



Landscape quality assessments using deep learning

Alex Levering

Propositions

1. Visual information is enough to predict the quality of landscapes.
(this thesis)
2. More research emphasis on small datasets is essential for deep learning methods to become commonplace in landscape quality assessments.
(this thesis)
3. Working from home is detrimental for academic output.
4. Having an uncommonly occurring last name is an unfair advantage in scientific publishing.
5. Society has not retained any of the lessons learned from the Covid-19 pandemic.
6. Relying on navigation systems rather than on maps and landmarks causes one to become out of touch with their surroundings.

Propositions belonging to the thesis, entitled

Landscape Quality Assessments using Deep Learning

Alex Levering

Wageningen, 26 January 2024

Landscape quality assessments using deep learning

Alex Levering

Thesis committee

Promotor

Prof. Dr D. Tuia
Professor of Geo-information Sciences
Wageningen University & Research

Co-promotor

Dr D. Marcos
Assistant Professor
Inria, Montpellier, France

Other members

Prof. Dr E.S. van Leeuwen, Wageningen University & Research
Prof. Dr A. Singleton, University of Liverpool, United Kingdom
Dr C.M. Gevaert, University of Twente, Enschede
Prof. Dr J. Chanussot, Inria, Grenoble, France

This research was conducted under the auspices of the C.T. de Wit Graduate School of Production Ecology & Resource Conservation (PE&RC)

Landscape quality assessments using deep learning

Alex Levering

Thesis

submitted in fulfilment of the requirements for the degree of doctor at

Wageningen University

by the authority of the Rector Magnificus

Prof. Dr A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Friday 26 January 2024

at 4 p.m. in the Omnia Auditorium.

Alex Levering
Landscape quality assessments using deep learning,
134 pages.

PhD thesis, Wageningen University, Wageningen, NL (2024)
With references, with summary in English

ISBN 978-94-6447-168-7
DOI <https://doi.org/10.18174/644684>

Contents

	Page
Contents	v
Acronyms	vii
Chapter 1 Introduction	1
Chapter 2 Predicting liveability with semantic intermediate concepts	13
Chapter 3 On the relation between landscape beauty and land cover	41
Chapter 4 Cross-modal learning of housing quality in amsterdam	61
Chapter 5 Scenicness assessments with vision-language models	71
Chapter 6 Synthesis	89
References	99
Summary	117
Acknowledgements	119
About the author	121
PE&RC Training and Education Statement	123

Acronyms

CLC	Corine Land Cover
CNN	Convolutional Neural Network
DL	Deep Learning
GSV	Google Street View
HMI	Human-Machine Interaction
LBM	Leefbaarometer
LPE	Landscape Prompt Ensembling
LQ	Landscape Quality
ML	Machine Learning
MLP	Multi-Layer Perceptron
NIR	Near-Infrared
PP2	Place Pulse 2
RS	Remote Sensing
RMSE	Root Mean Squared Error
SON	ScenicOrNot
VLM	Vision-Language Model

Chapter 1

Introduction

1.1 Context

Landscapes are an ever-present aspect of life, and the way in which humans experience these landscapes can shape their lives in various ways. Therefore, they are increasingly being recognised as an important resource for many facets of life. Landscapes are an important regulator for our emotional and physical well-being. For instance, high-quality natural landscapes correlate with positive emotions such as warmth and cheerfulness (Daniel and Vining, 1983), comfortableness, tranquillity, and safety (Galindo and Rodriguez, 2000), and happiness (Daniel and Vining, 1983; Seresinhe et al., 2019). The importance of landscapes transcends personal health, as scenic landscapes are a driver for tourism (Krippendorf, 1984), as well as cultural ecosystem services (Daniel et al., 2012; Havinga et al., 2021). Similar patterns are observed for urban residential spaces. Living in destitute neighbourhoods is associated with higher mortality rates (Haan et al., 1987), worse dietary and physical activity patterns (Thompson and Kent, 2014), and an increase in morbidity (Barber et al., 2016). Living in lower-quality housing is also detrimental for mental well-being (Evans, 2003). In aggregate, landscape perception and appreciation are crucial for land management activities (Solecka, 2019). Evidently, understanding the qualities of our landscapes is important for our overall well-being, and it may help guide decisions on how to manage landscapes.

The definition of a landscape can vary greatly depending on the research intent. For the purposes of assessing the qualities of landscapes based on perception alone, the definition is adjusted to leave out non-visible landscape elements that cannot be comprehended from single images, such as cultural or historical values or species distributions (Amir and Gidalizon, 1990). When considering visual factors, landscapes can be defined as *"a portion of a territory that the eye can comprehend in a single view"* (Daniel, 2001), or alternatively, *"the outdoor environment, natural or built, which can be directly perceived by a person visiting and using that environment"* (Hull and Revell, 1989). This definition

can be extended to describe *landscape qualities* (LQs). There is much debate over the nature of LQs, specifically about the degree of consensus between people and the degree to which they can be measured and compared. (Shuttleworth, 1979; Jacques, 1980). In practice, a mixture of methods inspired by both mindsets can be found in existing research. For instance, descriptive models can be defined by experts to be comprised of a set of visible aspects that are known to correlate with a particular LQ. Such methods excel at performing analyses at large spatial scales, as they can use commonly available digital methods, such as geographic information systems (Bubalo et al., 2019; Huang and Liu, 2022). However, a typical downside of such methods is a lack of input from the public, which would reflect the inherent subjectivity of LQs. Instead, *public preference* research approaches derive LQ opinions from the public through methods such as surveys, interview panels, or research site visits (Arthur, 1977; Schroeder and Daniel, 1981; Nahuelhual et al., 2018). More recently, the internet has enabled crowdsourcing efforts at unprecedented scales, allowing for the collection of many first-hand accounts at once (Wherrett, 1998; Naik et al., 2014; Seresinhe et al., 2015; Bubalo et al., 2019). Such methods have reduced the costs and the amount of manual processing needed. However, as a result of crowdsourcing being performed anonymously, there is little insight into the study participants, while the veracity of first-hand accounts depends on factors such as the socio-economic status of respondents and the duration of the observation (Jacques, 1980; Amir and Gidalizon, 1990). As a result, studies performed through crowdsourcing are less informative, as respondents cannot be studied in conjunction with their responses.

