34	0.47	0.39	0.25	PT AFR	0.53	0.37	0.43	0.27
EST NRV					**ESP GIR**				
PT USA	**0.58**	**0.68**	**0.63**	**0.46**	PT USA	**0.45**	**0.27**	**0.34**	**0.2**
PT AFR	0.4	0.56	0.47	0.3	PT AFR	0.37	0.15	0.21	0.12
LTV DGV					**ESP BAR**				
PT USA	**0.63**	**0.58**	**0.61**	**0.43**	PT USA	**0.18**	**0.08**	**0.11**	**0.06**
PT AFR	0.54	0.4	0.46	0.3	PT AFR	0.06	0.01	0.02	0.01
SLK KOS					**USA NY I**				
PT USA	**0.56**	**0.27**	**0.37**	**0.22**	PT USA	**0.58**	**0.76**	**0.66**	**0.48**
PT AFR	0.25	0.13	0.17	0.09	PT AFR	0.23	0.26	0.24	0.14
GER CHM					**USA NY II**				
PT USA	**0.37**	**0.27**	**0.37**	**0.23**	PT USA	**0.74**	**0.73**	**0.74**	**0.68**
PT AFR	0.2	0.18	0.19	0.11	PT AFR	0.28	0.17	0.21	0.12

Table 3: Building extraction results of PT USA and PT AFR. The highest scores are depicted in bold.

The Figures 11 and 12 present some examples of the building extraction results of PT USA and PT AFR. Both models were able to extract single family houses relatively well. PT USA performed well on apartment blocks with regular shapes. However, PT USA struggles with long apartment blocks with irregular shapes. PT AFR fails to detect numerous apartment blocks and apartment buildings. Interestingly, the model detects small rooftop dwellings on top of the apartment blocks but fails to extract the building per se (Figure 11). PT AFR performed particularly weak on adjacent apartment buildings with specific forms (Figure 11) and it produced numerous false positives on sport courts or parking lots (Figure 12). Both models perform weak on large industrial and commercial buildings. Parts are missed out and, more critically, the models fragmented those buildings in multiple parts. Similar extraction problems on those object types were observed by Prakash et al. (2022) and Sawa et al. (2024).

Figure 11: Building extraction results on the Vienna target area. The top row provides the satellite image and the ground reference of the target area. The bottom row depicts the results of PT USA (in red) and PT AFR (in purple).

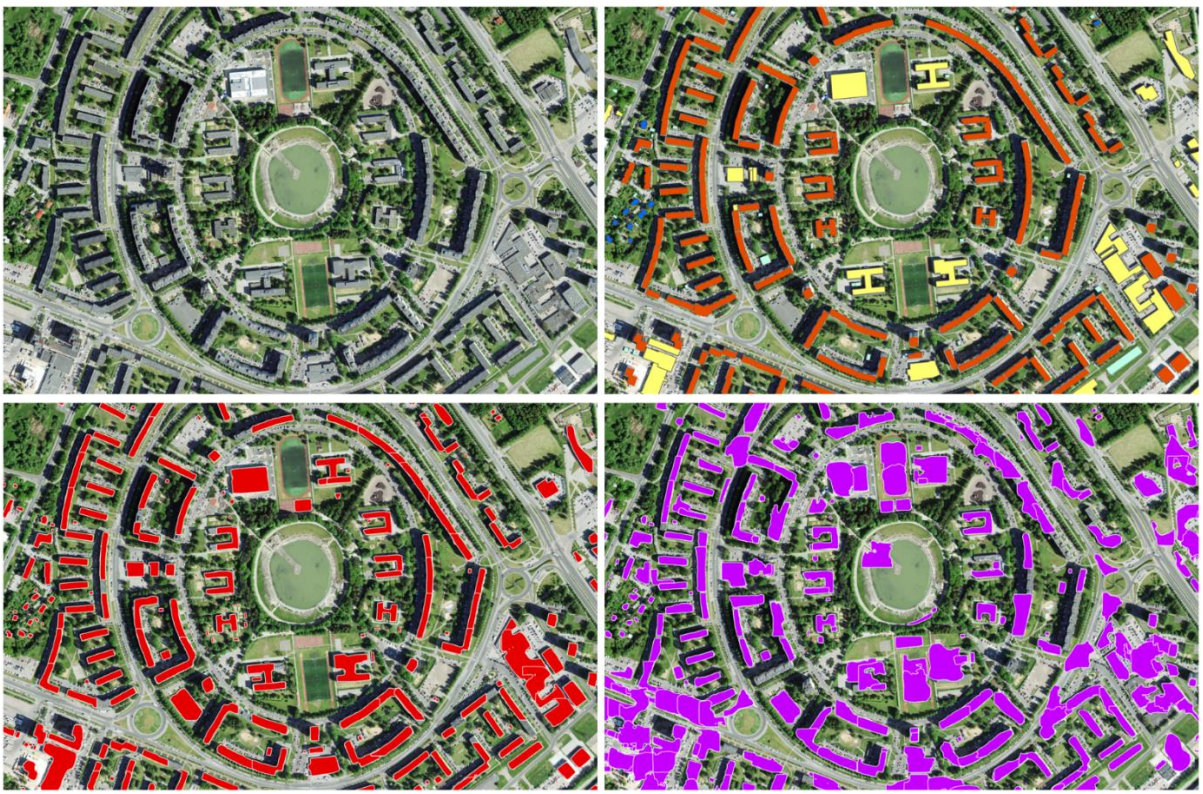

Figure 12: Building extraction results on the Tallinn target area. The top row shows the satellite image (left) and the ground reference (right). The bottom row depicts the extraction result of PT USA (red) and PT AFR (purple).

In the next research experiment, the three fine-tuned models FT USA TLN, FT USA VNA, and FT AFR VNA were executed on the same ten target areas to test H 1.1 and H 1.2. Table 4 provides the evaluation metrics of the fine-tuned models. The results of the pre-trained models, PT USA and PT AFR, are depicted in the table for better comparison. The F1 scores of 0.66 and 0.74 of PT USA on the two New York target areas remain the best recorded result also after the introduction of the three fine-tuned models. FT USA VNA reaches a maximum F1 score of 0.64 on the New York II target area, meanwhile FT USA TLN performs best on the Daugavpils target area with an F1 score of 0.62. Table 4 shows that the performance of PT USA and its two fine-tuned variants are on most of the European target areas, relatively close together. PT AFR and FT AFR VNA achieve significantly lower results on all but one target area.

Region/Model	Precision	Recall	F1	IoU	Region/Model	Precision	Recall	F1	IoU
EST TLN					**AUS VNA**				
PT USA	**0.64**	0.65	**0.64**	**0.48**	PT USA	**0.61**	0.43	0.5	0.34
FT USA TLN	0.47	**0.75**	0.57	0.4	FT USA TLN	0.5	0.58	**0.54**	**0.37**
FT USA VNA	0.46	0.72	0.56	0.39	FT USA VNA	0.45	**0.64**	0.53	0.36
PT AFR	0.34	0.47	0.39	0.25	PT AFR	0.53	0.37	0.43	0.27
FT AFR VNA	0.28	0.64	0.39	0.24	FT AFR VNA	0.33	0.49	0.39	0.24
EST NRV					**ESP GIR**				
PT USA	**0.58**	0.68	**0.63**	**0.46**	PT USA	**0.45**	0.27	0.34	0.2
FT USA TLN	0.33	**0.79**	0.46	0.3	FT USA TLN	0.39	0.37	**0.38**	**0.23**
FT USA VNA	0.34	**0.79**	0.48	0.31	FT USA VNA	0.3	**0.39**	0.34	0.2
PT AFR	0.4	0.56	0.47	0.3	PT AFR	0.37	0.15	0.21	0.12
FT AFR VNA	0.18	0.68	0.28	0.16	FT AFR VNA	0.19	0.31	0.24	0.13
LTV DGV					**ESP BAR**				
PT USA	**0.63**	0.58	0.61	0.43	PT USA	0.18	0.08	0.11	0.06
FT USA TLN	0.54	0.74	**0.62**	**0.45**	FT USA TLN	**0.21**	**0.12**	**0.15**	**0.08**
FT USA VNA	0.46	**0.76**	0.57	0.4	FT USA VNA	0.1	0.1	0.1	0.05
PT AFR	0.54	0.4	0.46	0.3	PT AFR	0.06	0.01	0.02	0.01
FT AFR VNA	0.37	0.63	0.47	0.3	FT AFR VNA	0.06	0.07	0.07	0.03
SLK KOS					**USA NY I**				
PT USA	**0.56**	0.27	0.37	0.22	PT USA	**0.58**	**0.76**	**0.66**	**0.48**
FT USA TLN	0.36	**0.42**	**0.39**	**0.24**	FT USA TLN	0.36	0.68	0.47	0.31
FT USA VNA	0.41	0.35	0.38	0.23	FT USA VNA	0.34	0.74	0.47	0.3
PT AFR	0.25	0.13	0.17	0.09	PT AFR	0.23	0.26	0.24	0.14
FT AFR VNA	0.18	0.33	0.23	0.13	FT AFR VNA	0.13	0.46	0.2	0.11
GER CHM					**USA NY II**				
PT USA	**0.37**	0.27	0.37	0.23	PT USA	**0.74**	0.73	**0.74**	**0.68**
FT USA TLN	**0.37**	**0.57**	**0.45**	**0.29**	FT USA TLN	0.5	0.64	0.56	0.39
FT USA VNA	0.35	0.5	0.41	0.26	FT USA VNA	0.53	**0.81**	0.64	0.47
PT AFR	0.2	0.18	0.19	0.11	PT AFR	0.28	0.17	0.21	0.12
FT AFR VNA	0.14	0.43	0.21	0.12	FT AFR VNA	0.21	0.41	0.27	0.16

Table 4: Building extraction results of all five examined models. The highest scores are depicted in bold.

An unexpected result of the presented research experiment is the fact, that PT USA outperforms FT USA TLN on the Tallinn and Narva target areas by a large margin. In accordance with the formulated research hypotheses and numerous research publications on transfer learning, fine-tuning on the Tallinn training area should have significantly increase the performance on those target areas. However, the fine-tuning even decreased the building extraction result significantly. Model fine-tuning on the Vienna target area, which has a significantly higher geographical proximity to the Estonian target areas, likewise underperformed the unmodified pretrained US model. These findings are in clear contradiction to the formulated research hypotheses H 1.1 and H 1.2.

Table 4 additionally reveals a significant increase of the recall through model fine-tuning on all examined target areas except on NY I and NY II. The recall values of FT USA TLN and FT USA VNA are relatively close together on almost all target areas. Both outperform in terms of the recall FT AFR VNA significantly on all target areas.

Fine-tuning likewise caused a significant decrease of the precision scores of all three fine-tuned models on all target areas.

Appendix 4 provides the exact numbers of the correctly extracted buildings as well as the number of false predictions. The fine-tuned models predict in general up to 10% more building footprints correctly (i.e. TP) than the two unmodified, pre-trained models. At the same time, the fine-tuning caused also a massive increase of falsely predicted buildings (i.e. FP). FT USA TLN, FT USA VNA and FT AFR VNA doubled or even tripled the number of FP in comparison to PT USA and PT AFR on most of the examined target areas. Finally, the fine-tuned model variants missed in general fewer buildings (i.e. FN) than their pre-trained counterparts. The increase of the true positives and the decrease of false negatives led to the increase o the recall values. The massive increase of false positives, on the other hand, led to a significant deterioration of the fine-tuned model's precision.

This general pattern of increasing recall and decreasing precision through the application of model fine-tuning does not count for the New York target areas. PT USA generates on that particular locations the most true positives and least false negatives as well as false positives with the exception of FT USA VNA on New York II. Another exception of the described general trend is observable on the Chemnitz target area. Here, the precision of PT USA is on the same level with PT USA TLN and only slightly better than PT USA VNA. Annex 4 reveals that PT USA generates an unusual high number of false positives and relatively low number of true positives on the Chemnitz target area.

The F1-score is the harmonious average of the recall and precision values. It takes the true positives, false positives and false negatives into account. On the Daugavpils, Kosice, Chemnitz, Vienna and Girona target areas, the higher number of correctly detected buildings by FT USA TLN and FT USA VNA was sufficient enough to compensate for the larger number of false predictions in order to outperform PT USA. The performance of FT USA TLN and FT USA VNA is close together. Nevertheless, FT USA TLN achieves slightly higher F1 scores on all target areas except on New York II.

The fine-tuning of PT AFR led to ambiguous results. The performance gap between PT AFR and the fine-tuned variant FT AFR VNA is, in general, low. The fine-tuning of PT AFR resulted in the exact same pattern like the previously described fine-tuning of PT USA. The number of true positives increased significantly. On six target areas, FT AFR VNA extracted more than twice buildings correctly. It likewise reduced the number of false negatives on all target areas. However, the number of false positives tripled in comparison to PT AFR. Consequently, the fine-tuning resulted only in a minor increase of the F1 scores on seven target areas. FT AFR VNA still underperforms PT USA and its three fine-tuned variants significantly on all target areas.

The following figures provide some visual examples of the research findings. Figure 13 shows that the pre-trained models struggled with the coastal areas of Tallinn. PT USA and PT AFR, both produced numerous false positives on the water body. The fine-tuning on the Tallinn training area prevented such false positives successfully. The fine-tuning on the Vienna training area led to a significantly better performance on the coastal zone of Tallinn for FT AFR VNA in comparison to PT AFR.

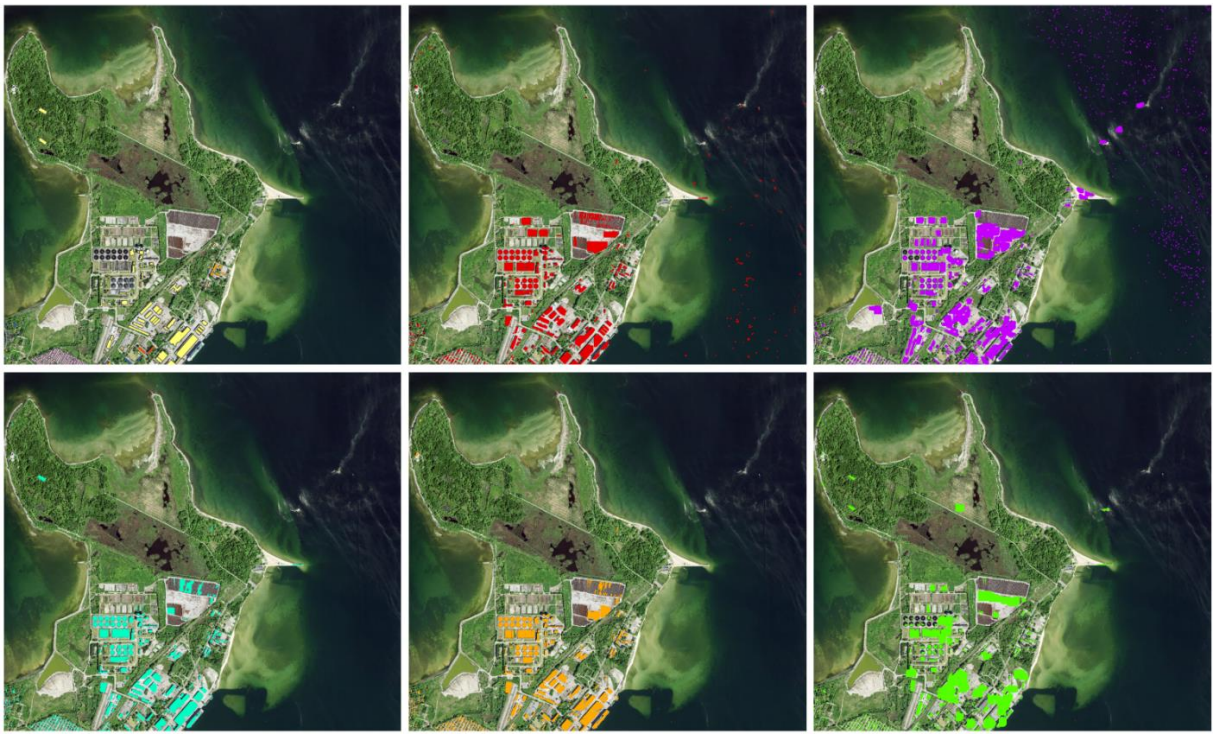

Figure 13: Building extraction results on the Tallinn target area. PT USA (red), PT AFR (purple), FT USA TLN (turquoise), FT USA VNA (orange) and FT AFR VNA (green). Ground reference is shown on the top left. Appendix 5 provides an enlarged format.

Figure 14 provides a visual insight of the fine-tuning results on the Vienna target area. PT AFR and FT AFR VNA, both perform relatively well on single family houses. However, the results significantly improved on the apartment blocks through the fine-tuning.

Figure 14: Building extraction results on the Vienna target area. PT AFR (purple) and FT AFR VNA (green). Both models extract the single family houses well. PT AFR fails to extract the apartment blocks in the middle.

Figure 15 provides another visual example for performance improvement through model fine-tuning. PT USA detects on the Chemnitz target area the apartment buildings correctly but it does not separate the adjacent buildings well. However, the reference dataset labels every adjacent building separately. Consequently, PT USA is penalized with a large number of falsely missed buildings (i.e. FN). FT USA TLN, on the other hand, extracts significantly more of the adjacent apartment buildings individually. This results in more true positives and less false negatives.

Figure 15: Building extraction results on the Chemnitz target area. PT USA (red) and FT USA TLN (turquoise). The top row provides the ground reference.

Figure 16 demonstrates the observed increase of false positives through the model fine-tuning. The Narva target area covers some suburban areas characterized by irregularly arranged single family houses, small dwellings and a large amount of plastic sheet greenhouses. PT USA extracts most of the detached houses and some of the dwellings and garages. FT USA TLN, on the other hand, extracts additionally most of the dwellings and greenhouses. However, numerous dwellings and none of the plastic greenhouses is labelled as buildings in the reference dataset. This increased the number of false positives of the fine-tuned models significantly.

Figure 16: Building extraction on the Narva target area. PT USA (red) and FT USA TLN (turquoise). The top row provides the ground reference.

4.2 RESEARCH QUESTION R 2

The previously reported research experiment revealed that model fine-tuning leads to a significant increase in correct building extraction and reduction of missed buildings in comparison to the unmodified pre-trained models. However, the increase of true positives and decline of false negatives comes with a tremendous inflation of false positives. The fine-tuned models produce two to three times more false positives than their pre-trained counterparts. The second research question examines the effect of post-processing to improve the results of the building extraction tasks. Table 5 depicts the results on all target areas before and after the execution of the post-processing.