Of particular interest are so-called *psychophysical* approaches, which attempt to link subjective crowd consensus studies to physical, biological, or social features of the environment (Arthur, 1977; Wherrett, 1998). Such studies may use predictive models to derive respondent preferences for a demarcated set of objective landscape elements or variables, typically through regression models (Arthur, 1977; Buhyoff and Riesenman, 1979). As a hybrid approach, these methods therefore represent a middle ground between objective and subjective studies by linking subjective opinions to objective landscape elements. Therefore, they require extensive in-depth knowledge about photographic representations of landforms as well as a manual process of defining and extracting features from images (Wherrett, 1998). Because of these characteristics, psychophysical analyses can be performed within the framework of *Machine Learning* (ML) methods, which leverage data to construct models in order to generate predictions across a set of tasks (Jordan and Mitchell, 2015). The main purpose of psychophysical modelling using ML methods is to learn from *labelled* examples, which in the context of this thesis are images with matching LQ reference scores. Trained models can then be used to provide predictions on new data as well as for knowledge extraction. This thesis follows this line of reasoning.

1.2 LQ assessment using machine learning

1.2.1 Learning to predict LQs

In order to train ML models for visual LQ assessments, it is necessary to fine-tune them over the provided image examples. The most common approach is to train models through a process referred to as *supervised learning*, which consists of five steps:

1. Extract *features* from images. An image feature is any pattern found in that image that describes a landscape quality. Extracted features can be of variable complexity. For instance, in the context of landscape beauty, a meaningful but simple pattern could be the amount of greenness in an image, which is a low-level feature. A more complex feature could be the interaction between a verdant lake and a snowy mountain, since it requires both geometric information (e.g., the shape of a mountain), colour information, and their position relative to one another. In a classical ML workflow, the modeller needs to decide which patterns are useful to extract prior to the fine-tuning process.
2. Using the extracted features, fit a prediction algorithm to relate the extracted features to the dependent variable, such as LQ ratings. The model can be anything from simple linear least-squares regression to complex neural networks. The form of the dependent variable can vary greatly as well. Examples include binary classifiers (LQ is present or absent), regression (a continuous scale ranging over the expected prevalence of a LQ), or even relative examples (the preference of people for either image A or B). The best possible fit of the model is calculated according to the chosen learning objective. For instance, in a regression setting, if a given image has a reference value of 6 out of 10 and the fitted model predicts a value of 5, then it can be fitted to minimise the absolute difference. After fitting, the model is said to have been *trained*, with an optimal fit based on the features that were provided.
3. Use the model on a hold-out set to determine how well it predicts the LQs of unseen samples. A fitted model is used on a set of new images with reference values to calculate its expected performance on new examples. This process is known as *validation*. A poor validation performance is an indicator that the chosen features do not work well beyond the training set. The modeller can then repeat steps 1 through 3 to train a model with more suitable features.
4. Once the modeller is satisfied with the performance of the model on the validation dataset of step 4, it can then be used to predict new images. During this stage, the model can be used to perform *inference*, where the values of new images are predicted without access to reference scores. By extension, it can be *tested* on another hold-out set of examples. The performance on this set is reported as the performance of the model.

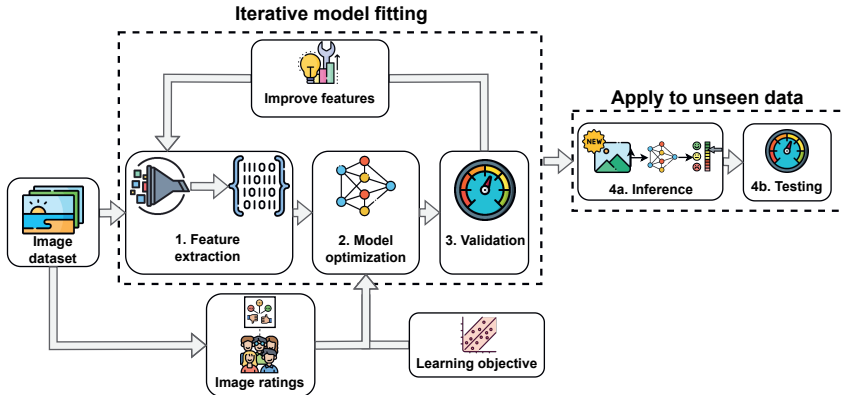


Figure 1.1: Supervised learning loop for training and ML models. The process requires a dataset of images, which are labelled by volunteers. During the model fitting process, the modeller iteratively extracts features (step 1), optimises a model on training examples according to a chosen learning objective (step 2), and assesses validation performance (step 3). When the modeller is satisfied with the model, it can be used on new examples (step 4a). The model can then also be tested for generalisation performance (step 4b) by comparing its predictions against new examples with labels.

A visual illustration of the supervised training process is given in Figure 1.1.