TA/Model	Precision				Recall				F1				IoU			
EST TLN	Basic	DB	X25	T80	Basic	DB	X25	T80	Basic	DB	X25	T80	Basic	DB	X25	T80
PT USA	**0.64**	**0.75**	**0.78**	**0.86**	0.65	0.77	0.8	0.77	**0.64**	**0.77**	**0.8**	**0.81**	**0.48**	**0.63**	**0.67**	**0.69**
FT USA TLN	0.47	0.54	0.7	0.76	**0.75**	**0.86**	**0.86**	**0.85**	0.57	0.66	0.77	0.8	0.4	0.49	0.63	0.67
FT USA VNA	0.46	0.53	0.74	0.8	0.72	0.85	0.84	0.83	0.56	0.65	0.79	**0.81**	0.39	0.48	0.65	**0.69**
PT AFR	0.34	0.4	0.41	0.63	0.47	0.54	0.55	0.27	0.39	0.46	0.47	0.38	0.25	0.3	0.31	0.24
FT AFR VNA	0.28	0.34	0.49	0.6	0.64	0.69	0.69	0.67	0.39	0.46	0.57	0.63	0.24	0.3	0.4	0.46
EST NRV																
PT USA	**0.58**	**0.66**	**0.71**	**0.78**	0.68	0.76	0.81	0.75	**0.63**	**0.71**	**0.75**	**0.77**	**0.46**	**0.55**	**0.61**	**0.62**
FT USA TLN	0.33	0.37	0.63	0.66	**0.79**	0.84	**0.84**	0.82	0.46	0.52	0.7	0.73	0.3	0.35	0.54	0.58
FT USA VNA	0.34	0.39	0.64	0.71	**0.79**	**0.85**	**0.84**	**0.83**	0.48	0.53	0.73	0.76	0.31	0.36	0.58	**0.62**
PT AFR	0.4	0.53	0.58	0.74	0.56	0.59	0.64	0.4	0.47	0.56	0.61	0.52	0.3	0.39	0.43	0.35
FT AFR VNA	0.18	0.23	0.43	0.41	0.68	0.68	0.67	0.68	0.28	0.35	0.52	0.51	0.16	0.13	0.35	0.34
LTV DGV																
PT USA	**0.63**	0.7	0.73	**0.81**	0.58	0.59	0.67	0.6	0.61	0.64	0.7	0.69	0.43	0.47	0.54	0.48
FT USA TLN	0.54	**0.71**	**0.74**	0.79	0.74	0.71	**0.74**	0.71	**0.62**	**0.65**	**0.73**	**0.74**	**0.45**	**0.49**	0.57	0.59
FT USA VNA	0.46	0.53	0.73	0.79	**0.76**	**0.74**	**0.74**	**0.72**	0.57	0.62	**0.73**	0.75	0.4	0.45	**0.58**	**0.6**
PT AFR	0.54	0.64	0.66	0.78	0.4	0.38	0.44	0.16	0.46	0.48	0.53	0.27	0.3	0.31	0.36	0.15
FT AFR VNA	0.37	0.47	0.58	0.69	0.63	0.46	0.49	0.54	0.47	0.47	0.53	0.62	0.3	0.31	0.36	0.44
SLK KOS																
PT USA	**0.56**	**0.74**	**0.75**	**0.8**	0.27	0.56	0.62	0.54	0.37	**0.69**	**0.68**	0.65	0.22	**0.47**	**0.52**	0.48
FT USA TLN	0.36	0.42	0.62	0.64	**0.42**	**0.69**	**0.69**	**0.67**	**0.39**	0.52	0.65	0.65	**0.24**	0.35	0.48	0.49
FT USA VNA	0.41	0.53	0.71	0.74	0.35	0.66	0.66	0.64	0.38	0.59	**0.68**	**0.68**	0.23	0.42	**0.52**	**0.52**
PT AFR	0.25	0.39	0.43	0.5	0.13	0.23	0.26	0.09	0.17	0.29	0.32	0.16	0.09	0.17	0.19	0.08
FT AFR VNA	0.18	0.24	0.45	0.54	0.33	0.44	0.42	0.5	0.23	0.31	0.44	0.52	0.13	0.18	0.28	0.35
GER CHM																
PT USA	**0.37**	**0.63**	**0.69**	**0.74**	0.37	0.6	0.6	0.51	0.37	**0.61**	**0.67**	**0.66**	0.23	**0.44**	**0.5**	0.49
FT USA TLN	**0.37**	0.45	0.61	0.63	**0.57**	**0.73**	**0.73**	**0.68**	**0.45**	0.56	0.66	0.65	**0.29**	0.38	**0.5**	**0.49**
FT USA VNA	0.35	0.45	0.63	0.67	0.5	0.66	0.66	0.63	0.41	0.54	0.64	0.65	0.26	0.37	0.47	0.48
PT AFR	0.2	0.34	0.38	0.43	0.18	0.3	0.33	0.13	0.19	0.32	0.35	0.21	0.11	0.19	0.21	0.11
FT AFR VNA	0.14	0.18	0.24	0.36	0.43	0.36	0.35	0.44	0.21	0.24	0.29	0.39	0.12	0.13	0.15	0.24
AUS VNA																
PT USA	**0.61**	**0.77**	**0.79**	**0.84**	0.43	0.6	0.82	0.77	0.5	0.67	**0.81**	0.8	0.34	**0.51**	**0.66**	0.67
FT USA TLN	0.5	0.6	0.72	0.79	0.58	0.76	**0.85**	**0.83**	**0.54**	**0.68**	0.78	**0.81**	**0.37**	**0.51**	0.64	**0.68**
FT USA VNA	0.45	0.53	0.75	0.78	**0.64**	**0.79**	0.84	0.82	0.53	0.63	0.79	0.8	0.36	0.46	**0.66**	0.67
PT AFR	0.53	0.63	0.64	0.82	0.37	0.51	0.68	0.38	0.43	0.56	0.66	0.52	0.27	0.39	0.5	0.35
FT AFR VNA	0.33	0.44	0.57	0.67	0.49	0.59	0.69	0.69	0.39	0.51	0.62	0.68	0.24	0.34	0.45	0.51
ESP GIR																
PT USA	**0.45**	**0.71**	**0.74**	**0.82**	0.27	0.58	0.65	0.6	0.34	**0.64**	**0.7**	0.65	0.2	**0.47**	**0.53**	**0.53**
FT USA TLN	0.39	0.47	0.57	0.71	0.37	0.63	0.66	0.6	**0.38**	0.54	0.61	0.65	**0.23**	0.37	0.44	0.48
FT USA VNA	0.3	0.44	0.62	0.69	**0.39**	**0.67**	**0.69**	**0.66**	0.34	0.53	0.65	**0.67**	0.2	0.36	0.48	0.51
PT AFR	0.37	0.57	0.61	**0.82**	0.15	0.35	0.4	0.16	0.21	0.44	0.48	0.27	0.12	0.28	0.32	0.16
FT AFR VNA	0.19	0.3	0.44	0.53	0.31	0.46	0.47	0.48	0.24	0.36	0.45	0.5	0.13	0.22	0.29	0.33
ESP BAR																
PT USA	0.18	**0.38**	**0.46**	**0.51**	0.08	**0.22**	**0.23**	**0.19**	0.11	**0.28**	**0.3**	**0.28**	0.06	**0.16**	**0.18**	**0.16**
FT USA TLN	**0.21**	0.21	0.27	0.34	**0.12**	0.21	0.21	0.12	**0.15**	0.21	0.21	0.12	**0.08**	0.11	0.13	0.09
FT USA VNA	0.1	0.16	0.27	0.27	0.1	0.21	0.21	0.16	0.1	0.18	0.24	0.2	0.05	0.1	0.13	0.11
PT AFR	0.06	0.05	0.07	0.13	0.01	0.03	0.03	0	0.02	0.04	0.04	0.01	0.01	0.02	0.02	0
FT AFR VNA	0.06	0.01	0.02	0.04	0.07	0.04	0.04	0.02	0.07	0.02	0.03	0.03	0.03	0.01	0.01	0
USA NY I																
PT USA	**0.58**	**0.71**	**0.78**	**0.82**	**0.76**	**0.79**	**0.8**	**0.77**	**0.66**	**0.75**	**0.79**	**0.8**	**0.48**	**0.6**	**0.66**	**0.66**
FT USA TLN	0.36	0.48	0.7	0.8	0.68	0.74	0.73	0.73	0.47	0.58	0.72	0.76	0.31	0.41	0.57	0.62
FT USA VNA	0.34	0.43	0.74	0.76	0.74	**0.79**	**0.8**	**0.77**	0.47	0.56	0.77	0.77	0.3	0.39	0.63	0.63
PT AFR	0.23	0.28	0.33	0.51	0.26	0.27	0.28	0.06	0.24	0.28	0.3	0.12	0.14	0.16	0.14	0.06
FT AFR VNA	0.13	0.17	0.31	0.42	0.46	0.31	0.31	0.37	0.2	0.22	0.31	0.4	0.11	0.12	0.18	0.25
USA NY II																
PT USA	**0.74**	**0.69**	**0.73**	0.72	0.73	0.69	**0.8**	0.74	**0.74**	**0.69**	**0.76**	**0.73**	**0.68**	**0.53**	**0.62**	**0.57**
FT USA TLN	0.5	0.53	0.69	**0.74**	0.64	0.75	0.77	0.72	0.56	0.62	0.73	**0.73**	0.39	0.45	0.57	**0.57**
FT USA VNA	0.53	0.47	0.63	0.61	**0.81**	**0.8**	**0.8**	**0.76**	0.64	0.59	0.7	0.67	0.47	0.43	0.54	0.51
PT AFR	0.28	0.41	0.46	0.62	0.17	0.26	0.31	0.06	0.21	0.32	0.37	0.12	0.12	0.19	0.22	0.06
FT AFR VNA	0.21	0.24	0.42	0.59	0.41	0.24	0.25	0.48	0.27	0.24	0.32	0.53	0.16	0.14	0.19	0.36

Table 5: Performance after post-processing. The "Basic" column provides the results without any post-processing. "DB" provides the results for the "Dissolve Boundaries" post-processing. "X25" provides the results for the following erasure of all objects smaller than 25 m². "T80" provides the results for the following increase of the prediction probability threshold from 50% to 80%.

The dissolve boundaries post-processing method aims to reduce false building predictions (i.e. reduce FP) and to reduce the number of missed buildings (i.e. reduce FN). One of the main sources for false positives in this research are large public, commercial and industrial buildings. On such buildings, it often occurs, that the model predicts different parts of the same buildings as separate relevant objects. In a pixel wise extraction, this would not be a problem as all positive pixels can be counted as correct as long as they are located inside a reference building polygon. This does not apply for the instance segmentation which is executed in this research. Here, if a building is segmented into 20 predicted objects of interest by the model, all polygons can be counted as false positives if none of them covers more than 50% of the reference building. Dissolving the boundaries between those predicted segments compiles them often into one large prediction polygon which represents the reference building sufficiently to count as true positive. At least, the number of polygons which still count as false positives, is through the merging process significantly reduced. FT USA TLN extracts multiple polygons for single buildings (Figure 17). This is particularly the case for the huge building on the right. The figure also depicts the positive effect of the dissolve boundaries post-processing. It merges the adjacent polygons and produces numerous true positives. The large building on the right is still not correctly extracted. However, the number of false negatives is significantly reduced, too.

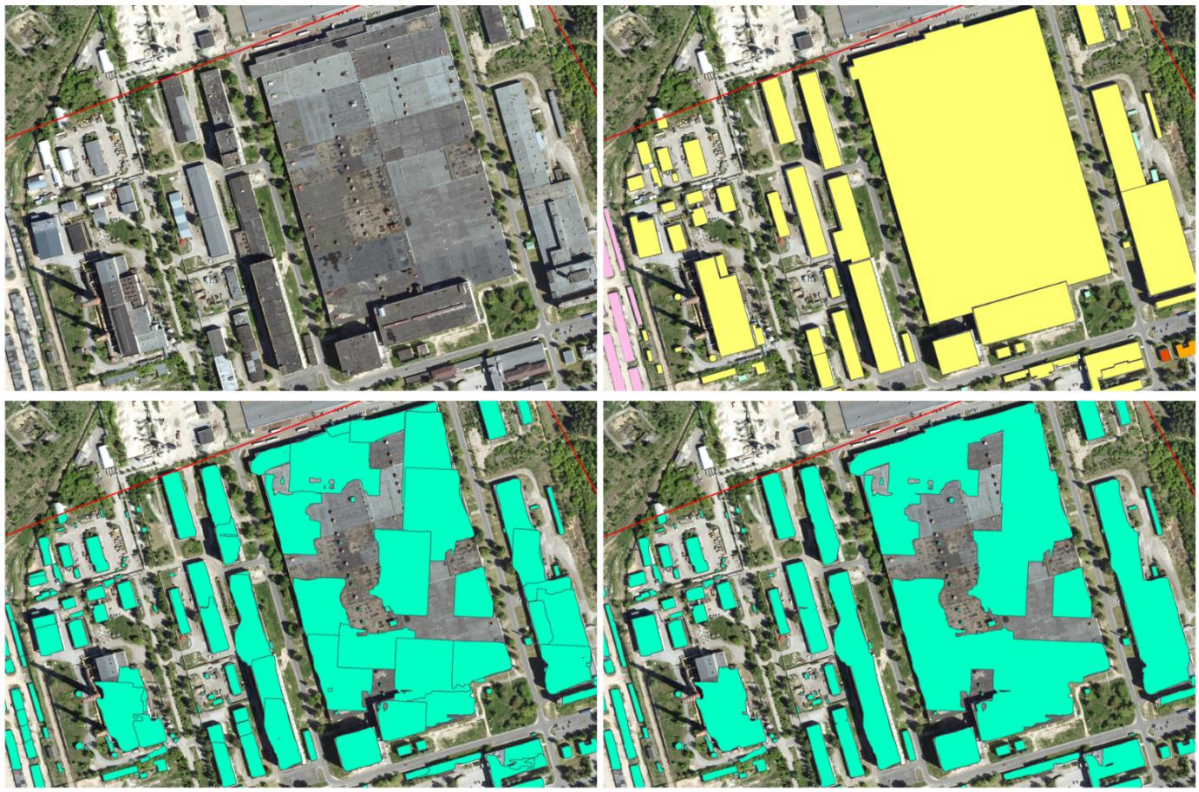

Figure 17: Building extraction on the Daugavpils target area. FT USA TLN (turquoise) on industrial buildings. On the bottom left, the extraction result is depicted without post-processing. On the bottom right, the result after the dissolve boundaries post-processing is depicted. The top row provides the ground reference.

Appendix 4 reveals that the dissolve boundaries method reduces the number of false positives of all five models significantly on all target areas. However, the greatest reduction of false positives, often by even more than 50%, occurs on the Kosice, Chemnitz, Girona, Barcelona and New York I target areas.

The precision score is the ratio between true positives and false positives. It provides a good estimation of the post-processing impact as it considers not only a potential reduction of false positives but also the number of true positives after the post-processing. Table 5 reveals a precision increase for all models on most of the target areas. However, the level of improvement varies significantly. The positive impact of the dissolve boundaries post-processing on the precision score is particularly high on the Daugavpils, Kosice, Chemnitz, Vienna and Girona target areas. The impact is smaller on the Narva and Tallinn target areas. The drop of the precision score on the Barcelona and New York II target areas is caused by the massive reduction of true positives.

The dissolve boundaries post-processing reduced not only false positives but also false negatives. The dissolvement of boundaries was executed on the model predictions as well as on the building reference dataset. Appendix 4 shows the drop in the number of reference buildings after the post-processing. The partition of individual residential buildings into multiple objects as well as the individual labelling of adjacent building parts in several reference datasets, is abrogated by the dissolve boundaries post-processing. The individual labelling of each garage within large garage blocks in the Kosice ground reference dataset, or the labelling of each individual building in long rows of terraced houses in Kosice or New York, are likewise dissolved. The reduction of the number of reference buildings through the dissolve boundaries post-processing ranges between less than 25% (e.g. Tallinn, Narva, Vienna) and up to 40% for the Kosice, Chemnitz, Girona and Barcelona target areas. Figure 18 depicts the effect of the dissolve boundaries post processing on the Kosice target area. The adjacent individual terraced houses and garages in the reference dataset are merged into single compact buildings. FT USA TLN extracted some but not all of the adjacent terraced buildings on the Kosice target area. It failed, like the other examined models, to segment the large garage blocks into separate adjacent garages. The dissolvement of those boundaries in the extraction result and in the ground reference dataset, led to a large decrease of false negatives.

Figure 18: Building extraction on the Kosice target area. FT USA TLN (turquoise). On the left before the post-processing and on the right after the execution of the dissolve boundaries step. The top row provides the ground reference, Ground reference after the application of the dissolve boundaries post-processing is shown on the tight.

Table 5 shows the spatial distribution of the significant increase of the recall scores. On the Kosice, Chemnitz, Vienna and Girona target areas, the recall of all models is improved by at least 10%, mostly by more than 20%. Contrariwise, on the Daugavpils and New York II target areas, the very same post-processing method causes a decrease of the recall scores of all models with only two exceptions. The F1 scores of all models improved significantly on the Kosice, Chemnitz, Vienna and Girona target areas. On the other target areas, the impact is only marginal, or a significant improvement is achieved only by individual models. A decrease of the F1 score occurs only on the Barcelona and New York II target areas.

In the next post-processing step, all predicted and referenced buildings smaller than 25 m², were deleted. This reduces for all models and on all target areas the number of true positives, false positives and false negatives. False positives are avoided when small building-like objects as garages or sheds are extracted correctly by the model but those objects are missing in the reference datasets. Small objects which were erroneously predicted as buildings like trucks or trains, are likewise removed. Furthermore, small building artefacts caused by the fragmented extraction of large buildings, which are not adjacent and therefore not eliminated through the dissolve boundaries post-processing, are erased during this second post-processing round, too.

The number of false negatives is reduced by not penalizing the missed extraction of small objects like sheds or small garages which are labelled as buildings in some of the reference datasets. However, the correct predictions of small objects which are defined as buildings in the respective reference datasets, are likewise eliminated. This leads to a simultaneous decrease of true positives.

Figure 19 demonstrates the positive effect of the deletion of small objects on FT USA TLN on the Narva target area. A large amount of extracted small building-like objects (e.g. small sheds and plastic greenhouses) are deleted.

Figure 19: Building extraction results on the Narva target area. PT USA (red) and FT USA TLN (turquoise) on the right after the deletion of small objects. The erased small objects are highlighted in yellow.

Figure 20 depicts the same pattern on the Vienna target area where FT AFR VNA benefits from post-processing stronger than PT AFR. Figure 21 provides an additional example, where the deletion of predicted objects smaller than 25 m² results in the erasure of the falsely extracted trucks and cars.

Figure 20: Building extraction results the Vienna target area. PT AFR (purple) and FT AFR VNA (green) on the right after the deletion of small objects. The erased small objects are highlighted in yellow.

Figure 21: Building extraction results on the New York II target area. PT USA (red) and FT USA VNA (orange) after the deletion of small objects. The erased small objects are highlighted in yellow.

The impact of the deletion of small objects does not follow a clear geographical distribution. A significant reduction of false positives by at least 50% is observable only for the three fine-tuned models. Table 5 depicts impressively that the precision rate of all three fine-tuned models increased after the second round of post-processing significantly by at least 10%, in numerous cases even by 20%, on all target areas except Barcelona. The impact of the elimination of small objects on the two pre-trained models is marginal.

The effect on the recall scores is less straightforward. A significant increase can be observed only on the Vienna target area. In a few cases, the reduction of false negatives cannot compensate the simultaneous drop of true positives resulting in a small decrease of the recall score.

The F1 scores reveal that significant impact of the removal of small objects follows rather the model characteristics than a distinct spatial distribution. Only the fine-tuned models show a significant increase of the F1 scores by more than 10% on almost all target areas. The Vienna target area is a positive exception where all models, including the two pre-trained models reveal a significant improvement of the F1 scores. None of the examined models benefit on the Barcelona target area significantly from the second post-processing method.

In the last post-processing experiment, the prediction confidence threshold was increased from 50% to 80%. That means, building extraction is executed only in case of a much greater certainty. This results generally in a decrease of the true positives by avoiding the extraction of actually correct prediction due to a lower certainty and the respective increase of false negatives. However, the increase of the confidence threshold can significantly reduce the number of false positives by avoiding the extraction of uncertain and thus often wrong predictions.

Appendix 4 reveals that the increase of the confidence threshold is an effective post-processing step to reduce the number of false positives. However, the impact on the false positives has to be considered together with the simultaneous increase of false

negatives and the parallel decrease of true positives. This is particularly the case for PT AFR, where the massive reduction of false positives comes with a likewise tremendous loss of true positives.

The precision scores of all models increase relatively uniformly on all target areas. Major improvements are recorded only for PT AFR and FT AFR VNA on most of the target areas. FT USA TLN shows, as an exception of the general pattern, on two target areas an increase by more than 10%.

The increase of the false negatives is evident on the relatively uniform moderate decrease of the recall scores of almost all models on all target areas. PT AFR, however, reveals a major reduction of the recall score by more than 20% on most of the target areas. FT AFR VNA, on the other hand, shows a remarkable increase of the recall on the Daugavpils, Kosice, Chemnitz and the two New York target areas caused by the respective decrease of false negatives on those target areas.

The comparison of the F1 scores between the second and third post-processing round depicts a heterogenous result (Table 5). The positive effects of the increase of the confidence threshold are largely counterbalanced through the simultaneous surge of false negatives. The fine-tuned variants of PT USA reveal in most cases a small increase of the F1 scores meanwhile the model´s F1 score alternates between a small increase and a small decrease. FT AFR VNA benefits on all target area from the increase of the probability threshold, in two cases even with a significant increase of the F1 score by more than 10%. PT AFR, on the other hand, experiences a critical deterioration of the F1 score on all target areas.

A distinct spatial pattern is not apparent. The impact on individual models varies significantly throughout the different target areas. The lowest increase of the F1 scores is recorded on the New York II, Kosice and Barcelona target areas.

Figure 22 shows the positive effect of the increase of the prediction confidence threshold on FT USA TLN. Industrial waste sites, which were predicted as buildings, are eliminated through the post-processing. However, the likewise false positives of the oval sewage treatment basins still remain because the model identified those objects as buildings with a confidence of over 90%. Figure 23 provides another example. PT USA extracted large freight wagons as buildings. Those objects were not eliminated through the previous post-processing round because their size is larger than 25 m². However, most of the false positives were successfully deleted through the increase of the confidence threshold. Figure 24, on the other hand, highlights the negative potential of the increase of the probability threshold. PT AFR extracts buildings on the Daugava target area with a significantly lower confidence than PT USA. The increase of the confidence threshold results in a massive reduction of true positives and a parallel increase of false negatives.

Figure 22: Large industrial area in Tallinn.

Figure 23: Building extraction result on the Tallinn target area. FT USA TLN (turquoise) after the increase of the prediction confidence threshold. The prediction confidence of FT USA TLN is shown on the right. The dark red colour depicts an extraction confidence of over 90%, the light red colour a confidence of over 80%, the pink, orange and yellow colours depict confidence levels below 80%. The falsely extracted waste deposit objects are erased but the treatment basins remained.

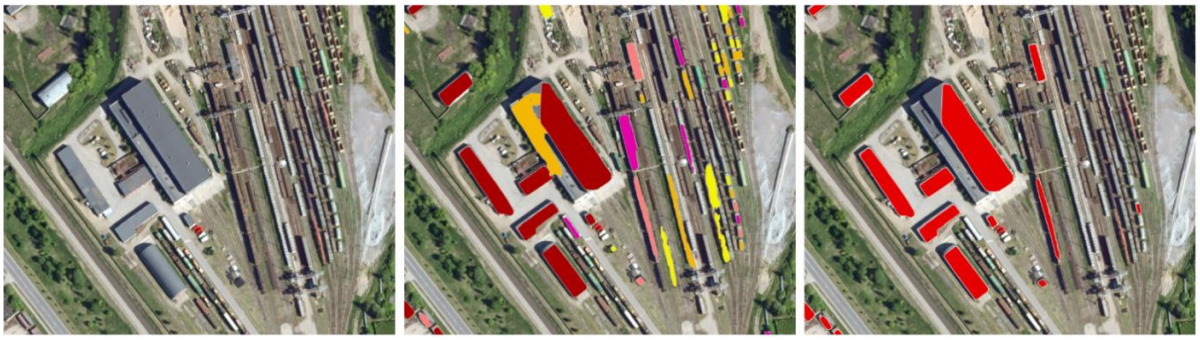

Figure 24:Building extraction results on the Daugavpils target area. PT USA (red). The satellite image of the Daugavpils industrial train station is depicted on the left. The extraction confidence of PT USA is depicted in the middle with the same colour code as above.

Figure 25: Building extraction results on the Daugavpils target area. The top row provides the ground reference. The prediction confidence level of PT USA (left bottom) and PT AFR (right bottom) highlight significant differences between the two models. PT AFR extracts the majority of buildings on the Daugavpils target area with a confidence below 80% resulting in a large number of false negatives after the application of the third post-processing step.

The following two figures provide an interesting example of cross-regional model performance variance. PT USA and its fine-tuned variants, FT USA TLN and FT USA VNA, perform relatively well on a challenging industrial area in New York. The various industrial buildings are largely extracted and none of the models are confused by the chaotic open areas covered by cars, garages and small sheds. However, all three models extracted a large number of white trucks on the highway as well as stabled white containers as buildings. Post-processing eliminated only a part of those false positives because most of the extracted trucks and containers are larger than 25m² and the confidence level of those extractions is often higher than 80%. This is particularly true for FT USA VNA. PT AFR and FT AFR VNA do not extract the white containers and trucks but those two models fail to reach an acceptable result on the New York target area in general.

A similar urban scene with a large amount of parked white trucks or busses in one of Tallinn´s industrial zone is depicted by Figure 27. Here, PT USA and FT USA TLN perform significantly better without extracting any of the vehicles. FT USA VNA extracts some of the parked white vehicles. Those false extractions are erased through the

application of post-processing. PF AFR fails again to produce an acceptable extraction result. FT AFR VNA fails likewise to extract most of the buildings. However, the model extracts some of the white vehicles with such a high confidence that the positive impact of post-processing remains low.