Deep Learning (DL) models are a particularly prominent class of ML models, which are also referred to as *deep neural networks*. Such models iteratively learn which features are most useful for the prediction objective (Goodfellow et al., 2016), which obviates the need for handcrafted features described in step 1. They can operate over many different types of modalities, such as natural images (Alzubaidi et al., 2021) and satellite images (Zhu et al., 2017), but also non-visual modalities such as textual information (Khurana et al., 2023). A *Convolutional Neural Network* (CNN) is a feed-forward, hierarchical DL model architecture that is adapted to handle image data. The main feature-learning layer in these models are *convolutional* layers, which consider features over a local neighbourhood of pixels, structured as a matrix. They learn the weights of such convolutions, or filters. Early layers extract simplistic patterns such as colour transitions or edge detectors. By applying non-linearities and further recombinations of features, deeper layers gradually refine these features to learn increasingly complex features, such as the interaction between trees and their surrounding landscapes. Many of the essential theories for DL models were proposed decades ago, such as the fundamental linear layer (Rosenblatt, 1958), the automated updating of model layers through backpropagation (Linnainmaa, 1976), and the convolutional layer for images (Lecun et al., 1998). However, due to the computational cost of CNNs, they were not frequently used. Once CNN operations were implemented on graphics processing units and a sufficient amount of reference data became available (Krizhevsky et al., 2012), they became a popular choice of model for learning from image



Figure 1.2: Comparison of image modalities used in this thesis. Natural images (left, courtesy of Philip Halling (Halling, 2011)) capture fine detail about the subject (the O2 Arena in London), while the Sentinel-2 satellite image (Copernicus Project, 2023) (right) gives a broad overview of the surroundings (the Thames area of London).

modalities. Recent DL models are successfully applied to a wide variety of topics, such as wildlife conservation (Tuia et al., 2022), urban streetscape analysis (Biljecki and Ito, 2021), and earth systems (Reichstein et al., 2019).

1.2.2 Data modalities for LQ assessments

ML and DL models have also been used for a variety of LQs and with a variety of image types, or *modalities*. There are two main types of modalities considered in ML-based LQ assessments. The first type of modality is *natural images*, which are photographs typically taken with a ground-level perspective and with colour bands (red-green-blue, or monochrome) that appear natural to humans. As such photos naturally convey what humans can see, it is possible to use volunteers to provide impressions for them.

This approach has been used to measure the perceived safety of streets (Naik et al., 2014), the perceived quality of facades (Law et al., 2018), and the scenicness of landscapes (Seresinhe et al., 2015). The second type of modality that is commonly used is *remote sensing* (RS) imagery. Remote sensing images are taken from an overhead perspective and may use spectral bands beyond the spectrum of colours visible to the human eye, such as those in the infrared spectrum. The main benefit of remote sensing imagery is the amount of spatial coverage it provides, meaning that it can be used to assess large areas at once. The drawback of using RS is that certain landscape elements cannot be seen from above, such as the façades of buildings or the interplay between landscapes, such as the picturesque lake pictured in front of a mountain. LQ assessments using this modality make use of aggregate spatial scores, such as the vitality of neighbourhoods at the block level (Scepanovic et al., 2021) or urban deprivation using a spatial grid (Arribas-Bel et al., 2017). A visual comparison of both modalities is given in Figure 1.2.

1.2.3 Natural and urban LQ assessments

In order to train ML and DL models to reproduce LQs, a dataset of labelled examples is needed. LQ assessments often concern the study of the scenicness of natural vistas (Arthur, 1977; Wherrett, 1998; Nahuelhual et al., 2018). This is a well-studied task with datasets that are available to the public. The ScenicOrNot dataset (SON) is a dataset of rated landscape images across the entirety of the United Kingdom gathered through crowdsourcing efforts (Seresinhe et al., 2015). It contains enough images to train DL models on (Seresinhe et al., 2017), and it has been extensively tested in existing research (Marcos et al., 2019; Arendsen et al., 2020; Marcos et al., 2021). As such, this particular dataset is well-suited for studying the role of DL modelling experiments in a typical applied setting for natural environments.

Urban environments are complex and concern many different facets of life. As a result, there are many different LQs that can be measured, such as safety (Naik et al., 2014; Dubey et al., 2016), deprivation (Suel et al., 2021; Singleton et al., 2022), walkability (Christman et al., 2020), and aesthetic quality (Dubey et al., 2016; Biljecki and Ito, 2021). A LQ that encompasses many qualities is *liveability*, which is “*the degree to which its provisions and requirements fit with the needs and capacities of its members*” (Veenhoven et al., 1993). Understanding if liveability can be assessed from images can therefore help to ensure that cities are able to match the needs of their inhabitants. While no labelled image datasets exist, it is possible to combine liveability reference data with images to create labelled image datasets. In doing so, it is possible to study how well liveability can be assessed using images.

1.2.4 Combining modalities

Previous sections have considered natural and remote sensing images separately from one another. In reality, they are complementary, as they describe different perspectives of the same landscape, and both modalities contain LQ information that is not visible to the other modality (Gómez-Chova et al., 2015). For instance, façades contain information about the state of maintenance of a building (Law et al., 2018), which is not visible in overhead aerial images. Likewise, a clear view of rooftops is often not available from ground-level images. Furthermore, ground-level images taken in urban areas are often hindered by a limited viewshed. As a result, LQ assessments benefit from having multiple viewpoints involved. It is possible to learn from multiple modalities at the same time with a *multimodal* approach. Such approaches leverage the availability of multiple data sources to understand how they complement one another. Examples of multimodal challenges include learning shared features between modalities, fusing features for improved predictive performance, and co-learning, which allows for the transfer of information about the dependent variable from one modality to the other (Baltrusaitis et al., 2019). Recently, multimodal learning approaches have proven successful in the geospatial domain (Tuia

et al., 2021) and have shown significant advances in image-text retrieval (Radford et al., 2021; Bommasani et al., 2022), visual-question answering (Lobry et al., 2021; Chappuis et al., 2022), urban function recognition (Srivastava et al., 2019; Sapena et al., 2021; Fan et al., 2022; Workman et al., 2022), and urban deprivation (Suel et al., 2021).

1.2.5 Interpretability

Being able to predict and monitor LQs can help provide timely information about the state of our landscapes. However, DL models are notoriously opaque in their workings, as they may base their decisions on abstract patterns that humans may not be able to understand. This lack of transparency and a lack of explanations given for predictions can undermine trust in models (Miller, 2019). Furthermore, the black box nature of ML models makes it difficult to discover new relationships from them, potentially limiting the creation of new knowledge from studies (Gevaert, 2022). To overcome these shortcomings, studies have attempted to improve the *interpretability* of deep learning models. There is no strict definition of the term explainability, but it is generally understood to be “*the degree to which an observer can understand the cause of a decision*” (Biran and Cotton, 2017; Miller, 2019; Gevaert, 2022). This can be achieved in many ways and with many different intentions. For instance, methods such as *class activation maps* (Zhou et al., 2016) and *integrated gradients* (Sundararajan et al., 2017) estimate the importance of image regions or individual pixels. Other models seek to determine which concepts are important for the task that is being predicted (Kim et al., 2018). For the case of RS imagery, there is an increased emphasis on the importance of *domain knowledge*, where the aim is to integrate existing knowledge into ML approaches (Roscher et al., 2020; Gevaert, 2022). For instance, prediction output ranges may be constrained to established physical models, where they may be used to find model-observation mismatches or to constrain models (Reichstein et al., 2019). Despite a growing selection of interpretability methods, the consensus remains that current interpretability methods are inadequate at explaining predictions in a way that is sufficient for scientific knowledge extraction (Miller, 2019; Roscher et al., 2020), as well as existing and upcoming regulatory frameworks, such as the European AI Act (Gevaert, 2022).

In the context of LQs, a mix of methods has previously been used in order to understand the prediction patterns of models. Two prominent approaches emerge in the current literature. Firstly, post-hoc interpretation methods are prominently used for both natural images and RS imagery. On natural images, post-hoc methods are commonly used to relate LQ predictions to objects seen in the image. For instance, urban perception factors can be related to the presence of objects in urban spaces (Zhang et al., 2018; Zhang et al., 2019; Qiu et al., 2022). Research using RS imagery often relates predictions to spatial datasets, such as demographics. For instance, local climate zones have been post-hoc compared to quality-of-life factors (Sapena et al., 2021), and urban vitality has been related to neighbourhood characteristics (Scepanovic et al., 2021). A second

prominent direction of interpretability methods concerns *intrinsically interpretable* models (Marcos et al., 2019; Koh et al., 2020; Gevaert, 2022). The intent is to design model architectures that give explanations as part of their design, rather than to implement interpretability approaches only once a model has already been trained. One such class of models is *semantic bottlenecks* (Marcos et al., 2019; Nguyen et al., 2022), also referred to as *Concept Bottlenecks* (Koh et al., 2020). These models use a set of predictive tasks related to the detection of the presence of concepts understandable by humans and then use them as a starting point to constrain the prediction of a final dependent variable. For instance, a landscape is made of distinct objects (trees, rocks, and a lake), materials, and textures. Deep learning models have been trained to learn the relation between objects, concepts, and scenicness (Marcos et al., 2019), groupings of objects, concepts, and scenicness (Marcos et al., 2021), and to discover relevant concepts (Arendsen et al., 2020). However, models with intrinsic interpretability have not been attempted yet for RS imagery for LQ assessments.

1.3 Research gaps

Landscapes are an important resource for human activities and ecosystems alike. Studying the qualities they exhibit can aid in understanding how they are used. While current research has established that LQs can be predicted from both natural and remote sensing images using ML methods, there are several issues limiting the impact and practical applicability of such approaches.

There is a notable lack of interpretability methods tailored for DL methods, in particular for RS imagery (Tuia et al., 2021; Gevaert, 2022). The lack of existing methods makes it difficult to discover new knowledge from studies that use DL methods. This lack of interpretability limits the amount of new insights that can be acquired, which makes them less attractive for end-users. While existing research has considered the interpretability of LQ assessments from natural images, more research is needed to determine which interpretability methods are suitable for assessments involving RS images.

Recent research has proven that multimodal approaches can result in significant performance improvements and new insights, whether fusing image modalities (Gómez-Chova et al., 2015), or combining image modalities with other modalities, such as text (Lobry et al., 2021). Natural images with location information are abundant in the form of Google Street View imagery, as well as through social media platforms, which may be combined with RS imagery to cover the weaknesses of each modality (Munoz et al., 2021; Zhu et al., 2022). However, few LQ assessment studies have attempted to use multimodal approaches, and as such, it is not yet clear which modalities should be used and which benefits they may bring.