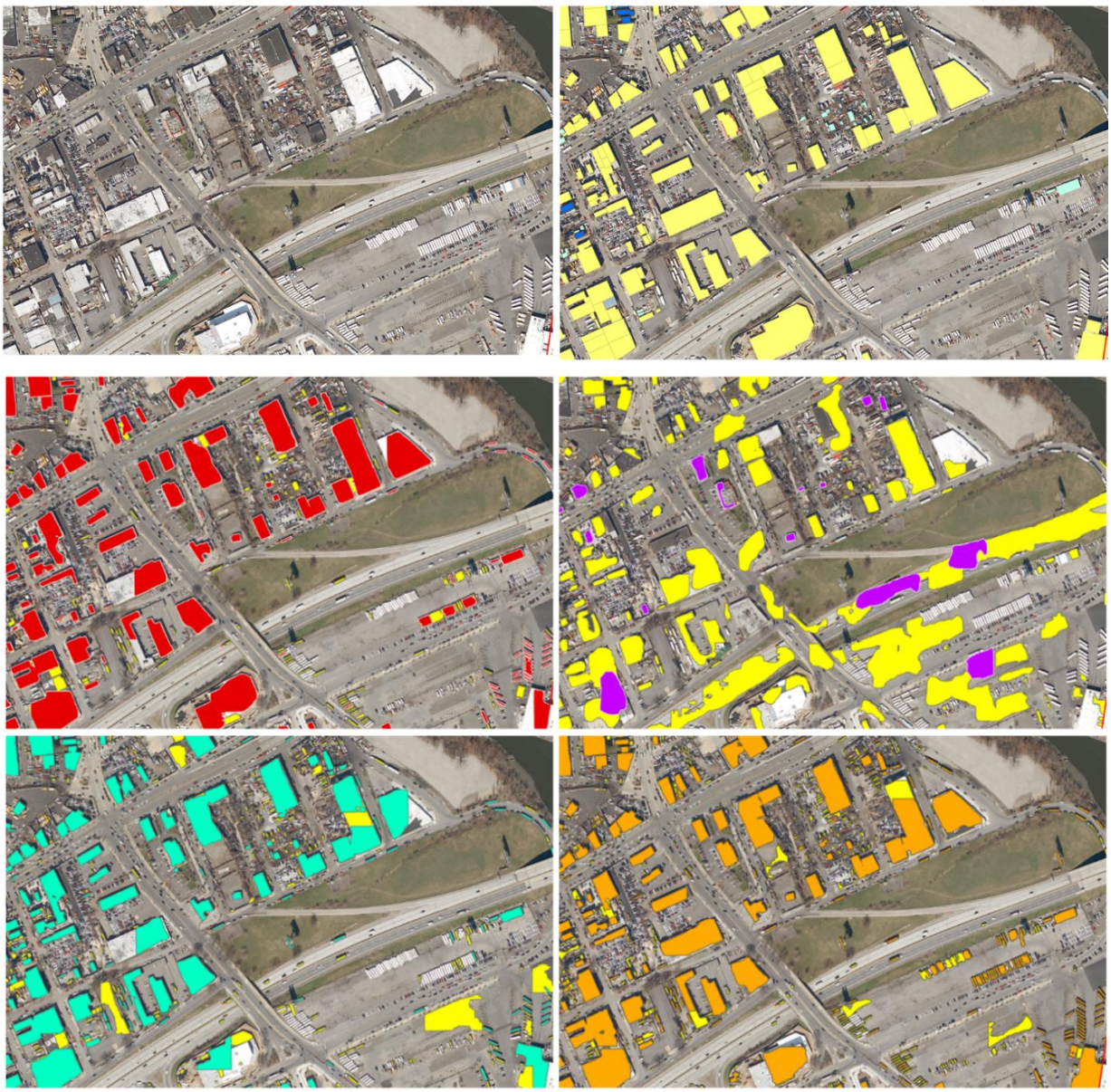

Figure 26: Building extraction results on the New York II target area. PT USA (red), PT AFR (purple), FT USA TLN (turquoise) and FT USA VNA (orange) on an industrial zone after the execution of all three post-processing steps. Extracted objects which were erased through the application of post-processing are highlighted in yellow. The top row provides the ground reference.

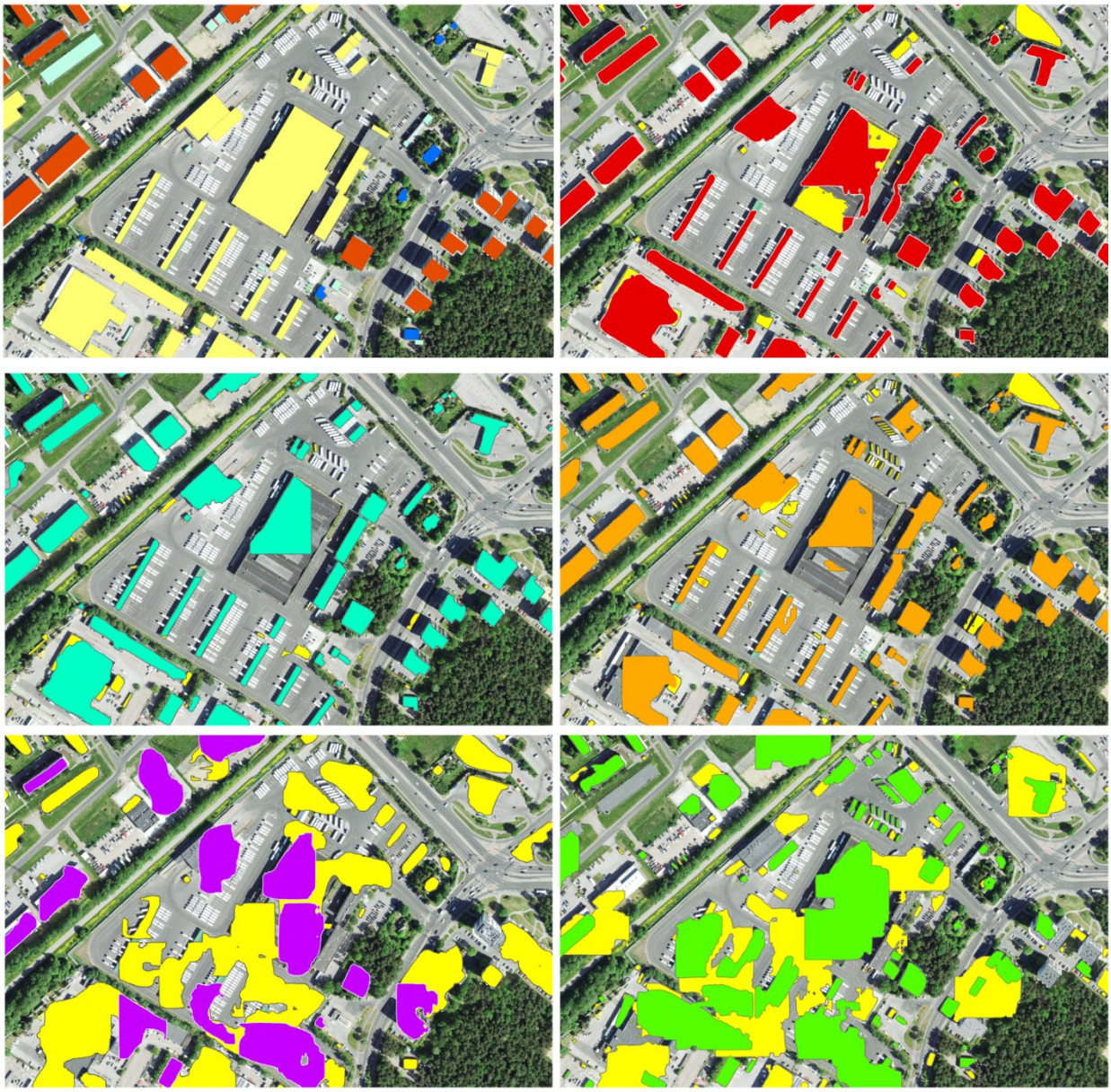

Figure 27: Building extraction results on the Tallinn target area. PT USA (red), FT USA TLN (turquoise), FT USA VNA (orange) and PT AFR (purple) on a commercial zone after the execution of all three post-processing steps. Extracted objects which were erased through the application of post-processing are highlighted in yellow. The top row provides the ground reference.

The Figures 28-30 illustrate the results of the post-processing research experiment. The achieved F1-scores of the five examined models are depicted for each post processing round. The diagrams highlight the positive impact of post-processing.

Post-processing managed to increase the results of PT USA and its fine-tuned variants to remarkable high F1 scores on previously unseen target areas. The weak performance of all models on the Chemnitz, Kosice and Girona target areas improved to an acceptable level. Post processing fails to elevate the poor results on Barcelona; none of the models achieve an F1 score above 0.3. Post-processing works excellent on the Vienna target area. The F1 scores of all models are improved by at least 0.24. FT AFR VNA benefits on all target areas significantly from post-processing, and it

reaches acceptable F1 scores on most of the target areas. Post-processing has a far lower positive impact on PT AFR. The model F1 score remains even after the application of post-processing methods on seven target areas under 0.5.

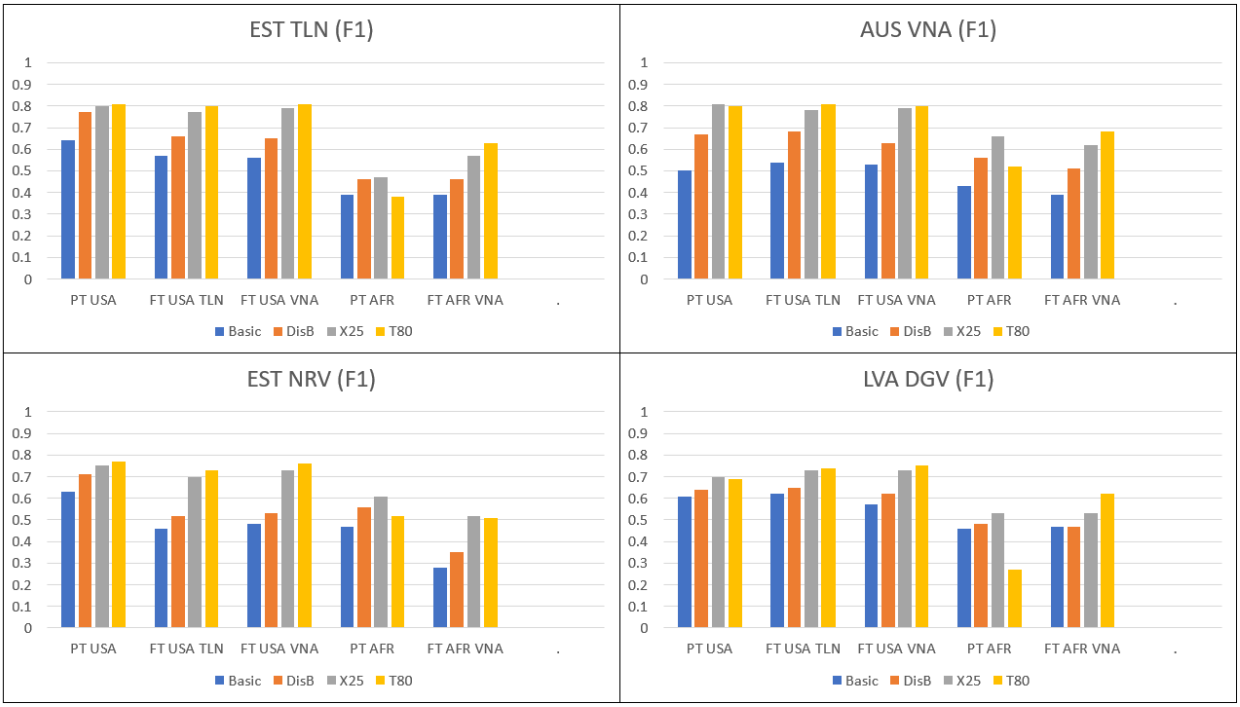

Figure 28: F1 scores of the examined models on the Tallinn, Vienna, Narva and Daugavpils target areas. F1 scores without post-processing ("Basic"), after the application of the dissolve boundaries "DB", erasure of objects smaller than 25 m² ("X25") and the increase of the confidence threshold from 50% to 80% "T80".

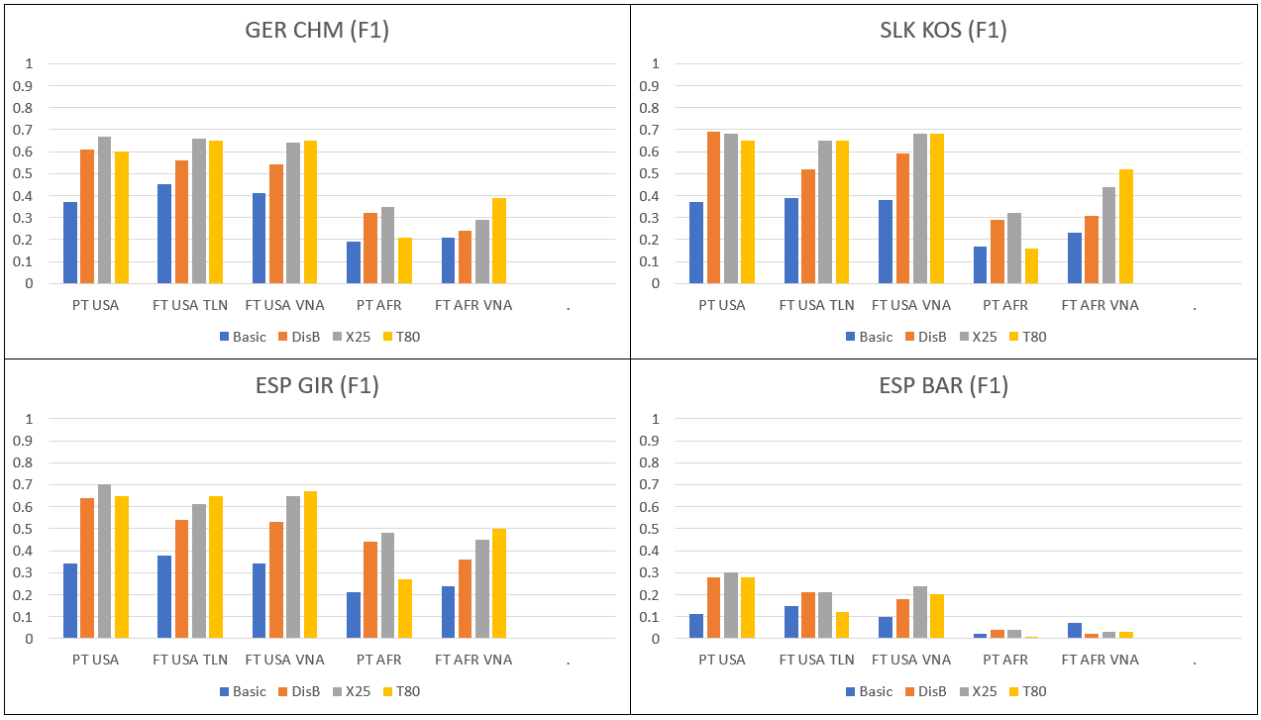

Figure 29: F1 scores of the examined models on the Chemnitz, Kosice, Girona and Barcelona target areas. F1 scores without post-processing ("Basic"), after the application of the dissolve boundaries "DB", erasure of objects smaller than 25 m² ("X25") and the increase of the confidence threshold from 50% to 80% "T80".

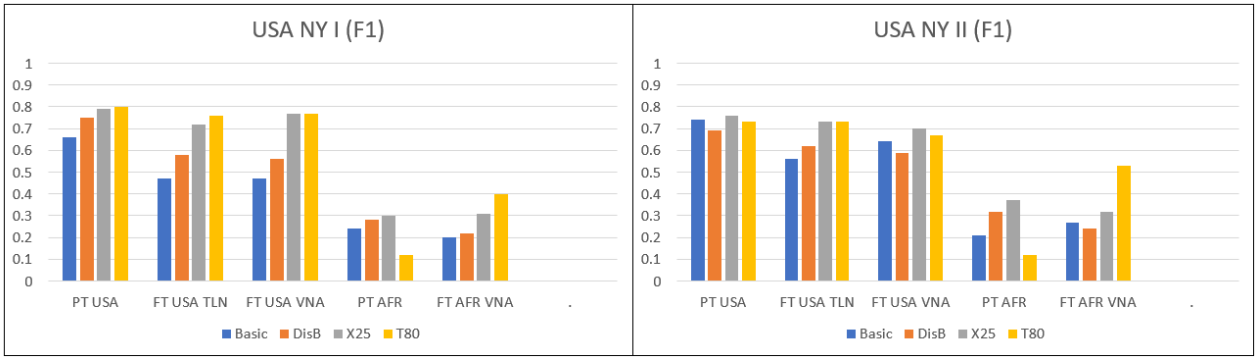

Figure 30: F1 scores of the examined models on the New York I and New York II target areas, F1 scores without post-processing ("Basic"), after the application of the dissolve boundaries "DB", erasure of objects smaller than 25 m² ("X25") and the increase of the confidence threshold from 50% to 80% "T80".

Table 6 provides an overview of the positive impact of post-processing on the extraction results. It depicts the improvement of the F1 scores through the application of post-processing in comparison to the results without any post-processing.

EST TLN	F1 basic	Dif.	F1 best
PT USA	**0.64**	0.17	**0.81**
FT USA TLN	0.57	0.23	0.8
FT USA VNA	0.56	**0.25**	**0.81**
PT AFR	0.39	0.08	0.47
FT AFR VNA	0.39	0.24	0.63
EST NRV			
PT USA	**0.63**	*0.14*	**0.77**
FT USA TLN	0.46	0.27	0.73
FT USA VNA	0.48	**0.28**	0.76
PT AFR	0.47	0.14	0.61
FT AFR VNA	0.28	0.24	0.52
LTV DGV			
PT USA	0.61	0.09	0.7
FT USA TLN	**0.62**	0.12	0.74
FT USA VNA	0.57	**0.18**	**0.75**
PT AFR	0.46	0.07	0.53
FT AFR VNA	0.47	0.15	0.62
SLK KOS			
PT USA	0.37	**0.32**	**0.69**
FT USA TLN	**0.39**	0.26	0.65
FT USA VNA	0.38	0.3	0.68
PT AFR	0.17	0.15	0.32
FT AFR VNA	0.23	0.29	0.52
GER CHM			
PT USA	0.37	**0.24**	**0.67**
FT USA TLN	**0.45**	0.21	0.66
FT USA VNA	0.41	**0.24**	0.65
PT AFR	0.19	0.16	0.35
FT AFR VNA	0.21	0.18	0.39

AUS VNA	F1 basic	Dif.	F1 best
PT USA	0.5	**0.31**	**0.81**
FT USA TLN	**0.54**	0.27	**0.81**
FT USA VNA	0.53	0.27	0.8
PT AFR	0.43	0.23	0.66
FT AFR VNA	0.39	0.29	0.68
ESP GIR			
PT USA	0.34	**0.36**	**0.7**
FT USA TLN	**0.38**	0.27	0.65
FT USA VNA	0.34	0.33	0.67
PT AFR	0.21	0.27	0.48
FT AFR VNA	0.24	*0.26*	0.5
ESP BAR			
PT USA	0.11	**0.19**	**0.3**
FT USA TLN	**0.15**	0.06	0.21
FT USA VNA	0.1	0.14	0.24
PT AFR	0.02	0.02	0.04
FT AFR VNA	0.07	*0*	0.07
USA NY I			
PT USA	**0.66**	0.14	**0.8**
FT USA TLN	0.47	0.29	0.76
FT USA VNA	0.47	**0.3**	0.77
PT AFR	0.24	0.06	0.3
FT AFR VNA	0.2	0.2	0.4
USA NY II			
PT USA	**0.74**	0.02	**0.76**
FT USA TLN	0.56	0.17	0.73
FT USA VNA	0.64	0.06	0.7
PT AFR	0.21	0.16	0.37
FT AFR VNA	0.27	**0.26**	0.53

Table 6: Comparison of the achieved F1 scores before and after the application of post processing. The initial F1 scores without post-processing are depicted in the left columns ("F1 basic"). The highest F1 scores of each model after the application of post-processing are depicted in the right columns ("F1 best"). The highest F1 score can be achieved after the application of the dissolve boundaries, deletion of small objects or the increase of the confidence threshold post-processing steps, depending on the model at hand and the target areas. The highest F1 scores and the highest increase of the F1-scores are depicted in bold.

4.3 RESEARCH QUESTION R 3

The third research question focuses on the model´s performance variance within the individual target areas. It examines the impact of specific geographic features on the general model performance on the sub-region level. The research hypotheses H 3.1 and H 3.2 are tested on 23 selected micro-regions from the Tallinn, Vienna, Daugavpils, Chemnitz, Girona and the New York target areas. For this research experiment, the three best performing models (i.e. PT USA, FT USA TLN and FT USA

VNA) are utilized. Additionally, PT AFR is added to the research experiment as the negative benchmark.

The Tables 7-13 show the extraction results on relatively homogenous micro-regions predominantly consisting of apartment buildings (“AptBldg”), large apartment building blocks with flat rooftops (“AptBlock”), single family houses (“SF_House”), historical buildings (“Historical”) and areas with large industrial, commercial or public buildings (“ICP_Bldg”). Like in the previous research experiment, the same three post-processing methods were applied to examine the impact of those methods on the respective sub-regions.

On the Tallinn target area, all four tested models achieve the best results on the apartment building micro-region. The F1-scores of 0.9 and above after the application of post-processing can be considered as excellent results on previously unseen target areas. The performance on the single family house micro-region is likewise good with F1-scores of above 0.8 with the exception of PT AFR. Model performance on the apartment blocks and ICP buildings micro-regions with F1-scores of around 0.7, are likewise acceptable. All models achieve the lowest results on the historical centre of Tallinn. Post-processing was particularly successful on the single family house and ICP buildings micro-regions while it was least effective on the apartment blocks micro-region. Interestingly, FT USA VNA achieved after post-processing the best result on the historical centre of Tallinn although the Vienna training area for model fine-tuning does not encompass any historical buildings.