Lastly, **LQ prediction using DL models, using either natural images or RS imagery, depends on the availability of datasets with thousands of examples.** As a result, any study attempting to measure new LQs will require a substantial data collection effort and may have latent biases due to a lack of oversight into the many images and labelers used to create the datasets (Gebru et al., 2021), which gets exacerbated by processes such as crowdsourced data collection (Deng et al., 2009). This is especially the case for recent models trained on captioned images scraped from the internet, which are often comprised of hundreds of millions of examples (Radford et al., 2021). The dependence of DL models on large datasets also makes it more difficult to undertake small-scale studies with fewer participants, which are typical for applied LQ research (Jacques, 1980; Wherrett, 1998). As a result, there is a considerable mismatch between DL methods and applied LQ research regarding dataset requirements. To bridge this gap, studies are needed into data-efficient training regimes for the DL-based prediction of LQs.

1.4 Objectives

While previous research has proven that assessing LQs through DL methods is feasible, several issues are limiting their practical usefulness. Large datasets are needed in order to train models, and their prediction patterns are difficult to understand. As such, the objective of this thesis is to study and address these shortcomings through the following four research questions:

- RQ 1:** Which patterns can be modelled and reproduced through DL-based LQ assessments?
- RQ 2:** How can LQ assessment workflows using DL be made more interpretable so that it is easier to acquire new knowledge?
- RQ 3:** What are the benefits and challenges of multimodal DL approaches for LQ assessments?
- RQ 4:** Which approaches are effective at reducing the dependence on large datasets for LQ assessments using DL models?

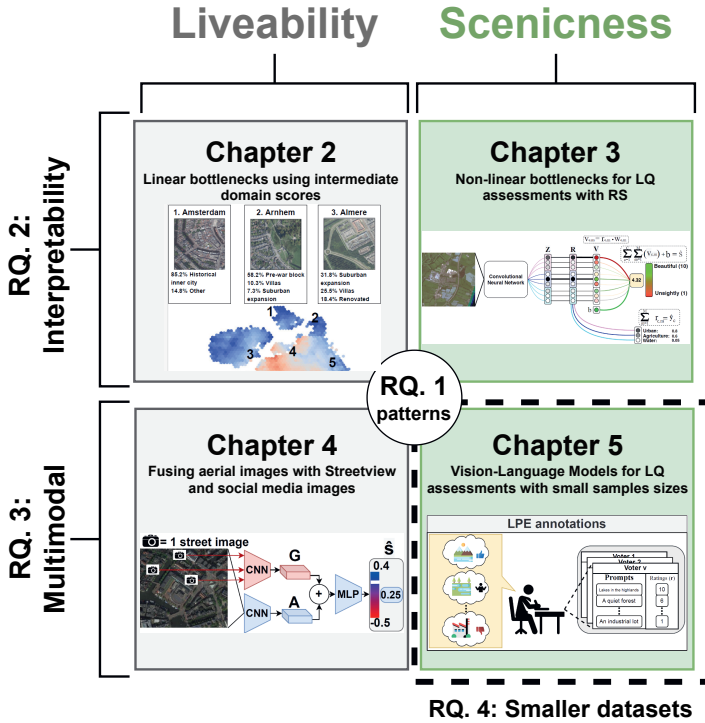


Figure 1.3: Conceptual framework of the thesis. The thesis first presents two chapters on interpretability (RQ. 2), and the remaining two are about multimodal research (RQ. 3). RQ.1 is studied in all research questions, while Chapter 4 is dedicated to RQ. 4.

1.5 Contributions

The research questions posed in the previous section are answered through four research papers, which are presented as chapters in this thesis. Figure 1.3 provides a graphical overview of how each chapter interacts in a conceptual framework. The first two chapters of the thesis aim to better understand the task of LQ assessment using RS images. The latter two chapters explore multimodal approaches to explore their performance potential, as well as multimodal co-learning in order to use smaller datasets.

Chapter 2 introduces the task of liveability prediction from aerial images for the entirety of the Netherlands. Firstly, it assesses the feasibility of liveability prediction using DL methods from RS imagery (**RQ. 1**). This chapter also considers the interpretability of models trained on RS data. For this purpose, an *interpretable-by-design* linear semantic bottleneck model is designed that uses concept classes that contribute to liveability.

Experiments are also performed for the post-hoc interpretation of the model (**RQ. 2**). This chapter uses contents from the following publication:

Levering, A., Marcos, D., van Vliet, J., Tuia, D., 2023. Predicting the liveability of Dutch cities with aerial images and semantic intermediate concepts. *Remote Sensing of Environment* 287, 113454. <https://doi.org/10.1016/j.rse.2023.113454>

Chapter 3 studies how well scenicness can be predicted from an overhead perspective (**RQ. 1**), and it addresses the lack of interpretation methods tailored for LQ assessments from RS images (**RQ. 2**). It extends the use of semantic bottlenecks for RS data to the prediction of scenicness from satellite imagery, and improvements are made to their ability to model complex patterns. The architecture uses land cover as an intermediate concept, which makes it possible to study how land cover classes relate to scenicness. This chapter was published as follows:

Levering, A., Marcos, D., Tuia, D., 2021. On the relation between landscape beauty and land cover: A case study in the U.K. at Sentinel-2 resolution with interpretable AI. *ISPRS Journal of Photogrammetry and Remote Sensing* 177, 194–203. <https://doi.org/10.1016/j.isprsjprs.2021.04.020>

In **Chapter 4**, a multimodal learning approach for housing quality is attempted. Features extracted from natural images are fused with aerial overhead images in order to test how well housing quality at the grid level can be predicted for the city of Amsterdam (**RQ. 1**). It considers the performance benefits that multimodal approaches to natural and remote sensing images may bring (**RQ. 3**). This work has been published as follows:

Levering, A., Marcos, D., Havinga, I., Tuia, D., 2021. Cross-Modal Learning of Housing Quality in Amsterdam, in: *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GeoAI 2021*, pp. 1–4. <https://doi.org/10.1145/3486635.3491067>

In **Chapter 5**, several strategies for data-efficient workflows are tested for the task of scenicness prediction from natural images. Firstly, methods are tested for learning and predicting using small labelled image datasets (**RQ. 4**). Secondly, a small dataset consisting of text descriptions of scenicness is gathered from volunteers and used to predict scenicness as an alternative to image ratings, which tests the use of multimodal approaches for the creation of new datasets (**RQ. 3**), as well as the use of text as a modality for scenicness prediction (**RQ. 1**). This chapter has been submitted as follows:

Levering, A., Marcos, D., Tuia, D., Jacobs, N., 2023. Prompt-guided and multimodal landscape scenicness assessments with vision-language models, *PLOS One*.

Chapter 2

Predicting the liveability of dutch cities with aerial images and semantic intermediate concepts

This chapter is based on:

Levering, A., Marcos, D., van Vliet, J., Tuia, D., 2023. Predicting the liveability of Dutch cities with aerial images and semantic intermediate concepts. *Remote Sensing of Environment* 287, 113454. <https://doi.org/10.1016/j.rse.2023.113454>

Abstract

In order to provide urban residents with suitable living conditions, it is essential to keep track of the liveability of neighbourhoods. This is traditionally done through surveys and by predictive modelling. However, surveying on a large scale is expensive and hard to repeat. Recent research has shown that deep learning models trained on remote sensing images may be used to predict liveability. In this paper we study how well a model can predict liveability from aerial images by first predicting a set of intermediate domain scores. Our results suggest that our semantic bottleneck model performs equally well as a model that is trained only to predict liveability. Secondly, our model extrapolates well to unseen regions (R^2 between 0.45 and 0.75, Kendall's τ between 0.39 and 0.57), even to regions with an urban developmental context that is different from areas seen during training. Our results also suggest that domains which are directly visible within the aerial image patches (physical environment, buildings) are easier to generalize than domains which can only be predicted through proxies (population, safety, amenities). We also test our model's perception of different neighbourhood typologies, from which we conclude that our model is able to predict the liveability of neighbourhood typologies though with a varying accuracy. Overall, our results suggest that remote sensing can be used to extrapolate liveability surveys and their related domains to new and unseen regions within the same cultural and policy context.

2.1 Introduction

The living standards of a neighbourhood may have a significant effect on the health of its residents. Residents of destitute neighbourhoods are prone to several health risks, such as increased morbidity rates (Barber et al., 2016), mortality rates (Haan et al., 1987), and worse dietary and physical activity patterns (Thompson and Kent, 2014). Similar patterns are observed for housing, where lower-quality housing also results in worse mental well-being (Evans, 2003). As such, it is important to monitor the wellbeing of a neighbourhood for the benefit of urban residents. For this purpose, researchers have studied how factors relate human wellbeing to their living environments using the *liveability* framework. The liveability of a society can be understood as *"the degree to which its provisions and requirements fit with the needs and capacities of its members"* (Veenhoven et al., 1993). In the context of living environments, examples of the needs and capacities required may be housing that is of adequate size and quality for its residents, provision for adequate travel to work, and sufficient green space in the neighbourhood. Research has since advanced the theoretical underpinnings of liveability research. Kamp et al. (2003) argue that a conceptual framework of liveability would *"allow for a more theory-based choice of indicators, and for the development of tools to evaluate multidimensional aspects of urban environmental quality"*. The leefbaarometer project (referred to as *LBM* hereafter) initiated by the Dutch government (Leidelmeijer et al., 2014) follows up on that suggestion. The LBM project was set up to survey the liveability of neighbourhoods across the Netherlands, and to subsequently model the liveability using variables that can be applied nation-wide, such as housing quality and greenspace proximity. In doing so, the authors assess which variables are relevant for liveability on a nation-wide scale. Linking such survey data to empirical and statistical data may improve our understanding of what makes cities liveable. However, a notable drawback to using manually collected data, such as surveys, is the difficulty of upscaling and repeating results.

Remote sensing methods have long been used to extract intermediate variables for liveability prediction, such as the prediction of urban greenery (Jensen et al., 2004; Li and Weng, 2007; Rahman et al., 2011), rather than the prediction of liveability directly from imagery. Studies attempting to recognise the qualities of cities have considered various intermediate variables, such as urban morphology (Taubenböck et al., 2012; Rodriguez Lopez et al., 2017; Tian et al., 2022), local climactic conditions (Bechtel et al., 2015; Qiu et al., 2019; Liu and Shi, 2020), and urban land use (Srivastava et al., 2019; Rosier et al., 2022). Recent advances in machine and deep learning have enabled research that predicts liveability variables directly from overhead imagery. Remote sensing models have the benefit of high scalability and better monitoring in regions with poor data availability (Kuffer et al., 2020, p. 18). In regions with greater data availability, much research has gone into hedonic housing pricing as a means of predicting the attractiveness of neighbourhoods. Hedonic housing pricing attempts to capture the value of a property based on its intrinsic value

as well as external factors affecting it. The main value of remote sensing for hedonic pricing is the inclusion of contextual information about the immediate and larger area of surroundings (Bency et al., 2017, p.5). Yao et al. (2018), for example, fuse remote sensing imagery with social media data to predict housing prices in Shenzhen, China, with highly accurate results.