ESTONIA	Precision				Recall				F1				IoU			
TALLINN	Basic	DB	X25	T80	Basic	DB	X25	T80	Basic	DB	X25	T80	Basic	DB	X25	T80
AptBldg																
PT USA	0.93	0.93	0.92	**0.97**	0.93	0.94	0.95	0.92	0.92	**0.94**	**0.94**	**0.94**	0.86	0.89	0.88	**0.9**
FT USA TLN	0.59	0.65	0.86	0.9	0.94	0.94	0.94	0.94	0.73	0.77	0.9	0.92	0.57	0.63	0.82	0.85
FT USA VNA	0.61	0.69	0.85	0.89	0.92	0.95	**0.96**	**0.96**	0.73	0.8	0.9	0.92	0.58	0.67	0.82	0.86
PT AFR	0.66	0.76	0.77	0.93	0.82	0.82	0.81	0.49	0.73	0.79	0.79	0.64	0.58	0.65	0.66	0.47
AptBlock																
PT USA	0.72	0.77	0.77	**0.78**	0.69	0.75	0.75	0.73	0.71	**0.76**	**0.76**	**0.76**	0.55	0.62	0.62	0.61
FT USA TLN	0.5	0.54	0.71	0.75	0.76	0.81	0.8	0.76	0.6	0.65	0.75	0.75	0.43	0.48	0.6	0.6
FT USA VNA	0.44	0.52	0.75	0.76	0.75	0.8	**0.83**	0.79	0.55	0.64	0.79	0.77	0.38	0.47	**0.65**	0.63
PT AFR	0.5	0.58	0.64	0.6	0.53	0.56	0.56	0.13	0.51	0.57	0.6	0.22	0.34	0.4	0.43	0.12
SF_House																
PT USA	0.8	0.87	0.89	**0.94**	0.57	0.76	0.81	0.75	0.66	0.81	0.85	0.83	0.5	0.69	0.74	0.72
FT USA TLN	0.58	0.6	0.81	0.84	0.72	0.87	**0.87**	0.85	0.64	0.71	0.84	0.84	0.48	0.55	0.72	0.73
FT USA VNA	0.56	0.59	0.86	0.89	0.67	0.83	0.85	0.84	0.61	0.69	**0.86**	**0.86**	0.44	0.53	0.74	**0.76**
PT AFR	0.78	0.81	0.83	0.92	0.52	0.66	0.68	0.39	0.62	0.73	0.75	0.54	0.45	0.57	0.6	0.37
Historical																
PT USA	0.33	0.52	0.52	0.56	0.24	0.64	0.64	0.61	0.28	0.57	0.57	0.59	0.16	0.4	0.4	0.42
FT USA TLN	0.39	0.34	0.45	0.4	0.48	0.7	0.7	0.55	0.43	0.46	0.55	0.46	0.27	0.3	0.38	0.3
FT USA VNA	0.26	0.47	0.6	**0.7**	0.35	**0.73**	**0.73**	0.7	0.3	0.57	0.66	**0.7**	0.17	0.4	0.5	**0.54**
PT AFR	0.28	0.25	0.28	0.5	0.14	0.32	0.32	0.08	0.18	0.23	0.3	0.15	0.1	0.16	0.17	0.08
ICP_Bldg																
PT USA	0.62	0.77	0.79	**0.84**	0.67	0.81	0.85	0.78	0.65	0.79	**0.82**	0.81	0.48	0.65	**0.7**	0.68
FT USA TLN	0.39	0.49	0.67	0.76	0.73	**0.86**	0.85	0.85	0.51	0.62	0.75	0.8	0.34	0.45	0.43	0.67
FT USA VNA	0.36	0.47	0.67	0.73	0.72	0.83	0.81	0.79	0.48	0.6	0.73	0.76	0.35	0.45	0.58	0.61
PT AFR	0.3	0.5	0.56	0.63	0.38	0.38	0.37	0.17	0.33	0.43	0.44	0.28	0.2	0.28	0.28	0.16

Table 7: Building extraction results on the Tallinn target area. The highest F1 scores are depicted in bold.

On the Daugavpils target area, the models perform best on the apartment blocks micro-region. Relatively good results are achieved on the single family house micro-region. PT USA, FT USA TLN and FT USA VNA perform solid on the apartment buildings and ICP buildings micro-regions, while PT AFR fails to achieve F1-scores above 0.5 on those micro-regions.

LATVIA	Precision				Recall				F1				IoU			
DGV	Basic	DB	X25	T80	Basic	DB	X25	T80	Basic	DB	X25	T80	Basic	DB	X25	T80
AptBldg																
PT USA	0.6	0.75	0.76	**0.86**	0.47	**0.69**	**0.69**	0.68	0.53	0.72	0.72	**0.76**	0.36	0.56	0.57	**0.62**
FT USA TLN	0.5	0.67	0.77	0.78	0.54	0.68	0.68	0.65	0.52	0.68	0.72	0.71	0.35	0.51	0.57	0.55
FT USA VNA	0.42	0.61	0.72	0.75	0.59	**0.69**	**0.69**	0.66	0.49	0.65	0.71	0.7	0.33	0.46	0.55	0.54
PT AFR	0.29	0.46	0.53	0.62	0.24	0.31	0.32	0.12	0.26	0.37	0.4	0.2	0.15	0.23	0.25	0.11
AptBlock																
PT USA	0.64	0.68	0.68	0.72	0.72	0.71	0.71	0.69	0.68	0.69	0.7	0.71	0.51	0.53	0.54	0.55
FT USA TLN	0.61	0.66	0.74	**0.77**	**0.85**	0.84	0.83	0.82	0.71	0.74	0.78	**0.8**	0.55	0.58	0.65	**0.66**
FT USA VNA	0.46	0.54	0.75	0.76	0.84	0.79	0.79	0.78	0.6	0.64	0.77	0.77	0.43	0.47	0.62	0.63
PT AFR	0.42	0.52	0.56	0.65	0.53	0.52	0.53	0.22	0.47	0.52	0.54	0.33	0.3	0.35	0.37	0.2
SF_House																
PT USA	0.64	0.63	0.65	0.72	0.49	0.49	0.58	0.49	0.56	0.55	0.61	0.58	0.39	0.38	0.44	0.41
FT USA TLN	0.67	0.67	0.74	0.83	0.76	0.73	0.77	0.61	0.71	0.7	0.75	0.7	0.55	0.54	0.6	0.55
FT USA VNA	0.59	0.6	0.81	**0.84**	**0.78**	0.77	0.75	0.72	0.67	0.67	0.78	**0.77**	0.51	0.51	**0.63**	**0.63**
PT AFR	0.72	0.73	0.75	0.83	0.44	0.42	0.49	0.14	0.55	0.53	0.59	0.25	0.38	0.36	0.42	0.14
ICP_Bldg																
PT USA	0.36	0.55	0.58	**0.64**	0.72	0.78	0.78	0.75	0.48	0.65	0.67	0.69	0.32	0.48	0.5	0.53
FT USA TLN	0.37	0.45	0.56	0.63	**0.81**	0.78	0.76	0.79	0.51	0.57	0.64	**0.7**	0.35	0.4	0.48	**0.54**
FT USA VNA	0.24	0.34	0.46	0.52	0.7	0.75	0.73	0.73	0.36	0.47	0.56	0.61	0.22	0.31	0.39	0.43
PT AFR	0.29	0.38	0.43	0.62	0.49	0.51	0.53	0.28	0.36	0.43	0.47	0.38	0.22	0.28	0.31	0.24

Table 8: Building extraction results on the Daugavpils target area.

In Vienna, good results are achieved for the apartment blocks and single family house micro-regions, followed by relatively good results for the IPC region. In contrast, all models performed much weaker on the apartment buildings region with scores ranging from 0.61 to 0.64.

:

AUSTRIA	Precision				Recall				F1				IoU			
VIENNA	Basic	DB	X25	T80	Basic	DB	X25	T80	Basic	DB	X25	T80	Basic	DB	X25	T80
AptBldg																
PT USA	0.52	0.66	0.66	**0.7**	0.37	0.61	0.63	0.63	0.3	0.57	0.64	**0.66**	0.2	0.47	0.48	**0.5**
FT USA TLN	0.23	0.52	0.62	0.58	0.45	0.58	0.6	0.55	0.31	0.55	0.61	0.56	0.18	0.38	0.44	0.39
FT USA VNA	0.23	0.52	0.64	0.66	0.45	**0.64**	0.63	0.63	0.3	0.57	0.64	0.64	0.18	0.4	0.47	0.48
PT AFR	0.25	0.35	0.4	0.5	0.32	0.41	0.42	0.07	0.28	0.3	0.41	0.13	0.16	0.23	0.25	0.07
AptBlock																
PT USA	0.74	0.84	0.85	**0.87**	0.51	0.7	0.83	0.78	0.6	0.77	**0.84**	0.82	0.43	0.62	**0.73**	0.7
FT USA TLN	0.58	0.69	0.81	0.83	0.64	0.81	**0.86**	0.83	0.61	0.74	**0.84**	0.83	0.44	0.59	0.72	0.71
FT USA VNA	0.58	0.62	0.81	0.83	0.64	0.82	0.84	0.83	0.61	0.71	0.82	0.83	0.44	0.55	0.7	0.72
PT AFR	0.59	0.64	0.68	0.78	0.43	0.56	0.66	0.27	0.5	0.6	0.67	0.4	0.33	0.43	0.51	0.25
SF_House																
PT USA	0.81	0.87	0.88	**0.92**	0.5	0.59	0.89	0.83	0.62	0.7	0.88	0.87	0.44	0.54	0.79	0.77
FT USA TLN	0.64	0.67	0.81	0.87	0.68	0.76	**0.9**	0.89	0.66	0.71	0.85	0.88	0.5	0.56	0.74	0.79
FT USA VNA	0.61	0.62	0.88	0.89	0.72	0.78	0.91	0.89	0.66	0.69	**0.89**	**0.89**	0.49	0.53	**0.81**	**0.81**
PT AFR	0.79	0.81	0.82	0.93	0.46	0.53	0.77	0.45	0.58	0.64	0.79	0.61	0.41	0.47	0.68	0.44
ICP_Bldg																
PT USA	0.7	0.79	0.8	**0.83**	0.62	0.68	0.8	0.74	0.66	0.73	0.8	0.78	0.49	0.57	0.67	0.65
FT USA TLN	0.5	0.58	0.71	0.79	0.74	0.77	0.8	**0.82**	0.6	0.66	0.75	0.8	0.43	0.49	0.6	0.67
FT USA VNA	0.36	0.43	0.72	0.8	0.73	0.8	**0.82**	0.81	0.48	0.56	0.76	**0.81**	0.32	0.39	0.62	**0.68**
PT AFR	0.47	0.53	0.56	0.76	0.46	0.52	0.6	0.38	0.46	0.53	0.58	0.51	0.3	0.35	0.41	0.34

Table 9: Building extraction results on the Vienna target area.

The Chemnitz target area contains large homogeneous areas covered by apartment buildings. Therefore, two micro regions with very similar geographic characteristics were selected to compare model performance variance within the same target area. On the first apartment buildings micro region, FT USA TLN and FT USA VNA perform relatively well and achieve a F1-score of 0.69. On the second apartment buildings micro-region, all models show significantly lower results. PT USA and its finetuned variants perform solid on the ICP buildings micro-region meanwhile PT AFR fails to achieve an acceptable result on any of the examined micro-regions.

GERMANY	Precision				Recall				F1				IoU			
CHEMNITZ	Basic	DB	X25	T80	Basic	DB	X25	T80	Basic	DB	X25	T80	Basic	DB	X25	T80
AptBldg I																
PT USA	0.31	0.6	0.66	**0.72**	0.3	0.62	0.61	0.57	0.31	0.61	0.63	0.64	0.18	0.44	0.47	0.47
FT USA TLN	0.3	0.5	0.66	0.64	0.42	**0.74**	0.72	0.65	0.35	0.6	**0.69**	0.65	0.21	0.43	**0.53**	0.47
FT USA VNA	0.31	0.48	0.64	**0.72**	0.39	0.65	0.63	0.66	0.35	0.55	0.63	**0.69**	0.21	0.38	0.46	**0.53**
PT AFR	0.12	0.29	0.33	0.37	0.14	0.33	0.32	0.11	0.13	0.31	0.33	0.17	0.07	0.18	0.19	0.09
AptBldg II																
PT USA	0.29	0.55	0.64	**0.67**	0.24	0.58	0.6	0.53	0.26	0.56	**0.62**	0.59	0.15	0.39	**0.45**	0.42
FT USA TLN	0.38	0.34	0.45	0.43	0.54	0.68	**0.69**	0.55	0.45	0.45	0.55	0.48	0.29	0.29	0.38	0.32
FT USA VNA	0.3	0.39	0.57	0.5	0.38	0.63	0.64	0.56	0.33	0.48	0.6	0.53	0.2	0.32	0.43	0.36
PT AFR	0.14	0.22	0.27	0.23	0.1	0.16	0.17	0.08	0.11	0.19	0.21	0.12	0.06	0.1	0.12	0.06
ICP_Bldg																
PT USA	0.41	0.52	0.57	**0.59**	0.57	0.49	0.6	0.39	0.48	0.5	0.58	0.47	0.31	0.34	0.41	0.31
FT USA TLN	0.31	0.39	0.51	0.57	0.76	0.78	0.79	**0.8**	0.44	0.52	0.62	**0.67**	0.28	0.35	0.45	**0.5**
FT USA VNA	0.29	0.35	0.47	0.55	0.7	0.64	0.7	0.63	0.41	0.46	0.56	0.59	0.25	0.3	0.39	0.42
PT AFR	0.22	0.36	0.38	0.38	0.35	0.34	0.43	0.17	0.27	0.35	0.4	0.23	0.16	0.21	0.25	0.13

Table 10: Building extraction results on the Chemnitz target area.

The already observed weak performance of all examined building extraction models on the Girona target could be caused by the densely build up areas with its characteristic Catalonian apartment building blocks. This is approved by the micro-region research experiment. None of the models achieve a F1-score above 0.37 on the apartment buildings. On the other hand, all four models perform extraordinary well on the single family house micro-region within the Girona target area. The difference between the highest F1-score on the apartment buildings micro region (i.e. 0.37) and the best result on the single house micro region (i.e. 0.87) is large. PT USA, FT USA TLN and FT USA VNA achieve also good results on the ICP buildings micro-region. Figure 31 visualizes the observed performance difference between the apartment buildings and single family houses micro-regions within the Girona target area.

SPAIN	Precision				Recall				F1				IoU			
GIRONA	**Basic**	**DB**	**X25**	**T80**	**Basic**	**DB**	**X25**	**T80**	**Basic**	**DB**	**X25**	**T80**	**Basic**	**DB**	**X25**	**T80**
AptBldg																
PT USA	0.3	0.5	0.52	0.62	0.13	0.27	0.28	0.24	0.18	0.35	**0.37**	0.35	0.1	0.21	**0.4**	0.21
FT USA TLN	0.35	0.3	0.38	0.41	0.24	0.32	**0.33**	0.26	0.29	0.31	0.36	0.32	0.17	0.18	0.22	0.19
FT USA VNA	0.21	0.27	0.41	0.42	0.25	0.37	0.33	0.32	0.23	0.29	0.36	0.36	0.13	0.17	0.22	0.22
PT AFR	0.22	0.3	0.32	**0.66**	0.05	0.12	0.13	0.03	0.08	0.18	0.19	0.07	0.04	0.09	0.1	0.03
SF_House																
PT USA	0.77	0.88	0.88	0.9	0.56	0.76	0.88	0.83	0.65	0.81	0.88	0.86	0.47	0.69	**0.79**	0.76
FT USA TLN	0.57	0.61	0.73	0.85	0.6	0.83	0.88	0.87	0.59	0.71	0.8	0.86	0.41	0.55	0.67	0.76
FT USA VNA	0.54	0.66	0.81	0.85	0.62	0.87	**0.92**	0.89	0.58	0.75	0.86	**0.87**	0.4	0.6	0.76	0.77
PT AFR	0.72	0.86	0.87	**0.97**	0.44	0.69	0.8	0.45	0.55	0.76	0.83	0.61	0.38	0.62	0.72	0.44
ICP_Bldg																
PT USA	0.35	0.7	0.79	0.8	0.39	0.65	0.74	0.61	0.37	0.67	0.76	0.69	0.23	0.51	**0.62**	0.53
FT USA TLN	0.42	0.64	0.77	**0.82**	0.51	0.77	**0.81**	0.71	0.46	0.7	**0.79**	0.76	0.3	0.54	0.6	**0.62**
FT USA VNA	0.29	0.42	0.66	0.76	0.48	0.7	0.76	0.73	0.36	0.53	0.71	0.74	0.22	0.36	0.55	0.59
PT AFR	0.29	0.42	0.46	0.46	0.24	0.29	0.33	0.18	0.26	0.34	0.39	0.26	0.15	0.21	0.24	0.15

Table 11: Building extraction results on the Girona target area.

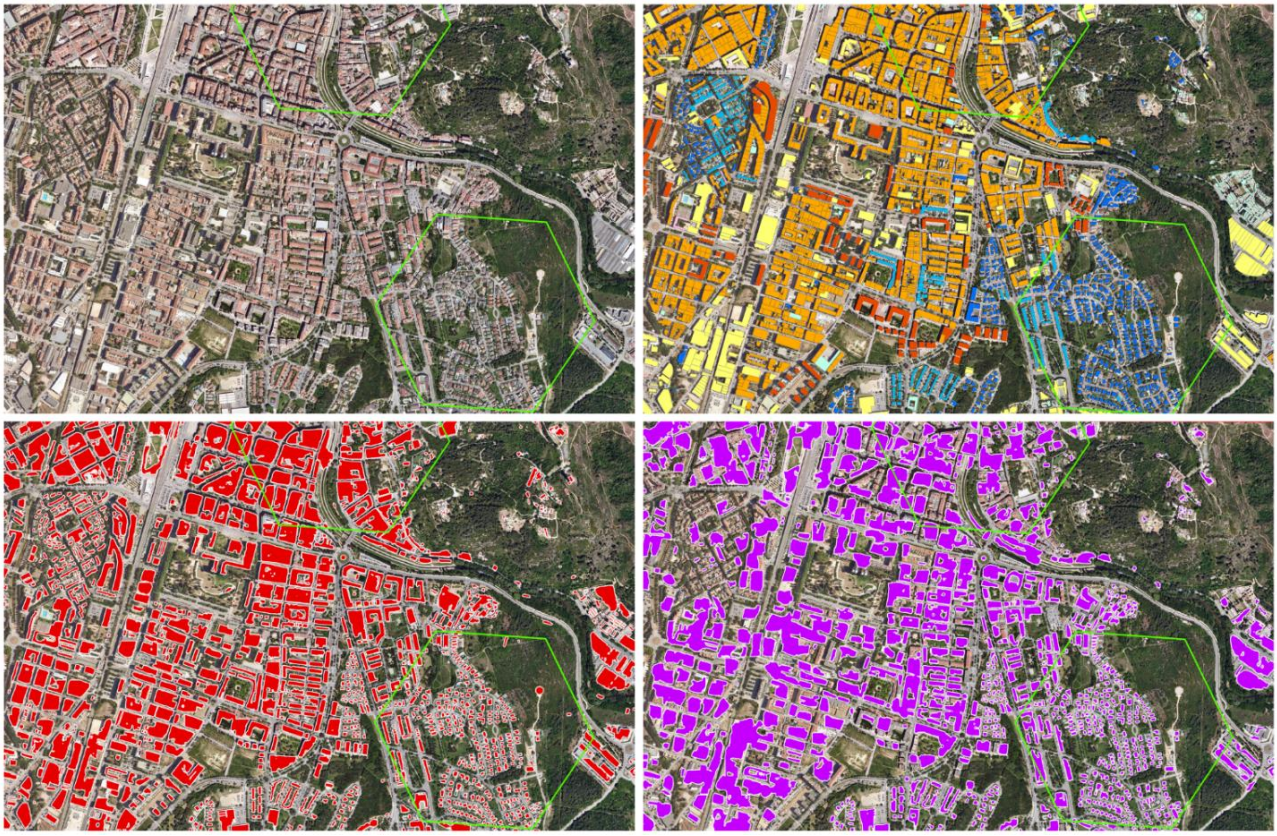

Figure 31: Building extraction results on the Girona target area. PT USA (red) and PT AFR (purple). Both models performed well on the single house micro-region located in the south-east of the target area (highlighted green hexagon). Both models failed to achieve good results on the apartment buildings in the northern part of the Girona target area (highlighted green hexagon). The top row provides the ground reference.

Remarkable performance variance exists also within the New York I target area. PT USA and its finetuned variants achieve F1-scores of above 0.9 on the single family house micro-regions. The results on the apartment blocks micro-region are around 30% lower. Performance differences exist also between the apartment blocks micro-regions located within the two New York target areas. PT USA achieves after the

application of post-processing a significantly better result on the New York II apartment blocks micro-region meanwhile PT AFR performs much worse in New York II in comparison to the apartment blocks micro-region in New York I. The performance on the ICP buildings in New York II is solid with the exception of PT AFR.

USA	Precision				Recall				F1				IoU			
NY I	**Basic**	**DB**	**X25**	**T80**	**Basic**	**DB**	**X25**	**T80**	**Basic**	**DB**	**X25**	**T80**	**Basic**	**DB**	**X25**	**T80**
AptBlock																
PT USA	0.58	0.65	0.7	**0.74**	0.7	0.65	0.64	0.61	0.64	0.65	**0.67**	**0.67**	0.47	0.48	0.5	**0.51**
FT USA TLN	0.4	0.4	0.58	0.72	0.61	0.53	0.52	0.52	0.48	0.46	0.55	0.61	0.32	0.3	0.38	0.43
FT USA VNA	0.36	0.38	0.63	0.61	**0.72**	0.71	0.7	0.67	0.48	0.5	0.67	0.64	0.31	0.33	0.5	0.47
PT AFR	0.22	0.16	0.23	0.29	0.28	0.16	0.16	0.04	0.25	0.16	0.19	0.07	0.14	0.09	0.1	0.04
SF_House																
PT USA	0.8	0.86	0.94	**0.97**	**0.92**	0.89	**0.92**	0.87	0.85	0.88	**0.93**	0.92	0.74	0.78	**0.87**	0.86
FT USA TLN	0.52	0.61	0.87	0.93	0.91	0.85	0.89	0.89	0.66	0.71	0.88	0.91	0.5	0.55	0.78	0.83
FT USA VNA	0.47	0.55	0.88	**0.97**	0.89	0.87	0.87	0.86	0.61	0.68	0.88	0.91	0.44	0.51	0.79	0.84
PT AFR	0.51	0.53	0.58	0.75	0.45	0.4	0.42	0.06	0.48	0.46	0.49	0.12	0.31	0.3	0.32	0.06

Table 12: Building extraction results on the New York I target area.

USA	Precision				Recall				F1				IoU			
NY II	**Basic**	**DB**	**X25**	**T80**	**Basic**	**DB**	**X25**	**T80**	**Basic**	**DB**	**X25**	**T80**	**Basic**	**DB**	**X25**	**T80**
AptBlock																
PT USA	0.22	0.49	**0.63**	0.56	0.76	0.87	0.89	0.82	0.34	0.63	**0.74**	0.66	0.2	0.46	**0.59**	0.5
FT USA TLN	0.11	0.21	0.41	0.38	0.67	0.8	0.79	0.69	0.19	0.34	0.54	0.49	0.1	0.2	0.37	0.32
FT USA VNA	0.1	0.2	0.53	0.43	0.74	**0.92**	**0.92**	0.74	0.17	0.33	0.67	0.54	0.09	0.19	0.51	0.37
PT AFR	0.02	0.03	0.04	0.18	0.13	0.15	0.15	0.05	0.04	0.06	0.07	0.08	0.02	0.03	0.04	0.04
ICP_Bldg																
PT USA	0.49	0.55	0.63	**0.76**	0.62	0.64	0.75	0.67	0.55	0.59	0.62	**0.72**	0.38	0.42	0.53	**0.56**
FT USA TLN	0.28	0.32	0.57	0.64	0.69	0.73	0.75	0.73	0.4	0.45	0.65	0.68	0.25	0.29	0.48	0.52
FT USA VNA	0.23	0.25	0.55	0.59	0.73	0.78	**0.79**	0.75	0.35	0.38	0.65	0.66	0.21	0.23	0.48	0.5
PT AFR	0.27	0.36	0.41	0.48	0.35	0.39	0.46	0.09	0.31	0.37	0.43	0.16	0.18	0.23	0.28	0.08

Table 13: Building extraction results on the New York II target area.

5. DISCUSSION

The final step in the automated building extraction process is the interpretation of the results and the subsequent evaluation of the utilized building extraction models. This step is essential in order to assess, if the model at hand fits the purpose or, in other words, if the model performance is acceptable for real world applications.

5.1 ANSWERING THE RESEARCH QUESTIONS AND HYPOTHESES

In the following sections, the results of the research experiments are discussed along the formulated research questions and hypotheses.

5.1.1 RESEARCH QUESTION R 1

Two research experiments were conducted to examine the geographical transferability and cross-regional performance of two Mask R-CNN building extraction models.

Table 14 summarizes the results of the first and second research experiment. The five examined models are ordered by the achieved F1-scores. Results contradicting the research hypotheses are marked red or, in case of contradicting only H 1.1, in pink. The results which are in accordance with H1.1 and H.1.2 are highlighted in green.

EST TLN	F1	EST NRV	F1	LVA DGV	F1	AUS VNA	F1	GER CHM	F1
USA PT	0.64	PT USA	0.63	FT USA TLN	0.62	FT USA TLN	0.54	FT USA TLN	0.45
FT USA TLN	0.57	FT USA VNA	0.48	PT USA	0.61	FT USA VNA	0.53	FT USA VNA	0.41
FT USA VNA	0.56	PT AFR	0.47	FT USA VNA	0.57	PT USA	0.5	PT USA	0.37
PT AFR	0.39	FT USA TLNA	0.46	FT AFR VNA	0.47	FT AFR VNA	0.43	FT AFR VNA	0.21
FT AFR VNA	0.39	FT AFR VNA	0.28	PT AFR	0.46	PT AFR	0.39	PT AFR	0.19
SLK KOS	**F1**	**ESP GIR**	**F1**	**ESP BAR**	**F1**	**USA NY I**	**F1**	**USA NY II**	**F1**
FT USA TLN	0.39	FT USA TLN	0.38	FT USA TLN	0.15	PT USA	0.66	PT USA	0.74
FT USA VNA	0.38	FT USA VNA	0.34	PT USA	0.11	FT USA TLN	0.47	FT USA VNA	0.64
PT USA	0.37	PT USA	0.34	PT USA VNA	0.1	FT USA VNA	0.47	FT USA TLN	0.56
FT AFR VNA	0.23	FT AFR VNA	0.24	FT AFR VNA	0.07	PT AFR	0.24	FT AFR VNA	0.27
PT AFR	0.17	PT AFR	0.21	PT AFR	0.02	FT AFR VNA	0.2	PT AFR	0.21

Table 14. Initial F1 scores on all target areas.

Research hypothesis H 1.1 is confirmed by the results of PT USA on the two New York target areas. The significantly better performance of PT USA in comparison to PT AFR on all ten target areas is in accordance with H 1.2. The assumed higher geographical closeness between the model´s training areas in the United States and the European target areas in comparison to the training areas located in Africa, is reflected in the extraction results. The weak performance of PT USA on the Girona and Barcelona

target areas is likewise in accordance with H 1.2 due to the lower geographical closeness between those two Mediterranean locations and the training areas of PT USA. The performance of PT USA on the two New York target areas likewise confirms H 1.2. It was expected that PT USA achieves the best results on the target areas which are located in the United States. However, none of the research hypotheses explain the low performance of PT USA on the Kosice, Chemnitz and Vienna target areas in comparison to the acceptable results on the Tallinn, Narva and Daugavpils target areas.

The fine-tuning of the pretrained models on two European training areas in Tallinn and Vienna led to inconsistent results. The expected performance increase occurred only on some of the European target areas. The surprising decrease of the F1 score on the Tallinn target areas after fine-tuning PT USA, could be interpreted as a model overfitting problem. However, the two fine-tuned versions of PT USA enhanced the building extraction results on most of the other European target areas. A closer visual examination of the extraction results of FT USA TLN and FT USA VNA reveals that model fine-tuning did not caused an increase of arbitrary errors but it raised the model´s sensitivity. The fine-tuning enabled the models to capture significantly more details on the satellite images and to extract more buildings precisely. This leads to a significant increase of true positives on all European target areas in full accordance with H 1.1 and H 1.2. However, the higher sensitivity of the fine-tuned models results in an additional extraction of numerous building-like objects like small sheds, bus stations or plastic greenhouses within the target areas. The ground reference datasets do not list such objects as buildings. Consequently, the fine-tuned models produce significantly more false positives. This is particularly fatal on the Narva target areas as shown by Figure 16. On other target areas, the increase of true positives and the decrease of false negatives outbalances the rising number of false positives which leads to an overall increase of the F1 scores in comparison to PT USA.

The formulated research hypotheses can only partially explain the better performance of FT USA TLN on most of the target areas in comparison to FT USA VNA. A larger geographical proximity exists between the Tallinn training area and the Tallinn and Daugavpils target areas. Additionally, a larger geographical closeness can be assumed between Tallinn and the Chemnitz and Kosice target areas by sharing a common communist architectural history. However, the results on the Girona and Barcelona target areas contradict H 1.1. The higher F1 score of FT USA VNA on the Narva target area and the better result of FT USA TLN on the Vienna target area likewise contradict H 1.1 and H 1.2.

The significantly lower performance of all five examined models on the Chemnitz, Kosice, Girona and Barcelona target areas, cannot be explained by the formulated

research hypotheses H 1.1 and H 1.2. However, the geographical characteristics of those four target areas could be an explanation for the poor results.

Although the discussed findings of the first and second research experiment partially refute the H 1.1 and H 1.2, the results also demonstrate that geographical closeness is relevant for the model´s extraction performance. PT AFR continuously underperforms PT USA on all examined target areas. This validates H 1.2. Secondly, the fine-tuning of PT AFR on the Vienna training area improved the performance on most but not all target areas. However, the performance gap to PT USA and its two fine-tuned variants remains large. Hence, the research findings point to the assumption that low geographical closeness between the training and target areas deteriorates the model´s performance meanwhile minor differences in the geographic characteristics between the training and target areas like between Tallinn and Vienna, do not have a significant impact on the extraction results. Furthermore, the presented results clearly show that the geographic proximity between the model´s training and test area is not a strong indicator for model performance on new target areas.

5.1.2 RESEARCH QUESTION R 2

The third research experiment examined the impact of post-processing on the building extraction results. Three different post-processing methods were tested on all target areas to answer the second research question.

The experiment results reveal that the fine-tuned models, in general, benefit more from the applied post-processing methods. All three executed post-processing rounds aim to reduce the number of false positives and false negatives. The higher sensitivity of the fine-tuned model variants leads to an increase of correctly predicted buildings, a reduction of missed buildings (i.e. false negatives) and, simultaneously, to a significant increase of false positives. Consequently, the reduction of the false positives has a large effect on the results of the three fine-tuned models (see Appendix 4).

The focus of this research is on the large cross-regional variance of the positive impact of post-processing, which is shown by the Figures 28-30. The increase of the F1 scores through the very same post-processing methods varies throughout the different target areas significantly. Table 15 reveals some general spatial patterns. Post-processing is particularly effective on the Girona, Vienna, Kosice, Narva and Chemnitz target areas. On those target areas, the F1 scores of all models improved significantly. The effect of post-processing remains particularly low on the Daugavpils; Barcelona and New York II target areas. Post-processing seems to be particularly efficient on target areas with initially lower model performance. The Barcelona target area is here an exception. The initial average F1 score of 0.1 is only marginally improved through the application of post-processing.

The interpretation of the observed cross-regional variance is challenging. The chosen research setup aims to reduce the number of possible impact factors on the model´s extraction results by testing the very same building extraction models on various satellite imagery with very similar technical parameters. The utilized ground reference datasets are one of the remaining variables of the presented experiment. On the Kosice target area, for example, every individual garage is listed as separate building. The over 4000 small adjacent objects challenged the building extraction capabilities of each examined model enormously. In some of the ground reference datasets, the apartment buildings are sub-divided into multiple buildings which are only partially distinguishable on the satellite imagery. However, variance in the ground reference datasets cannot explain the difference in post-processing efficiency between the Tallinn and Narva target areas or between the Girona and Daugavpils target areas.

Differences in the target areas geographic characteristics are another experiment variable that can impact the efficiency of the applied post-processing methods. To answer the second research question, the positive impact on the F1 scores of each post-processing method has to be analysed separately. The examination of the Figures 28-30 in combination with Table 2 and the visual inspection of the satellite images do not provide an adamant correlation between the frequency of certain building types on the one hand and the building extraction performance and post-processing efficiently on the other hand. However, some tendencies are recognisable. The dissolve boundaries post-processing method works particularly well on target areas with a large number of apartment buildings, terraced houses and garages. This applies for the Girona, Kosice, Chemnitz and Tallinn target areas. The examined areas in Narva, Daugavpils and New York have comparatively less apartment buildings and the efficiency of the first post-processing method is likewise low on those target areas. The deletion of small objects was especially beneficial on the Narva and Vienna target areas. Significant improvement can be also seen on the Tallinn, Daugavpils, Chemnitz and the two New York target areas. These results correspond largely with the occurrence of dwellings in the respective target areas.

The inspection of the achieved F1 scores provide valuable information about the efficiency of the specific post-processing methods on locations with particular geographic characteristics. However, a closer examination of the results is sometimes necessary to further disclose the interrelation between the target area´s geographic characteristics, model performance and the efficiency of post-processing methods. Table 15 and Figure 15, for example, show only a low positive impact of the post-processing on the New York II target area. Table 2, on the other hand, reveals that terraced houses constitute 1/3 of the buildings within the New York II target area. The high efficiency of the dissolve boundaries post-processing method on the extraction of terraced houses should have resulted in a significant increase of the F1 scores.

A closer examination of the extraction results (see Appendix 4) and the satellite image reveals that the initial extraction of the terraced houses worked relatively well on the New York II target area in contrary to the Girona, Tallinn or Vienna target areas. The dissolvement of boundaries can efficiently reduce the number of false negatives but it eliminates simultaneously the correctly extracted individual terraced houses (i.e. true positives). This example demonstrates, that the same post-processing method can be beneficial on some geographic locations meanwhile it can even deteriorate the extraction results on other target areas.

Target Area	min.	max.	avg.	F1 avg. (basic)	F1 avg. (best)
ESTONIA TALLINN	0.08	0.27	0.2	0.51	0.71
ESTONIA NARVA	0.14	0.32	0.23	0.45	0.68
LATVIA DAUGAVPILS	0.07	0.21	0.13	0.53	0.67
AUSTRIA VIENNA	0.23	0.31	0.26	0.48	0.75
GERMANY CHEMNITZ	0.16	0.24	0.21	0.33	0.55
SLOVAKIA KOSICE	0.15	0.32	0.25	0.32	0.58
SPAIN GIRONA	0.26	0.36	0.29	0.3	0.6
SPAIN BARCELONA	0	0.19	0.06	0.1	0.17
USA NEW YORK I	0.06	0.3	0.2	0.41	0.61
USA NEW YORK II	0.02	0.26	0.11	0.5	0.62

Table 15: Increase of the F1 scores after the application of post-processing The lowest achieved increase by a model is depicted in the left column ("min."). The highest achieved increase of the F1 score by a model on the individual target areas is depicted in the second column ("max."). The average of the max. increase scores of all five examined models on the individual target areas is shown in the third column ("avg."). Additionally, the average F1 scores of the five models ("F1 avg. basic") and the average of the highest F1 scores after post-processing ("F1 avg. best") of all five models on the individual target areas are depicted in the fourth and fifth columns .

In summary, the presented research findings support research hypothesis H 2 in general. Specific post-processing methods are particularly efficient on locations with certain geographic characteristics. However, the model architecture, its general performance on individual target areas and the utilized ground reference datasets can also affect the efficiency of post-processing.

After the application of post-processing on the examined building extraction models, the performance on each target area can be re-examined to answer the first research question more elaborately. Table 16 provides the highest achieved F1 scores after the application of the discussed post-processing methods. Although the results of the three best performing models are close together on most of the target areas, the number of red fields signalling the falsification of the formulated research hypotheses H 1.1 and H 1.2, further increased. Post-processing closed the gap between PT USA and its fine-tuned variants significantly. However, PT USA provides the best results on most of the

target areas. FT USA VNA benefits more from the applied post-processing methods and overtakes FT USA TLN on seven target areas. However, the performance difference is so small that a clear explanation cannot be provided. FT AFR VNA benefits significantly more from post-processing than its pre-trained model variant. However, the result gap between FT AFR VNA and PT USA and its two fine-tuned variants, remains large.

The execution of three post-processing methods on the building extraction results generally affirms that the geographic proximity between the training and target areas is not a strong indicator for the assessment of cross-regional model performance. Particularly low geographical closeness between the training and target areas seems to have a significant negative impact on the model´s results. However, the geographic transferability between Central- and Eastern Europe and the United States can be considered successful with the application of the presented post-processing methods.

EST TLN	F1	EST NRV	F1	LVA DGV	F1	AUS VNA	F1	GER CHM	F1
FT USA VNA	0.81	PT USA	0.77	FT USA VNA	0.75	FT USA TLN	0.81	PT USA	0.67
PT USA	0.81	FT USA VNA	0.76	FT USA TLN	0.74	PT USA	0.81	FT USA TLN	0.66
FT USA TLN	0.8	FT USA TLN	0.73	PT USA	0.7	FT USA VNA	0.8	FT USA VNA	0.65
FT AFR VNA	0.63	PT AFR	0.61	FT AFR VNA	0.62	FT AFR VNA	0.68	FT AFR VNA	0.39
PT AFR	0.47	FT AFR VNA	0.52	PT AFR	0.53	PT AFR	0.66	PT AFR	0.35
SLK KOS	**F1**	**ESP GIR**	**F1**	**ESP BAR**	**F1**	**USA NY I**	**F1**	**USA NY II**	**F1**
PT USA	0.69	PT USA	0.7	PT USA	0.3	PT USA	0.8	PT USA	0.76
FT USA VNA	0.68	FT USA VNA	0.67	FT USA VNA	0.24	FT USA VNA	0.77	FT USA TLN	0.73
FT USA TLN	0.65	FT USA TLN	0.65	FT USA TLN	0.21	FT USA TLN	0.76	FT USA VNA	0.7
FT AFR VNA	0.44	FT AFR VNA	0.5	FT AFR VNA	0.07	FT AFR VNA	0.4	FT AFR VNA	0.53
PT AFR	0.32	PT AFR	0.48	PT AFR	0.04	PT AFR	0.37	PT AFR	0.37

Table 16: F1 scores after the application of post-processing.

5.1.3 RESEARCH QUESTION R 3

The third research question examined the impact of certain building types and urban structures on the building extraction model´s cross-regional performance. Figure 32 summarizes the results (i.e. F1 scores) on the various examined micro-regions within the target areas after the application of post-processing.

The highest performance average is achieved on the single house family buildings. Even PT AFR, the negative reference, performs on those micro-regions generally good by reaching F1 scores of above 0.7. PT USA, FT USA TLN and FT USA VNA achieve an average F1 scores above 0.7 on locations covered predominantly with large industrial, commercial and public buildings (i.e. ICP buildings). The results of those three models are likewise similarly well on the apartment block micro-regions. The performance of PT AFR on those micro-regions is less constant. On some target areas, the extraction of apartment blocks and ICP buildings is satisfactory meanwhile the

model fails to achieve acceptable results on other target areas with F1 scores below 0.5. All examined models show the weakest average performance on the apartment building micro-regions.

The experimental findings support research hypothesis H 3.1. The compact, detached houses with simple rectangular shapes enable remarkable high building extraction results. The challenges for the extraction task posed by small dwellings, garages, and other small building-like objects like awnings were resolved through the application of post-processing. The performance on the apartment blocks with regular shapes and large open spaces between the individual buildings resulted in satisfactory results on all examined target areas. PT AFR achieved acceptable results on some of the target areas but failed to extract apartment blocks on other locations satisfactorily. The application of post-processing was indispensable on the large commercial and industrial zones containing large buildings with irregular shapes and multiple roof types, (i.e. ICP bldg.) to achieve acceptable results. PT AFR shows a more constant performance on the ICP buildings micro-region in comparison to the performance on the apartment blocks. However, the extraction results are poor.

All models achieve the weakest performance on the apartment micro-regions. The average is elevated through the extremely high F1 scores achieved on the Tallinn target area. Urban districts with apartment buildings within the examined target areas are characterized by high building density, complex building shapes consisting of multiple adjacent buildings with various roof types and colours. The applied post-processing methods, particularly the dissolvement of boundaries, could improve the extremely poor initial extraction results to an acceptable level. However, the performance on the apartment building micro-regions remain the lowest in comparison to the other micro-regions. PT AFR generally fail, even with the application of post-processing, to extract apartment buildings satisfactorily.

The experiment results depicted by Figure 32 validate research hypothesis H 3.2 only partially. Model performance on single family house micro-regions is constantly high on all examined target areas with the exception of Daugavpils. The performance on the apartment buildings micro-regions tends to be lowest with the exception of the Tallinn and Daugavpils target areas. Model performance on the ICP buildings and apartment block micro-regions varies significantly between the different target areas. A constant performance level and performance sequel on those micro-regions throughout the different target areas is not recognizable.

The discussed research findings suggest that geographic characteristics like the predominance of certain building types have a significant impact on the building extraction results. However, local peculiarities can cause significant cross-regional performance variances. The results on the apartment building micro-regions provide a striking example. The performance of all models on the Tallinn target area excels the

results on the other locations by a large margin. A closer examination of the satellite images reveals that the apartment buildings within the Tallinn micro-region are regularly arranged and clearly separated from each other. The examined micro-regions within the Chemnitz, Vienna and Girona target areas are characterized, on the other hand, by densely build up urban areas with adjacent apartment buildings.

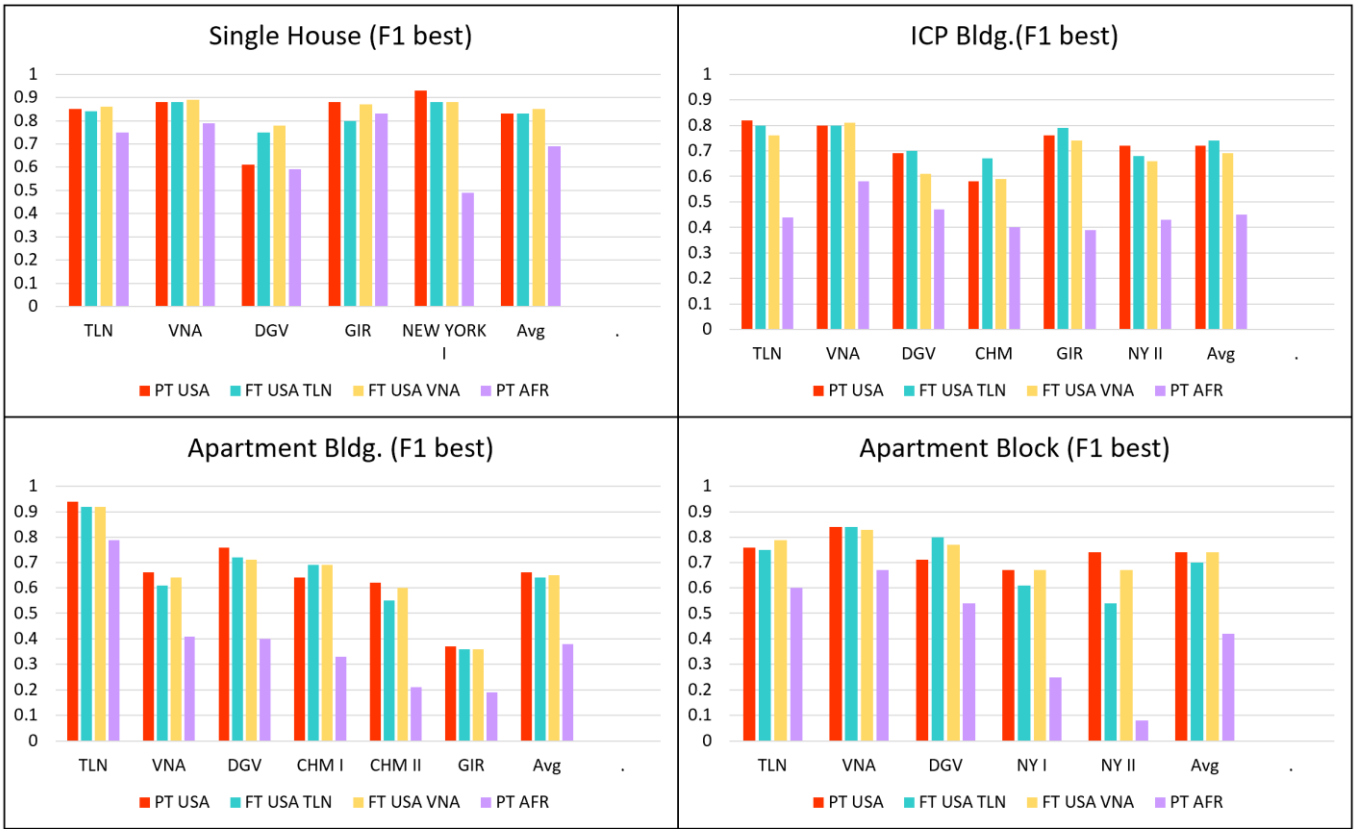

Figure 32: Summary of highest F1 scores after the application of post-processing on micro-regions.