Recent studies have attempted to directly predict variables relating to liveability in countries with high data availability. Arribas-Bel and colleagues trained machine learning models to recognise living environment deprivation from high-resolution aerial images over the city of Liverpool in the United Kingdom (Arribas-Bel et al., 2017). Singleton et al. (2022) use an autoencoder model to extract features describing Sentinel-2 satellite image tiles of neighbourhoods across the UK. These features were clustered to form neighbourhood typologies and subsequently related to urban deprivation data. However, the clustered neighbourhood representations proved insufficient to explain urban deprivation. Suel and colleagues study income, overcrowding, and environmental deprivation using a multimodal approach, using both Google Street View and 3m resolution Planet satellite images over the Greater London region (Suel et al., 2021). Their findings confirm that high-resolution aerial images on their own can approximate the trend of urban deprivation at the neighbourhood level. Scepanovic et al. (2021) use Sentinel-2 image tiles to predict the vitality (presence of people throughout the day) of Italian cities at the district level through several experiments. The authors predict six physical descriptors of urban form relating to land use and block size from Sentinel-2 image patches across Italian districts and infer their usefulness for predicting vitality. This first experiment showed limited accuracy, most likely due to the resolution of the Sentinel-2 image tiles. In their second experiment, the authors predict urban vitality (as measured by mobile internet usage) directly from Sentinel-2 image features and the capacity of models to generalise between cities. Their results indicate that generalisation of urban vitality is possible, but generalizing their model to Rome resulted in a notable decrease in accuracy, as it is historically, culturally, and naturally distinct from the other cities within their dataset. Huang and Liu (2022) use a deterministic approach to model the liveability of 101'630 communities in China in 42 major cities, guided by expert decisions. A total of 27 liveability factors are extracted using high-resolution satellite imagery and subsequently weighted according to expert opinions. Their work presents the first large-scale assessment of the liveability of urban communities in China.

Previous work has attempted to study remotely sensed liveability by observing a limited number of components relating to liveability at a time and without taking into account surveyed resident opinions. In doing so, they have confirmed that individual liveability factors such as income, environmental deprivation, and block size can be suitably predicted through optical remote sensing. Yet, it is unclear to what extent different domains relating to liveability can be predicted from remote sensing imagery. Therefore, in this paper, we study how well different domains of liveability may be predicted from high-resolution aerial imagery on a neighbourhood scale. We set out to determine the suitability of remote

sensing for interpolating and extrapolating large-scale inventories of liveability. Moreover, we explore how a model with a semantic intermediate layer compares to a model that only predicts liveability. Specifically, we compare how the liveability prediction as a linear combination of domain scores compares against a direct prediction of liveability. Lastly, we evaluate how well liveability domains can adapt to unseen geographical contexts as well as building typologies. We formulate and address two research questions for our research:

1. How well can we predict different domains of liveability?
2. How well does a bottleneck model predict compared to an unconstrained model?

The remainder of the paper is as follows: In section 2 we present the dataset used in our study and our model architecture. In section 3 we present the metrics and maps for our experiments. Lastly, in section 4 we reflect on our results and their relevance for liveability monitoring.

2.2 Material and methods

We are interested in training a deep learning model to predict liveability on a neighbourhood scale by first predicting domain-specific liveability contribution scores as a set of interpretable semantic intermediate concepts. For this purpose, we use a semantic bottleneck model (Marcos et al., 2021; Koh et al., 2020), which uses an intermediate linear layer with semantic concepts, which are then used to predict a final objective. For this purpose, we need a dataset of overhead aerial images, neighbourhood-scale labels of liveability, and a deep learning model architecture that can first predict individual domain scores and then regress the overall liveability score through the domain-specific scores. We discuss these requirements in order.

2.2.1 Dataset design

To train our model, we require a labelled dataset of liveability scores and overhead aerial imagery (Figure 2.1). Additionally, we make use of a series of domain scores, which decompose the liveability score into a series of explainable aspects. To build this dataset, we use nationally available data sources in the Netherlands. Specifically, we consider 13 built-up areas of varying sizes, ranging from village (Beesel) to metropolis (Amsterdam). Selected built-up areas are listed in Table 2.1.

Liveability reference data

The reference data for liveability used in our research is made available by the leefbaarometer (*LBM*) project (Leidelmeijer et al., 2014), an ongoing liveability monitoring project initiated by the Dutch government. For this purpose, the authors collected a dataset with over 100

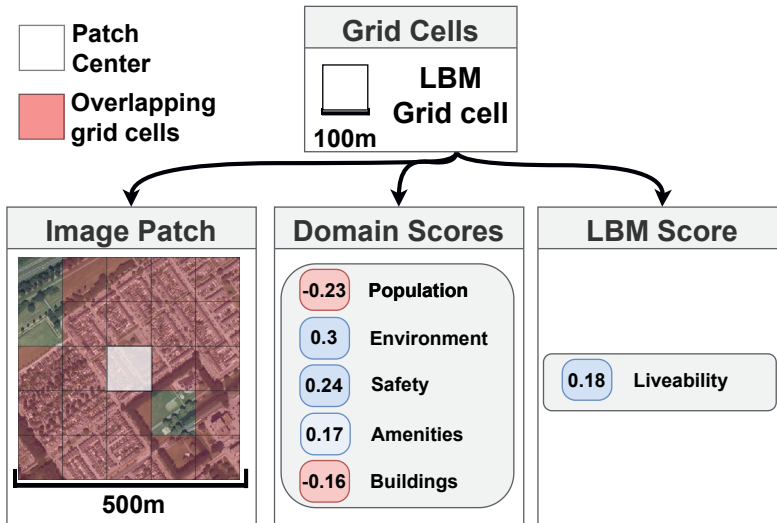


Figure 2.1: Workflow for generating our reference data. From the LBM dataset, we extract the domain scores and the final liveability score. The domain scores are a decomposition of the liveability score, which reflects how each domain contributes to the overall liveability of a grid cell. For our image patches, we use the grid cell as the centre for a 500 by 500-pixels patch at 1 metre resolution. The 400 by 400-metre overlap with other grid cells ensures that the patch size is equal to the spatial sum operation that was performed for the original variables of the LBM dataset.