5.2 RESULT COMPARISON WITH RELEVANT RESEARCH

A precise comparison of the presented research findings with the results of other studies, is challenging. Although numerous recent studies on deep learning based automated building extraction exist, different research setups reduce the comparability. Other studies examine different model architectures, use imagery with significantly lower or higher resolution, utilize other evaluation metrics or have different experiment approaches. Research on transfer learning and model performance examination on new, previously unseen target areas, often focuses on damaged building detection and extraction. Those results cannot be compared with the results of the general building extraction task. However, a few relatively similar studies on building extraction and transfer learning, are briefly discussed in the following section to place the obtained results and findings of this study in a broader research context. The Figures 33 and 34

provide a summary of the examined model´s results to ease the comparison with other studies.

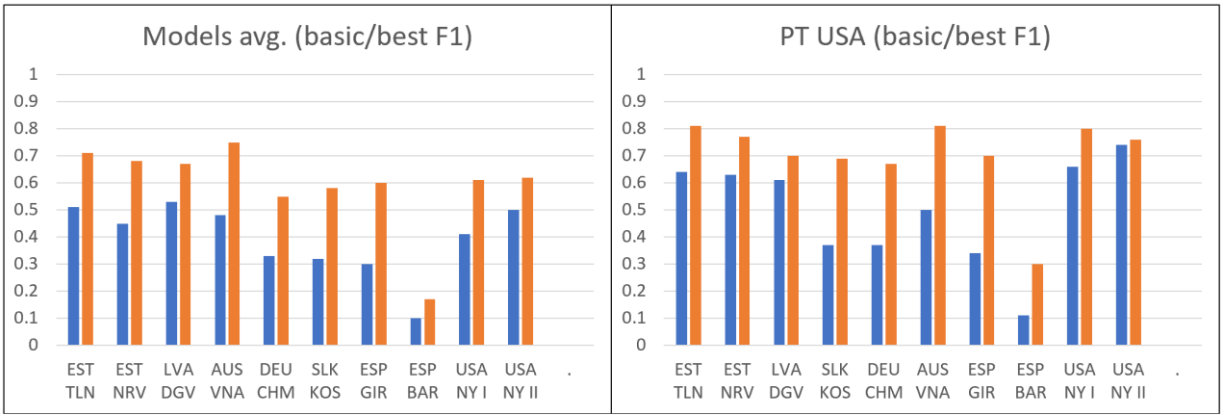

Figure 33: F1 scores before and after the application of post-processing. The initially achieved F1 scores are depicted in blue and the highest F1 scores after the application of post-processing are shown in red. On the left, the average F1 scores of all models are provided. On the right, the F1 scores of the best performing model, PT USA, are depicted.

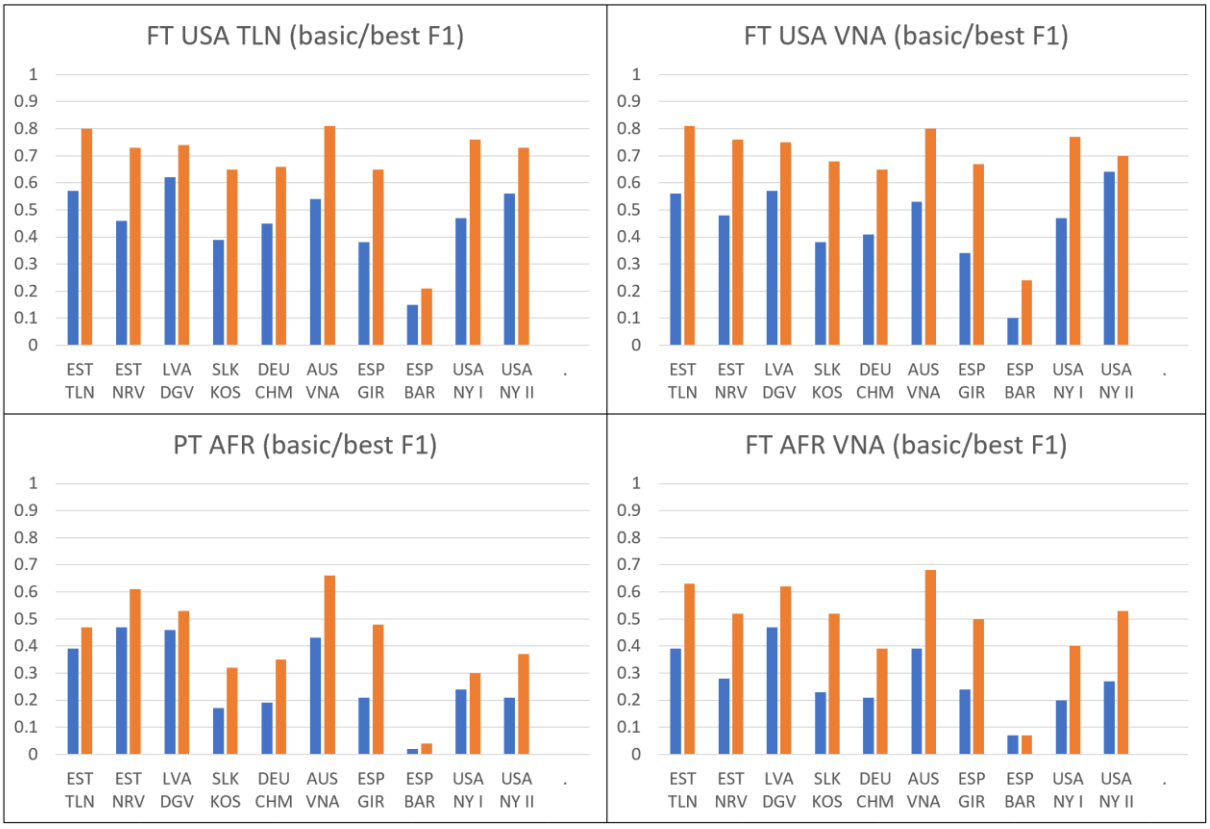

Figure 34: F1 scores before and after the application of post-processing of FT USA TLN, FT USA VNA, PT AFR and FT AFR VNA .The initially achieved F1 scores are depicted in blue and the highest F1 scores in red.

In April 2023, Sakeena et al. (2023) published a closely related research on the robustness and generalization ability of building footprint extraction, testing among others, a Mask R-CNN model. The model was examined on a large, high resolution (30 cm) dataset covering 24 different geographic locations across Europe and North America. The dataset covers, similarly to this research, multiple structurally different urban and suburban regions with various building types like residential and industrial buildings. The research aim was to derive guidelines for practical application in terms of model architecture, pre-training, model fine-tuning and transfer learning and to examine potential limitations of the employed building extraction models. Specific research questions cover the performance of models on previously unseen cities, the impact of transfer learning and performance consistency across different settlements and building structures.

Sakeena et al. (2023) report for the employed Mask R-CNN building extraction model F1 scores between 0.61 to 0.72 on new target areas within already seen cities. This category corresponds to the Tallinn and Vienna target areas of this research. The target areas are new for the models; however, they were fine-tuned on training areas within the same city. The reported F1 scores for new target areas within completely unseen cities range between 0.59 and 0.70. These scores are reached in this study only by PT USA on several target areas in the initial phase. FT USA VNA and FT USA TLN achieve an F1 score above 0.6 only on some target areas. However, after the application of post-processing, PT USA, FT USA VNA and FT USA TLN achieve regularly higher F1 scores on almost all target areas.

Interestingly, Sakeena et al. (2023) report, that the employed Mask R-CNN model detected buildings generally well but it mainly suffered from false positive detections. This reflects the findings of this research. Especially, the fine-tuned variants of the Mask R-CNN pre-trained models produced an enormous number of false positives. Furthermore, a Mask R-CNN model, trained only on North American training areas, was successfully employed on numerous European target areas. The authors conclude that Mask R-CNN models generalize well on data from the other continent. This affirms the strong performance of PT USA on most of the European target areas in this research.

Sakeena et al. (2023) experienced ambiguous results for the model fine-tuning. A performance improvement was not always achieved. They conclude that the effect of fine-tuning strongly depends on the employed model and the characteristics of the dataset. The results of this research confirm that model fine-tuning is a complex process with uncertain outcome. Further research on the impact of model architecture and the characteristics of the new target areas on the results of model fine-tuning, is required.

Finally, Sakeena et al. (2023) tested model performance also on smaller, more homogeneous areas predominantly covered by industrial buildings or multiple family buildings. The general finding is that the building structure has a strong impact on the model performance. The employed Mask R-CNN struggled particularly with large objects like industrial buildings. The applied model, which was trained specifically on large industrial buildings, achieved only an F1 score of 0.33. This pattern is reported by several studies and it is in line with the findings of this research, too. However, the relatively low performance on areas characterized by single family houses is surprising. The model, that was trained specifically on such buildings, achieved an F1 score of 0.47 on suburban areas predominantly covered by single family houses. In this research, all models achieve significantly higher results on the single house micro region, even in the initial phase without the application of post-processing.

Another closely related study was reported by Sawa et al. (2024). They used the Mask R-CNN building extraction model, which was utilized also in this research, in order to examine the effect of transfer learning on various, previously unseen, target areas. The pre-trained model achieved an F1 score of 0.71 on several target areas located in Ankara, Turkey. The target areas are characterized by a large building heterogeneity, comprised of high-rise buildings, old, densely build up residential areas, newly constructed medium sized buildings and numerous other buildings such as mosques, malls and large public buildings. Model fine-tuning was conducted separately on three different training areas with distinct geographical characteristics like high-rise buildings or densely build up areas. Sawa et al. (2024) report from successful model fine-tuning which elevated the results of the pre-trained model by around 20%. However, they utilized a specific calculation method for the accuracy measurement. which prevents a precise performance comparison.

Chen et al. (2022) report from the fine-tuning of a pre-trained Mask R-CNN building extraction model which was executed on previously unseen target areas in Japan. The model achieved an F1 score of 0.71 on urban areas and 0.75 on suburban regions. The higher density in urban built up areas posed a significant challenge for instance segmentation models like the employed Mask R-CNN. Similar experience was made in this research. The weakest performance was observed on densely built-up areas in Barcelona and Girona.

Li et al. (2019) observed for their building extraction model significant performance variances on different target areas. The U-Net based model achieved an F1 score of 0.89 on the Las Vegas target area, 0.75 in Paris, 0.62 in Shanghai and 0.54 in Khartoum. Model training and testing was conducted within the same cities. Applied post processing methods like the increase of the prediction confidence threshold from 0.45 to 0.55 and the removal of small objects resulted in slightly better results. Post-

processing was more beneficial for the results on the Shanghai and Khartoum target areas than on the Las Vegas and Paris target areas.

5.3 RESEARCH LIMITATIONS

In general, there is no research without limitations. This is particularly true for the deep learning and automated object detection and extraction research field. The goal of this research is to provide more insights into the impact of geography on the results of the automated building extraction. The examination of one influencing factor requires, ideally, the freezing of other potential factors. Such laboratory conditions are not realistic in the automated building extraction domain.

First of all, the testing on multiple target areas, located in different geographic regions, requires the utilization of different satellite imagery. Attention was paid in this research on similarity among the input imagery for the training and test areas, particularly regarding the spatial resolution, off-nadir angle, (no) cloud coverage, etc. However, it cannot be excluded that model performance variance is, partially, caused by sensor parameter variability of the input images. This is a widely discussed research challenge in the building extraction domain (Sakeena et al., 2023). In this research, significant model performance variance is recorded within the same country (i.e. Estonia) or even same cities (New York and Chemnitz), covered by the same satellite image sources. Thus, it is considered highly unlikely, that the recorded differences in model performance can be explained solely through sensor variance.

Second, this research is not using benchmarked reference labels. Model performance is assessed on governmental or open source (i.e. OpenStreetMap Project) building footprint labels. Although all reference datasets were manually reviewed and amended, errors like missing building labels or false building polygons can remain. However, errors in the benchmark reference labels exist, too (Bakiman et al., 2022). Furthermore, it is unlikely that the number of errors in the reference datasets are so high, that it significantly distorts the building extraction results. Finally, if significant distortions would occur, then all five examined models would be likewise affected.

Third, this research utilized relatively unsophisticated building extraction models. The majority of studies in the deep learning based object detection and extraction research domain focus on the model architectures. The main research goal is to improve existing models or to develop new models that outperform the existing ones. For this research, the two Mask R-CNN models, provided by the ESRI developer team, were used without any further improvements. Additionally, the model fine-tuning was conducted likewise relatively simple with a short training phase, using the default

settings and ESRI´s recommendations. In sum, the models examined for this research, are not considered as the highest standard. However, the aim of this research was to develop a sophisticated research setting to elaborately test cross-regional generalizability and transferability of automated building extraction models.

Fourth, this research tested five models sharing the same model architecture. Each of the models is trained or fine-tuned on different training areas located in various geographic regions. However, the location of the respective training areas is not the only variable. The training size differs between the fine-tuned versions. The size and characteristics of the training areas could be an additional impact factor on the model performance in addition to the geographic characteristics of the target areas.

6. CONCLUSION

The goal of this research was to get a deeper insight into the impact of geography on the deep learning based, automated building extraction. For elaborate testing, five Mask R-CNN models were tested on ten new, previously unseen, target areas within Europe and the United States. The model architecture is the same, however, the models are trained and fine-tuned on different regions.

Future users of AI-based building extraction models face an ever-growing number of available AI-models. The motivation for this research was to facilitate an elaborate assessment of the model´s expected performance for future applications. However, the presented research findings suggest that it is very difficult to estimate the geographical transferability and performance of automated building extraction models on new, previously unseen, target areas. Significant cross-regional performance variance could be observed for all five tested models in dependence of the utilized training areas, the target areas´ geographic characteristics and the applied post-processing methods. A model pre-testing on a small subset of the relevant new target area is likewise problematic due to the observed high performance differences within the very same geographic locations (i.e. cities). In summary, the presented research demonstrated the complexity and limited predictability of building extraction model´s cross-regional applicability and it calls for further research on that important topic.

6.1 MOST IMPORTANT RESEARCH FINDINGS

1) Acceptable building extraction results can be expected on new, unseen target areas, when the model was initially trained on locations with relatively similar geographic characteristics. Central- and Eastern European cities demonstrated in this research the required geographical closeness to the model training locations in the United States. The geographical proximity between the location of the fine-tuning training area and the target area is not decisive as long as the geographical closeness of the two locations is high enough. The results indicate that this is the case between Vienna and Tallinn. However, it is likely, that the sufficiency of geographical closeness for successful model fine-tuning, has to be determined case by case. Figure 35 provides some visual examples of the satisfactory building extraction results on new, previously unseen target areas.

2) The conducted model fine-tuning resulted in a higher sensitiveness of the examined pre-trained models. They extracted significantly more buildings correctly but, simultaneously, increased the number of false positives. The level of the recorded performance increase or decrease depends on the geographical characteristics and

the reference dataset characteristics of the individual target areas. Fine-tuned models regularly extracted additional small sheds and garages correctly but got penalized as false positives because those objects were not listed as buildings in the available reference datasets. Thus, the exploitation of the confusion matrixes alone is not sufficient to assess the model performance. A closer examination of the model predictions on the satellite imageries is indispensable.

3) Post-processing of the building extraction results significantly improves the outcome. However, the extent of improvement depends on the model characteristics as well as on the geographical characteristics of the individual target areas. Certain errors in the building extraction process occur particularly often on locations with specific geographical characteristics. Here, simple post-processing methods like the deletion of small, building-like objects, are particularly effective.

Post-processing is often double-edged. It can reduce the number of extraction errors only at the cost of eliminating valuable results, too. Thus, the use of post-processing should be chosen carefully, in dependence of the model at hand, the target areas and the users' requirements. This research highlighted the limits of post-processing. The large performance gap between PT USA and its fine-tuned variants on the one hand, and PT AFR on the other hand, could not be significantly reduced through the utilization of post-processing.

4) Significant model performance variance exists not only between different target areas but also within the target areas. All tested models, perform, for example, significantly better on single family house residential neighbourhoods than on areas characterized by apartment buildings. This finding underlines the importance of geographical heterogeneity within the target areas to ensure a profound testing of building extraction models.

Figure 35: Successful building extraction examples. FT USA VNA on the Tallinn target area is shown on the top (orange), FT USA TLN on the Vienna target area is shown in the middle (turquoise), PT USA on the New York II target area is shown on the bottom (red). Ground reference (yellow) is depicted on the left. Appendix 6 provides the images at a larger scale.

6.2 FUTURE RESEARCH

Deep learning-based, automated building extraction, and particularly, the transferability of established models on new, unseen, target areas is an active research field. (Sakeena et al., 2023) This research cannot answer all remaining questions, but it can contribute to a better understanding of the relevant impact factors through the development and implementation of a comprehensive building extraction model examination workflow and the subsequent testing of five research hypotheses.

It would be interesting to test the transferability of other building extraction models on the presented target areas in order to examine if similar impacts of the local geographic characteristics would occur. ESRI Inc. for example, also provides Mask R-CNN models trained on datasets from Saudi Arabia or Australia.

Another research expansion could be the addition of new target areas. Here, the impact of geographical closeness and proximity on model performance could be further investigated. Additional target areas within a close geographical distance to the already existing target areas would be particularly interesting to test the findings of this research. The geographical scope of this study could be extended with the addition of target areas from distant geographic regions but with similar geographical characteristics such as Auckland (New Zeeland), Oslo (Norway) or Vladivostok (Russia).

Finally, the established model performance test setting could be used to examine the geographic transferability of various types of building extraction models. Here, the newest developments in the automated object detection domain point towards large pre-trained models such as vision transformers as an alternative to CNNs (Angelis et al., 2022; Wiguna et al., 2024). In the last years, researcher started to adapt such transformer models, that were originally designed for natural language processing, for remote sensing image segmentation tasks, too.

First studies on the utilization of vision transformer for the building extraction tasks show promising results (Aleossanee et al. ,2023; Song et al., 2023, Yu et al., 2023). Few-shot learning or even zero-shot learning methods, that means the adaptation to new tasks or new geographic regions with extremely few – or even without any – new labelled samples, could replace the complex fine-tuning process of CNN building extraction models (Gella et al., 2023). Vision transformer could be potentially the solution to cope with the tremendous geographical heterogeneity of our world by providing a single building extraction model which can be applied worldwide. However, the extensive testing of such a model on various new, unseen, target areas would be still a precondition for its successful implementation and utilization.

7. BIBLIOGRAPHY

Abdi, G. & Jabari, S. (2021). A multi-feature fusion using deep transfer learning for earthquake building damage detection. *Canadian Journal of Remote Sensing*, 47, 337-352. https://doi.org/10.1080/07038992.2021.1925530

Abdollahi, A., Pradhan, B., Gite, S. & Alamri, A. (2020). Building footprint extraction from high resolution aerial images using Generative Adversial Network (GAN) architecture. *IEEE Access*, 8. 10.1109/ACCESS.2020.3038225

Abriha, D. & Szabo, S. (2023). Strategies in training deep learning models to extract building from multisource images with small training sample sizes. *International Journal of Digital Earth,* 16(1), 1707-1724.https://doi.org/10.1080/17538947.2023.2210312

Aleissaee, A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H. & Khan, F. S. (2023). Transformers in remote sensing: A survey. *Remote Sensing*, 15(7), 1860. https://doi.org/10.3390/rs15071860

Angelis, E., Domi, A., Zamichos, A., Tsouma, M., Drosou, A. & Tzovaras, D. (2022). On the exploration of vision transformers in remote sensing building extraction. *IEEE International Symposium on Multimedia*, 208-215. 10.1109/ISM55400.2022.00046

Arya, D., Maeda, H., Ghos, S. K., Toshniwal, A. M., Kashiyama, T. & Sekimoto, Y. (2021). Deep learning-based road damage detection and classification for multiple countries. *Automation in Construction,* 132, 1-18. https://doi.org/10.1016/j.autcon.2021.103935

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, 3135. https://doi.org/10.3390/rs13163135

Bai, Y., Hu, J., Su, J., Liu, Y., Liu, H., He., X., Meng, S., Mas, E. & Koshimura, S. (2020). Pyramid pooling module based Semi-Siamese Network. A benchmark model for assessing building damages from xBD satellite imagery datasets. *Remote Sensing*, 12(24). https://doi.org/10.3390/rs12244055

Bakirman, T., Komurcu, I. & Sertel, E. (2022). Comperative analysis of deep learning based building extraction methods with the new VHR Istanbul dataset. *Expert Systems With Applications,* 202. https://doi.org/10.1016/j.eswa.2022.117346

Bilkecki, F., Chow, Y. S. & Lee, K. (2023). Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes. *Building and Environment,* 237. https://doi.org/10.1016/j.buildenv.2023.110295

Bouchard, I., Rancourt, M. E., Aloise, D. & Kalitzis, F. (2022). On transfer learning for building damage assessment from satellite imagery in emergency contexts. *Remote Sensing,* 14, 2532. https://doi.org/10.3390/rs14112532

Calton, L. & Wei, Z. (2022). Using Artificial Neural Network models to assess hurricane damage through transfer learning. *Applied Sciences,* 12, 1466. https://doi.org/10.3390/app12031466

Chen, M., Wu, J., Liu, L., Zhao, W., Tian, F., Shen, Q., Zhao, B. & Du, R. (2021). DR-Net: An improved network for building extraction from high resolution remote sensing image. *Remote Sensing,* 13, 294. https://doi.org/10.3390/rs13020294

Chen, S., Ogawa, Y., Zhao, C. & Sekimoto, Y. (2022). Large-scale building footprint extraction from open-sourced satellite imagery via instance segmentation approach. *IGARSS 2022*, 6284.

Cheng, G., Xie, X., Han, J., Guo, L. & Xia, G. S. (2022). *IEEE Journal of Selected Topics in Applied Earth Observations in Remote Sensing*, 13, 3735-3756.

Duarte, D., Nex, F., Kerle, N. & Vosselmann, G. (2018). Multi-resolution feature fusion for image classification of building damages with convolutional neural networks. *Remote Sensing*, 10(10), 1636. https://doi.org/10.3390/rs10101636

El Asri, S. A., Negabi, I., El Adib, S. & Raissouni, N. (2023). Enhancing building extraction from remote sensing images through UNet and transfer learning. *International Journal of Computers and Applications,* 45(5), 413-419.

Gates, B. (2023, March 21). The age of AI has begun. *Gates Notes*. https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

Gao, Y., Lang, S., Tiede, D., Gella, G. W. & Wendt, L. (2022). Comparing OBIA-generated labels and manually annotated labels for semantic segmentation in extracting refugee-dwellings footprints. *Applied Sciences*, 12., 1126. https://doi.org/10.3390/app122111226

Gella, G. W., Tiede, D., Lang, S., Wendit, L. & Gao, Y. (2023). Spatially transferable dwelling extraction from multi-sensor imagery in IDP / refugee settlements: A meta-learning approach. I*nternational Journal of Applied Earth Observations and Geoinformation*, 117. https://doi.org/10.1016/j.jag.2023.103210

Gella, G. W., Pelletier, C., Lefevre, S., Wendt, L, Tiede, D. & Lang, S. (2024). Unsupervised domain adaptation for instance segmentation: Extracting dwellings in temporary settlements across various geographical settings. *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing*, 17, 1701-1718. 10.1109/JSTARS.2023.3336929

Ghaffarian, S., Farhadabad, A. R. & Kerle, N. (2020). Post-disaster recovery monitoring with Google Earth Engine. *Applies Sciences*, 10(13). https://doi.org/10.3390/app10134574

Ghorbanzadeh, O., Tiede, D., Wendt, L., Sundmanns, M. & Lang, S. (2020). Transferable instance segmentation of dwellings in a refugee camp – integrating CNN and OBIA. *European Journal of Remote Sensing,* 54, 127-140.

Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S. R., Tiede, D. & Aryal, J. (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing*, 11(2), 196. https://doi.org/10.3390/rs11020196

Gupta, R., Sajeev, S., Hosfeldt, R., Patel, N., Heim, E., Doshi, J., Lucas, K., Choset, H. & Gaston, M. (2019). Creating xBD: A dataset for assessing building damage from satellite imagery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10-17.

Han, Q., Yin, Q. & Chen, Z. (2022). Remote sensing image building detection method based on Mask R-CNN. *Complex & Intelligent Systems*, 8, 1847-1855.

Hoeser, T.& Kuenzer, C. (2020). Object detection and image segmentation with Deep Learning on earth observation data: A review – Part I: Evolution and recent trends. *Remote Sensing*, 12(10), 1667. https://doi.org/10.3390/rs12101667

Hoeser, T., Bachofer, F. & Kuenzer, C. (2020). Object detection and image segmentation with deep learning on earth observation data: A review – Part II: Applications. *Remote Sensing*, 12(18), 3053. https://doi.org/10.3390/rs12183053

Huang, H., Genyung, S., Zhang, Y., Hao, Y., Zhang, A., Ren, J. & Ma, H. (2019). Combined multiscale segmentation convolutional neural network for rapid damage mapping from post-earthquake very high-resolution images. *Journal of Applied Remote Sensing,* 13(02), 1. 10.1117/1.JRS.13.022007

Ji, S. Wei, S. & Lu, M. (2018). A scale robust Convolutional Neural Network for automatic building extraction from aerial and satellite imagery. *International Journal of Remote Sensing*, 40(9), 3308-3322. https://doi.org/10.1080/01431161.2018.1528024

Ji, M., Liu, L. & Buchroithner, M. (2018). Identifying collapsed buildings using post-earthquake satellite imagery and convolutional neural networks: A case study of the 2010 Haiti earthquake. *Remote Sensing,* 10(11). 1689. https://doi.org/10.3390/rs10111689

Li, W., He, C., Fang, J., Zheng, J., Fu, H. & Yu, L. (2019). Semantic segmentation-based building footprint extraction using very high resolution satellite images and multi-source GIS data. *Remote Sensing*, 11(4), 403. https://doi.org/10.3390/rs11040403

Li, Z., Xin, Q., Sun, Y. & Cao, M. (2021). A deep learning-based framework for automated extraction of building footprint polygons from very high-resolution aerial imagery. *Remote Sensing*, 13(18), 3630. https://doi.org/10.3390/rs13183630

Lin, Q., Ci, T., Wang, L., Mondal, S. K., Yin, H. & Wang, Y. (2022). Transfer Learning for improving seismic building damage assessment. *Remote Sensing*, 14(1), 201. https://doi.org/10.3390/rs14010201

Kang, J., Fernandez-Beltran, R., Sun, X., Ni, J. & Plaza, A. (2021). Deep learning-based building footprint extraction with missing annotations. *IEEE Geoscience and Remote Sensing Letters*, 19. 10.1109/LGRS.2021.3072589

Khelifi, L. & Mignotte, M. (2020). Deep learning for change detection in remote sensing images: A comprehensive review and meta-analysis. *IEEE Access,* 8, 126385-126400. 10.1109/ACCESS.2020.3008036

Kirilov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollar, P. & Girschick, R. (2023). Segment Anything. *ArXiv*. https://doi.org/10.48550/arXiv.2304.02643

Kirkpatrick, J., Pascanu, R., Rabinowitz., N. et al. (2017). Overcoming catastrophic forgetting in neural networks. *Applied Mathematics*, 114(13), 3521-3526. https://doi.org/10.1073/pnas.1611835114

Kucharczyk, M., Hay, G. J., Ghaffarian, S., & Hugenholtz, C. H. (2020). Geographic object-based image analysis. A primer and future directions. *Remote Sensing*, 12(12), 2012. https://doi.org/10.3390/rs12122012

Li, W., He, C., Fang, J., Zheng, J. Fu, H. & Yu, L. (2019). Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sensing*, 11(4), 403. https://doi.org/10.3390/rs11040403

Li, Q., Mou, L., Hua, Y., Shi, Y. & Zhu, X. X. (2022). CrossGeoNet: A framework for building footprint generation of label-scare geographical regions. *International Journal of Applied Earth Observations and Geoinformation*, 11. https://doi.org/10.1016/j.jag.2022.102824

Li, Q., Mou, L., Hua, Y., Shi, Y & Zhu, X. X. (2024). A review of building extraction from remote sensing imagery: Geometrical structures and semantic attributes. *IEEE Transactions on Geoscience and Remote Sensing,* 62 https://doi.org/10.1109/TGRS.2024.3369723

Li, Z. & Dong, J. (2022). A framework integrating Deeplab V3+, transfer learning, active learning, and incremental learning for mapping building footprints. *Remote Sensing,* 14, 4788. https://doi.org/10.3390/rs14194738

Liu, P., Liu, X., Liu, M., Shi, Q., Yang, J., Xu, X. & Zhang, Y. (2019) Building footprint extraction from high-resolution images via Spatial Residual Inception Convolutional Neural Network. *Remote Sensing*, 11(7), 830. https://doi.org/10.3390/rs11070830

Liu, Y., Zhou, J., Qi, W., Li, X., Gross, L., Shao, Q., Zhao, Z., Ni, L. & Li, Z. (2020). ARC-Net: An efficient network for building extraction from high-resolution aerial images. *IEEE Access*, 8.

Liu, Y., Chen, D., Ma, A., Zhong, Y., Fang, F. & Xu, K. (2021). Multiscale U-shaped CNN building instance extraction framework with edge constraint for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7), 6106-6120.

Luo, L., Li, P. & Yan, X. (2021). Deep learning-based building extraction from remote sensing images: A comprehensive review. *Energies*, 14(23), 7982. https://doi.org/10.3390/en14237982

Luo, M., Ji, S. & Wei, S. (2023). A diverse large-scale building dataseet and a novel plug and play domain generalization method for building extraction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16.

Ma, H., Liu, Y., Ren, Y. & Yu, J. (2019). Detection of collapsed buildings in post-earthquake remote sensing images based on the improved YOLOv3. *Remote Sensing*, 12(1), 44. https://doi.org/10.3390/rs12010044

Maggiori, E., Tarabalka, Y., Charpiat, G. & Alliez, P. (2017). Can semantic labelling methods generalize to any city? The INRIA aerial labelling benchmark. *IGARSS 2017*, 3226-3229.

Martin, S. S. & Pradhan, B. (2021). Challenges and limitations of earthquake-induced building damage mapping techniques using remote sensing images. A systematic review. *Geocarto International,* 37(21), 6181-6212.

Maxwell, A. E., Warner, T. A. & Guillien, L. A. (2021). Accuracy assessment on Convolutional Neural Network-based Deep Learning remote sensing studies – Part 2: Recommendations and best practices. *Remote Sensing*, 13(13), 2591. https://doi.org/10.3390/rs14010201

Minaee, S., Boykov, Y., Porikli, F., Plaza, A. & Terzopoulos, D. (2022). Image segmentation using Deep Learning. A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 3523.

Neupane, B., Horanont, T. & Aryal, J. (2021). Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sensing*, 13, 808. https://doi.org/10.3390/rs13040808

Nex, F., Duarte, D., Tonolo, F. G. & Kerle, N. (2019). Structural building damage detection with deep learning: Assessment of a state of the art CNN in operational conditions. *Remote Sensing*, 11(23), 2765. https://doi.org/10.3390/rs11232765

Nurkarim, W. & Wijayanto, A. W. (2023). Building footprint extraction and counting on very high-resolution satellite imagery using object detection deep learning framework. *Earth Science Informatics,* 16, 515-532.

Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P. & Vateekul, P. (2019). Semantic segmentation on remotely sensed images using an enhanced Global Convolutional Network with channel attention and domain specific transfer learning*. Remote Sensing*, 11(1), 83. https://doi.org/10.3390/rs11010083

Pi, Y., Nath, N. D. & Behzadan, A. H. (2020). Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics*, 43, 101009.

Prakash, P. & Aithal, B. H. (2022). Building footprint extraction from very high-resolution satellite images using deep learning. *Journal of Spatial Science,* 68(3), 487-503. https://doi.org/10.1080/14498596.2022.2037473

Rastogi, K., Bodani, P. & Sharma, A. (2022). Automatic building footprint extraction from very high-resolution imagery using deep learning techniques. *Geocarto International*, 37(5), 1501-1513. https://doi.org/10.1080/10106049.2020.1778100

Qiu, C., Li, H., Guo, W., Chen, X., Yu, A. & Tong, X. (2022). Transferring transformer-based models for cross-area building extraction from remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 15, 4104-4116. 10.1109/JSTARS.2022.3175200

Sakeena, M., Stumpe, E., Despotivic, M., Koch, D. & Zeppelzauer, M. (2023). On the robustness and generalization ability of building footprint extraction on the example of SegNet and Mask R-CNN. *Remote Sensing*, 15(8), 2135. https://doi.org/10.3390/rs15082135

Sawa, K., Yalcin, I. & Kocaman, S. (2024). Building detection from SkySat images with transfer learning: A case study over Ankara. *PFG Journal of Photogrammetry, Remote Sensing and Geoinformation Science*. https://doi.org/10.1007/s41064-024-00279-x

Shao, J., Tang, L, Liu, M., Shao, G., Sun, L. & Qiou, Q. (2020). BDD-Net: A general protocol for mapping buildings damaged by a wide range of disasters based on satellite imagery. *Remote Sensing*, 12(10), 1670. https://doi.org/10.3390/rs12101670

Song, J., Zhu, A. X. & Zhu, Y. (2023). Transformer-based semantic segmentation for extraction of building footprints from very-high resolution images. *Sensors*, 23, 5166. https://doi.org/10.3390/s23115166

Spasov, A. & Petrova-Antonova, D. (2021). Transferability assessment of open-source deep learning model for building detection on satellite data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 66(4), 107-110.

Stiller, D., Stark, T., Wurm, M. Dech, S. & Taubenbök, H. (2019). Large-scale building extraction in very high-resolution aeiral imagery using Mask R-CNN. *IEEE Xplore*.

Sublime, J. & Kalinicheva, E. (2019). Automatic post-disaster damage mapping using deep-learning techniques for change detection: A case study of the Tohoku Tsunami. *Remote Sensing,* 11(9), 1123. https://doi.org/10.3390/rs11091123

Tahir, A., Munawar, H. S., Akram, J., Adil, M., Ali, S., Kouani, A. Z. & Mahmud, M. A. (2022). Automatic target detection from satellite imagery using machine learning. *Sensors,* 22(3), 1147. https://doi.org/10.3390/s22031147

Tiede, D., Schwendemann, G., Alobaidi, A., Wendt, L. & Lang, S. (2021). Mask R-CNN-based building extraction from VHR satellite data in operational humanitarian action: An example related to Covid-19 response in Khartoum, Sudan. *Transactions in GIS*, 25, 1213-1227.

Touzani, S., & Granderson, J. (2021). Open data and deep semantic segmentation for automated extraction of building footprints. *Remote Sensing*, 13(3), 2578. https://doi.org/10.3390/rs13132578

Usmani, M., Bovolo, F. & Napolitano, M. (2023). Remote sensing and deep learning to understand noisy OpenStreetMap. *Remote Sensing*, 15(18), 4639. https://doi.org/10.3390/rs15184639

Valentijn, T., Margutti, J, von den Homberg, M. & Laaksonen, J. (2020). Multi-hazard and spatial transferability of a CNN for automated building damage assessment. *Remote Sensing*, 12(17), 2839. https://doi.org/10.3390/rs12172839

Wang, W., Shi, Y., Zhang, J., Hu, L., Li, S., He, D. & Liu, F. (2023). Traditional village building extractio based on improved Mask R-CNN: A case study of Beijing, China. *Remote Sensing*, 15(10), 2616. https://doi.org/10.3390/rs15102616

Wiguna, S., Adriano, B & Koshimura, S. (2024). Evaluation of deep learning models for building damage mapping in emergency response settings. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 5651-5667.

Wurm, M., Stark, T., Zhu, X. X., Weigand, M. & Taubenböck, H. (2019). Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150, 59-69.

Xie, Y., Cai, J., Bhojwani, R., Shekhar, S. & Knight, J. (2019). A locally-constrained YOLO framework for detecting small and densely-distributed building footprints. *International Journal of Geographical Information Science*, 34(4), 777-801. https://doi.org/10.1080/13658816.2019.1624761

Xie, Y., Zhung, Cao, Y., Feng, D., Hu, M., Li, W., Zhang, J. & Fu, L. (2020). Refined extraction of building outlines from high-resolution remote sensing imagery based on a Multifeature Convolutional Neural Network and morphological filtering. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 1842-1855.

Xu, J. Z., Lu, W., Li, Z., Khatian, P. & Zaytseva, V. (2019). Building damage detection in satellite imagery using Convolutional Neural Networks. 33rd *Conference on Neural Information Processing Systems*, Vancouver, Canada.

Yang, W, Zhang, X. and Luo, P. (2021). Transferability of Convolutional Neural Network models for identifying damaged buildings due to earthquake. *Remote Sensing,* 13(3), 504. https://doi.org/10.3390/rs13030504

Ye, Z., Fu, Y., Gan, M., Deng, J., Comber, A. & Wang, K. (2019). Building extraction from very high resolution aerial imagery using Joint Attention Deep Neural Network. *Remote Sensing,* 11(24), 2970. https://doi.org/10.3390/rs11242970

Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W. & Zhao, T. (2019). Semantic segmentation of urban buildings from VHR remote sensing imagery using a Deep Convolutional Neural Network. *Remote Sensing*, 11(15), 1774. https://doi.org/10.3390/rs11151774

Yu, T., Tang, P., Zhao, B., Bai, S., Gou, P., Liao, J & Jin, C. (2023). ConvBNet: A convolutional network for building footprint extraction. *IEEE Geoscience and Remote Sensing Letters*, 20.

Yuan, X., Shi, J. & Gu, L. (2021). A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems With Applications*, 169, 114417.

Zhan, Y., Liu, W. & Maruyama, Y. (2022). Damaged building extraction using modified Mask R-CNN model using post-event aerial images of the 2016 Kumamoto earthquake. *Remote Sensing*, 14(4), 1002. https://doi.org/10.3390/rs14041002

Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3), 107-115. https://doi.org/10.1145/3446776

Zhang, L, Wu, J., Fan, Y., Gao, H. & Shao, Y. (2020). An efficient building extraction model from high spatial resolution remote sensing images based on improved Mask R-CNN. *Sensors*, 20(5), 1465. https://doi.org/10.3390/s20051465

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43-76.

Zhou, R., Liu, J., Pan, H., Tang, D. & Zhou, R. (2024). An improved instance segmentation method for fast assessment of damaged buildigs based on post-earthquake UAV images. *Sensors*, 24(13), 4371. https://doi.org/10.3390/s24134371

Wang, Y., Cui, L., Zhang, C., Chen, W., Xu, Y. & Zhang, Q. (2022). A two-stage seismic damage assessment method for small, dense, and imbalanced buildings in remote sensing images. *Remote Sensing*, 14(4), 1012. https://doi.org/10.3390/rs14041012

Wu, X., Sahoo, D. & Hoi, S. C. (2020). Recent advances in deep learning for object detection. *Neurocomputing*, 396, 39-64.

8. APPENDIX

Appendix 1: Target area resolution and data sources

Barcelona 25 cm
https://visors.icgc.cat/appdownloads/index.html
Chemnitz 20 cm
https://www.geodaten.sachsen.de/downloadbereich-dop-4826.html
Daugavpils 25 cm
https://www.lgia.gov.lv/en/color-orthophoto-map-2016-2018-cycle-6
Girona 25 cm
https://visors.icgc.cat/appdownloads/index.html
Kosice 20 cm
https://www.geoportal.sk/en/zbgis/orthophotomosaic/2nd-cycle/
Narva 20 cm
https://geoportaal.maaamet.ee/est/Ruumiandmed/Ortofotod-p99.html
New York I&II 50 cm
https://data.cityofnewyork.us/browse?q=imagery&sortBy=relevance
Tallinn 25 cm
https://geoportaal.maaamet.ee/est/Ruumiandmed/Ortofotod-p99.html
Vienna 20 cm
https://www.wien.gv.at/ma41datenviewer/public/start.aspx

Appendix 2: Location of target areas

Figure 36: Location of the target areas. The colours indicate the geographical closeness between the individual target areas. Baltic region (dark blue), Central Europe (turquoise and green), Mediterranean (yellow) and United States (dark green).

Appendix 3 Target area image samples

Barcelona

Figure 37: Barcelona satellite imagery

Figure 38: Barcelona ground reference

Chemnitz

Figure 39: Chemnitz satellite imagery

Figure 40: Chemnitz ground reference

Daugavpils

Figure 41: Daugavpils satellite imagery

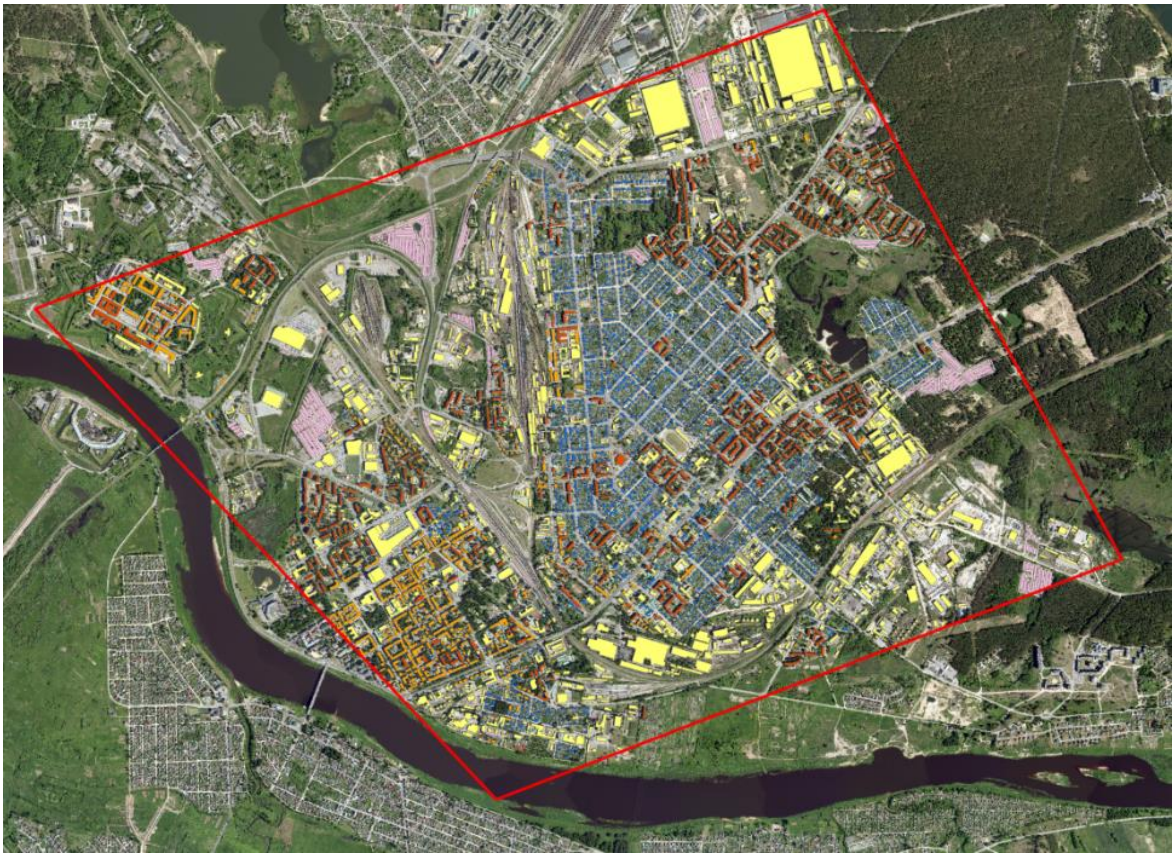

Figure 42: Daugavpils ground reference

Girona

Figure 43: Girona satellite imagery

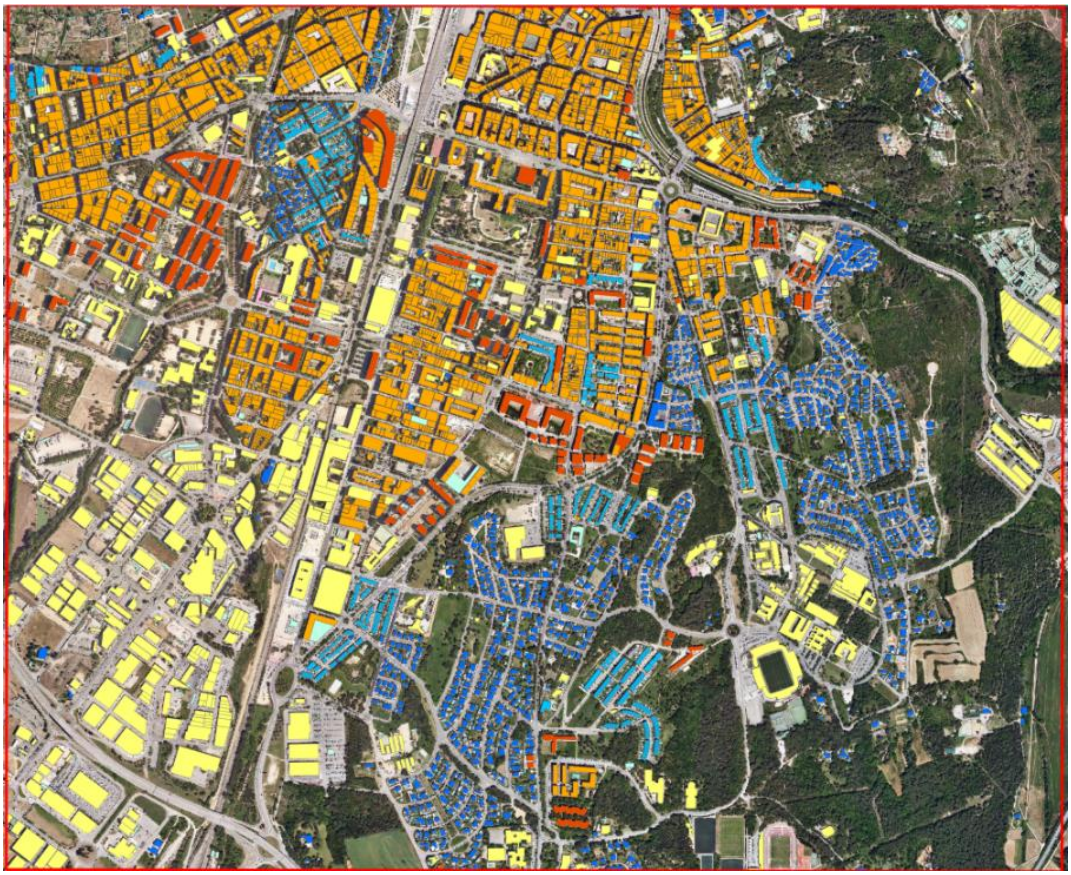

Figure 44: Girona ground reference

Kosice

Figure 45: Kosice satellite imagery

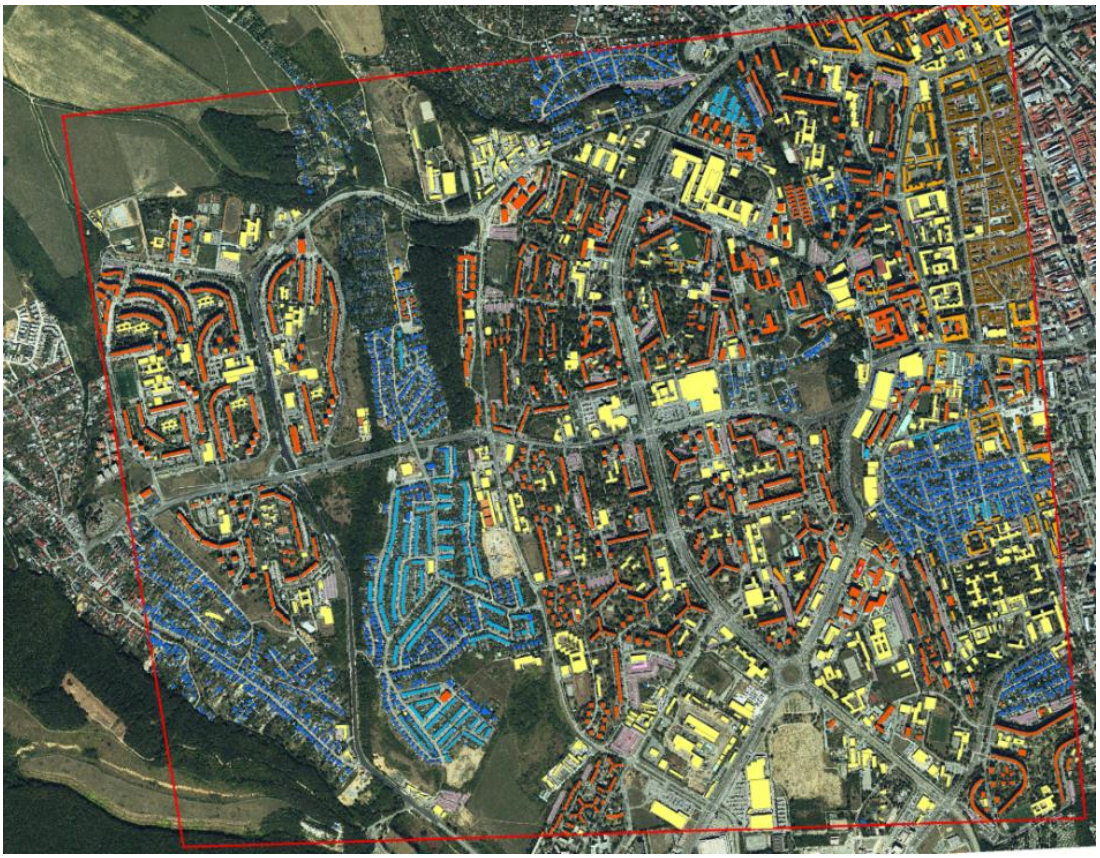

Figure 46: Kosice ground reference

Narva

Figure 47: Narva satellite imagery

Figure 48: Narva ground reference

New York I

Figure 49: New York I satellite imagery

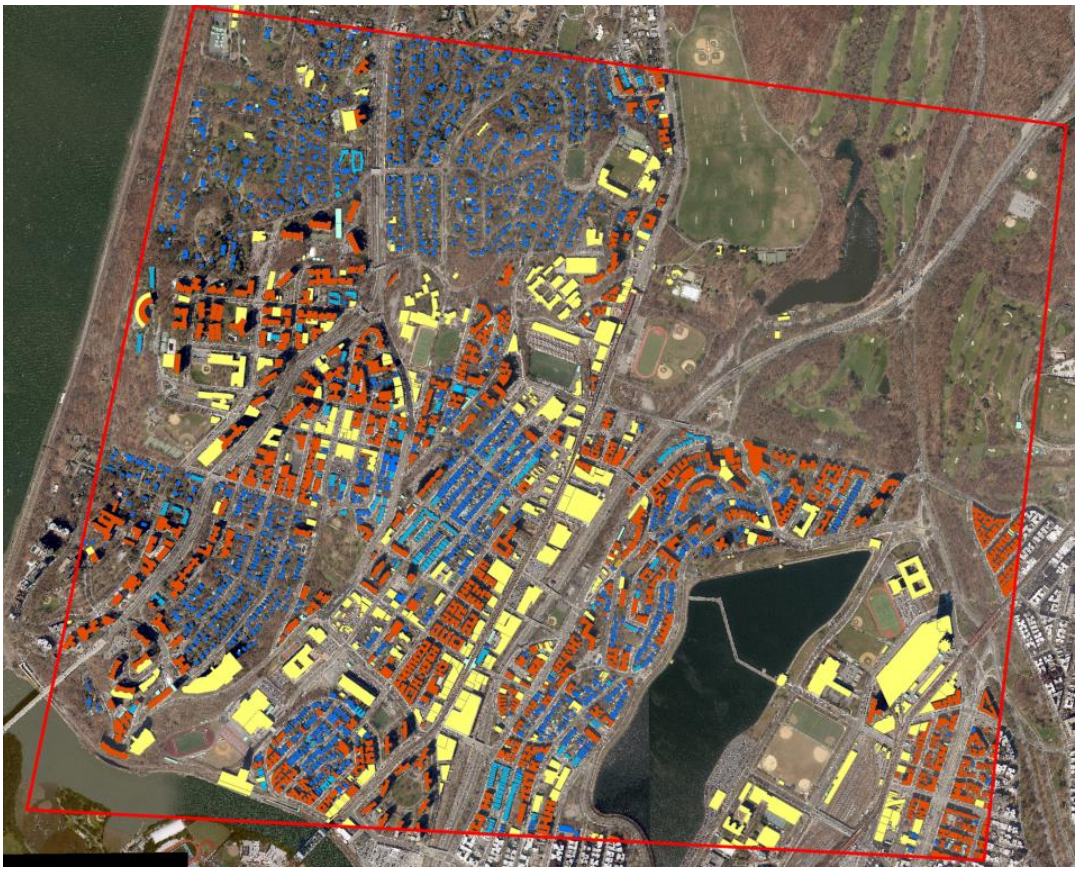

Figure 50: New York I ground reference

New York II

Figure 51: New York II satellite imagery

Figure 52: New York II ground reference

Tallinn

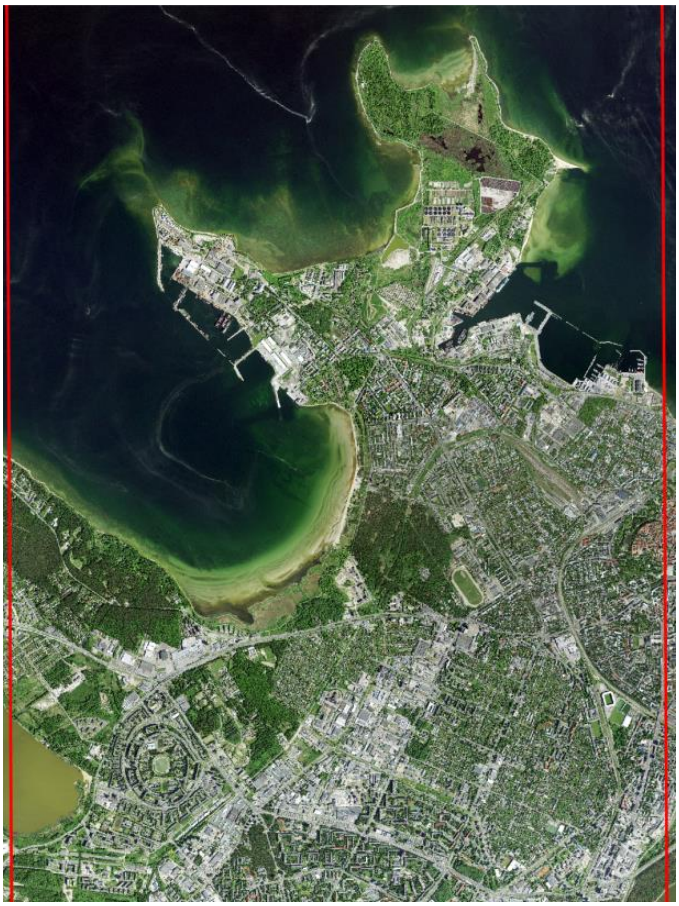

Figure 53: Tallinn satellite imagery

Figure 54: Tallinn ground reference

Vienna

Figure 55: Vienna satellite imagery

Figure 56: Vienna ground reference

Appendix 4: Building extraction results

Target Area/	TP				FP				FN				Ground Reference		
Model	Basic	DisB	X25	T80	Basic	DisB	X25	T80	Basic	DisB	X25	T80	Basic	DisB	X25/T80
EST TLN															
PT USA	9210	9240	9020	8439	**5171**	**3022**	**2453**	**1313**	4783	2230	1815	2396	13993	11470	10835
FT USA TLN	**10539**	**9922**	**9367**	**9223**	11833	8414	3835	2783	**3454**	**1548**	**1468**	**1612**	13993	11470	10835
FT USA VNA	10209	9753	9158	9002	11905	8478	3146	2160	3784	1717	1677	1833	13993	11470	10835
PT AFR	6689	6227	6007	3030	12976	9262	8315	1715	7304	5243	4828	7805	13993	11470	10835
FT AFR VNA	9056	7961	7488	7351	23080	15038	7582	4843	4937	3509	3347	3484	13993	11470	10835
EST NRV															
PT USA	2130	2083	1906	1782	**1490**	**1044**	***777***	500	964	653	442	566	3094	2736	2348
FT USA TLN	**2468**	2321	1973	1940	4945	3800	1244	972	**626**	415	375	408	3094	2736	2348
FT USA VNA	2459	**2328**	**1990**	**1962**	4687	3624	1077	796	635	**408**	**358**	**386**	3094	2736	2348
PT AFR	1749	1635	1510	**955**	2550	1428	*1054*	**327**	1345	1101	838	1593	3094	2736	2348
FT AFR VNA	2115	1872	1581	1597	9582	6075	2058	2289	979	864	767	751	3094	2736	2348
LVA DGV															
PT USA	5610	5382	5181	4614	**3187**	**2245**	***1844***	1014	3962	3646	2487	3054	9572	9028	7668
FT USA TLN	7080	6439	**5719**	**5446**	5914	4143	*2265*	1430	2492	2589	**1949**	2222	9572	9028	7668
FT USA VNA	**7310**	**6755**	5689	5531	8528	5894	2027	1419	**2262**	**2273**	2027	**2137**	9572	9028	7668
PT AFR	3847	3510	3426	1276	3186	1959	1691	**341**	5725	5518	4242	6392	9572	9028	7668
FT AFR VNA	6065	4218	3786	4188	10129	4571	*2665*	1827	3507	4810	3882	3480	9572	9028	7668
SLK KOS															
PT USA	3192	3001	2718	2379	**2480**	**1039**	**875**	584	8340	2325	1611	1950	11532	5326	4329
FT USA TLN	**4938**	**3718**	**3003**	**2903**	8682	5045	1819	1567	**6594**	1808	**1326**	**1426**	11532	5326	4329
FT USA VNA	4138	3526	2864	2774	5826	3010	1125	969	7394	**1800**	1465	1555	11532	5326	4329
PT AFR	1577	1249	1127	420	4569	1903	1455	**409**	9955	4077	3202	3909	11532	5326	4329
FT AFR VNA	3873	2391	1832	2191	17228	7322	2166	1796	7659	2935	2497	2138	11532	5326	4329
GER CHM															
PT USA	2779	2484	2268	2107	**4652**	**1458**	**976**	735	4617	1616	1241	1402	7396	4100	3509
FT USA TLN	**4226**	**3010**	**2563**	**2415**	6905	3636	1615	1396	**3170**	**1090**	**946**	**1094**	7396	4100	3509
FT USA VNA	3755	2736	2323	2229	6807	3228	1346	1096	3641	1364	1186	1280	7396	4100	3509
PT AFR	1397	1256	1178	486	5269	2370	1382	**623**	5999	2844	2331	3023	7396	4100	3509
FT AFR VNA	3189	1508	1232	1553	18610	6744	3735	2730	4207	2592	2777	1956	7396	4100	3509
AUS VNA															
PT USA	4249	4290	4068	3799	**2683**	**1253**	**1031**	693	5490	2813	858	1127	9739	7103	4926
FT USA TLN	5738	**5461**	**4191**	**4100**	5605	3494	1602	1031	4001	1642	**735**	**826**	9739	7103	4926
FT USA VNA	**6224**	5645	4171	4079	7381	4945	1376	1117	**3515**	**1458**	755	847	9739	7103	4926
PT AFR	3601	3660	3398	1880	3132	2126	1831	**405**	6138	3443	1528	3046	9739	7103	4926
FT AFR VNA	4842	4217	3404	3412	9698	5191	2496	1643	4897	2886	1522	1514	9739	7103	4926
ESP GIR															
PT USA	1564	1600	1582	1466	**1839**	**649**	**529**	**314**	4099	1125	827	943	5663	2725	2409
FT USA TLN	2095	1737	1592	1447	3269	1952	1184	567	3568	988	817	962	5663	2725	2409
FT USA VNA	**2222**	**1835**	**1670**	**1597**	4987	2309	1006	715	**3441**	**890**	**739**	**812**	5663	2725	2409
PT AFR	881	971	969	400	1491	712	603	**87**	4782	1754	1440	2009	5663	2725	2409
FT AFR VNA	1767	1261	1138	1162	7167	2888	1425	1027	3896	1464	1271	1247	5663	2725	2409
ESP BAR															
PT USA	612	**347**	**343**	**294**	**2779**	**562**	**388**	274	6759	**1211**	**1147**	**1196**	7371	1558	1490
FT USA TLN	**920**	337	323	181	3333	1257	833	346	**6451**	1221	1167	1309	7371	1558	1490
FT USA VNA	791	339	325	246	6556	1710	853	655	6580	1219	1165	1244	7371	1558	1490
PT AFR	112	53	53	9	1730	928	684	**56**	7259	1505	1505	1481	7371	1558	1490
FT AFR VNA	536	76	63	42	7511	4048	2376	860	6825	1482	1427	1448	7371	1558	1490
USA NY I															
PT USA	**2766**	2171	**2113**	2020	**1979**	**862**	**572**	418	**838**	**575**	**496**	**589**	3604	2746	2609
FT USA TLN	2446	2046	1965	1929	4197	2168	821	482	1158	700	644	680	3604	2746	2609
FT USA VNA	2680	**2196**	2097	**2035**	5074	2818	708	609	924	550	512	574	3604	2746	2609
PT AFR	960	763	743	180	3146	1913	1467	**169**	2641	1983	1866	2429	3604	2746	2609
FT AFR VNA	1654	858	824	982	10974	4116	1783	1312	1950	1888	1785	1627	3604	2746	2609
USA NY II															
PT USA	8539	4711	**4363**	4000	**2993**	**2029**	**1595**	1510	3007	2042	**1033**	1396	11546	6753	5396
FT USA TLN	7477	5114	4181	3912	7270	4437	1874	1365	4069	1639	1215	1484	11546	6753	5396
FT USA VNA	**9427**	**5464**	4332	**4118**	8151	6117	2509	2609	**2119**	**1289**	1064	**1278**	11546	6753	5396
PT AFR	2009	1817	1677	365	5093	2579	1924	**222**	9537	4936	3719	5031	11546	6753	5396
FT AFR VNA	4759	1680	1401	2596	17753	5120	1891	1742	6787	5073	3995	2800	11546	6753	5396

Table 17: Building extraction results overview. True positives (TP), False positives (FP). False negatives (FN) for the initial building extraction without post-processing, after application of dissolve boundaries (DisB), deletion of objects smaller than 25 m² (X25) and increase of the prediction confidence threshold to 80% (T80). Best results in bold.

Appendix 5: Building extraction on the Tallinn target area

Figure 57: PT USA on Tallinn

Figure 58: PT AFR on Tallinn

Figure 59: FT USA TLN on Tallinn

Figure 60: FT USA VNA on Tallinn

Figure 61: FT AFR VNA on Tallinn

Appendix 6: Building extraction overview samples

Figure 62: Tallinn ground reference

Figure 63: FT USA VNA on Tallinn

Figure 64: Vienna ground reference

Figure 65: FT USA TLN on Vienna

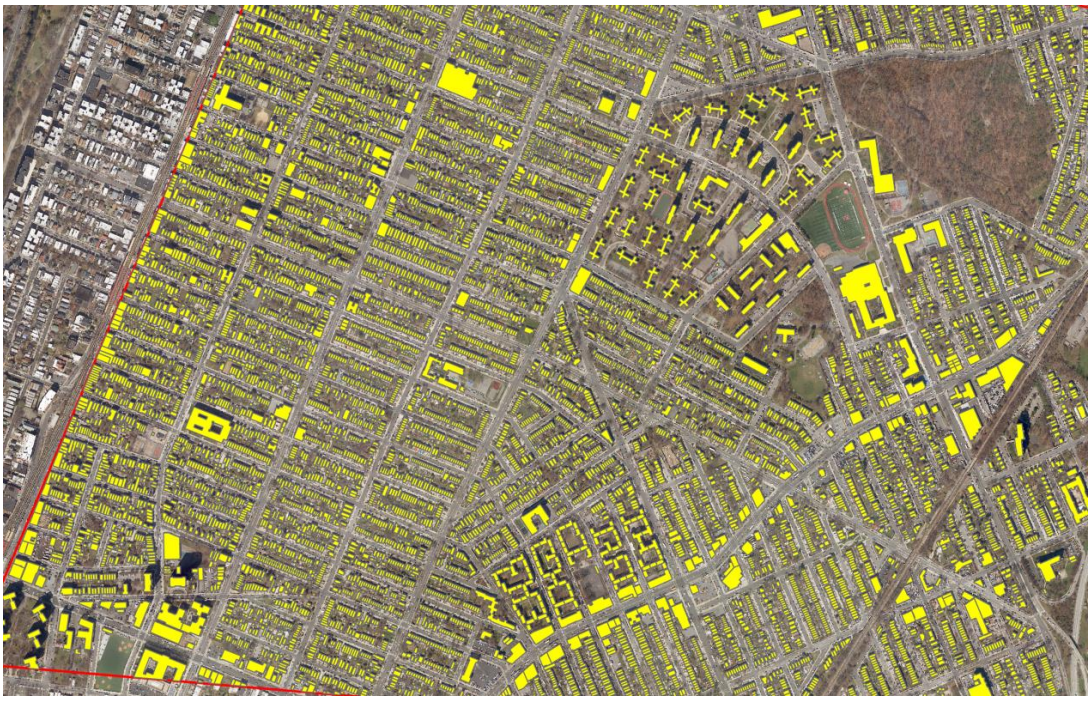

Figure 66: New York II ground reference

Figure 67: PT USA on New York II