variables for use in regression models to predict liveability. These variables are available for all neighbourhoods in the Netherlands at the scale of an individual street. Where applicable, variables are summed over a radius of 200m around each neighbourhood to reduce the occurrence of outlying neighbourhoods with few respondents in the dataset. The input variables can be assigned to five domains. The following broad groups of variables are considered for each domain:

- **Population:** Welfare factors, age groups, residuals for family composition, and ethnic composition after controlling for income
- **Physical Environment:** Green/grey area descriptors, proximity to water/green areas, proximity to nuisances (e.g., trains/roads)
- **Safety:** Number of occurrences for several broad crime categories
- **Amenities:** Amenities within 1-20km distance, e.g., cafes, hospitals, schools
- **Buildings:** Building age groups per 10 years after 1900, ownership status, simple typology descriptors e.g., pre/post-war

For a complete description of all 100 variables used by the LBM project we refer the reader to Table 7.1 of the documentation (Leidelmeijer et al., 2014, p.91). In the discussion, we

elaborate on the use of stigmatising variables for the population domain score and the problems arising from it.

These 100 variables belonging to five domains are then used as the input for two linear regression models. The first model regresses the surveyed resident liveability opinions. Respondents were asked three questions about their satisfaction with their living situation and asked to answer on a scale of 1–5 for each question, where 5 is "most satisfied". The average of these three questions is used as the response variable for the first regressor, after correcting for the age of residents. The second model uses a hedonic pricing approach to estimate housing prices for a neighbourhood derived from nationally available property value estimates. From these two linear regression models, each neighbourhood is assigned the averaged z-score of these models as the single overall liveability score, shown on the right side of Figure 2.1. Hereafter, we will refer to this averaged z-score simply as the *liveability score*. By grouping the 100 variables into five domains and averaging their coefficients, the contribution of each domain to the overall change in z-score can also be computed for each domain. We refer to these grouped scores as *domain scores*. The five domain scores are fundamentally different in nature. Some domain scores can be observed directly from aerial images. This group consists of the *buildings* and the *physical environment* domain scores. We refer to these scores as *direct scores*. The other three domain scores cannot be observed from aerial images but should instead be predicted by proxy correlations. For instance, for the *Population* domain score, the model could learn that large single-family houses generally have a more affluent population, thereby learning a correlation as a proxy for the prediction of the domain score. We refer to these domain scores (*Population*, *Safety*, and *Amenities*) as *indirect scores*.

The veracity of the outcomes of the LBM project was verified through interactions with policymakers. For all of the 13 built-up areas considered in this research, the results truthfully reflected the general liveability trends (Leidelmeijer et al., 2014, p.100). The liveability score and the domain scores are re-predicted bi-yearly from 2014 onwards. For privacy reasons, the dataset could not be made available at street level. Instead, all variables and scores are made available at a resolution of 100 metres through a gridded dataset. We use the grid cells made available in this research as the basis for our dataset, for both their spatial extent and as reference data.

Neighbourhood liveability patches

We use the gridded dataset provided by the LBM project as the starting point for our dataset of neighbourhood liveability patches. We use the liveability scores made available for the year 2016. We do so first because it is the closest year to which there is a nationally available aerial image (2016). In total, we use 51'781 grid cells from the dataset over the 13 built-up areas within our dataset. The samples used from each built-up area are shown

Table 2.1: Samples per split and municipal population census numbers for each built-up area. Population data is derived from the Dutch statistics agency (CBS, 2016)

Built-up area	Training	Validation	Testing	Population (2016)
Almere	1'856	1'206	-	198'145
Amsterdam	7'116	2'609	-	833'624
Arnhem	3'713	722	-	153'818
Beesel	-	-	388	13'388
Dordrecht	-	-	3'548	118'801
Eemsdelta	607	238	-	47'080
Eindhoven	-	-	6'490	224'755
Groningen	2155	718	-	200.952
Hengelo	-	-	3'034	81'075
Nijmegen	3'071	1'068	-	172'064
Rotterdam	8'439	1'823	-	629'606
Venlo	1'074	664	-	100'371
Weert	1'008	234	-	49'100
Total	29'039	9'282	13'460	-

in Table 2.1. We use the five domain scores and the overall liveability scores (middle and right columns of Figure 2.1 respectively) as the liveability labels of our patches.

For the overhead aerial imagery, we use images from the national composite aerial image from 2017, made available by the Dutch government (PDOK, 2017). The original composite image is available at $0.1m$ resolution with four bands (red, green, blue, and near-infrared (NIR)) and is entirely cloud-free. We do not perform additional pre-processing steps such as geometric correction, as this has already been done by the data provider. We downsample the pixel size to $1m$.

Beyond determining how well liveability can be predicted, we are interested in monitoring it over multiple timesteps. However, high-resolution imagery available for past years does not have NIR information. To ensure the compatibility of our analyses with historical aerial image data in the Netherlands, we exclude the NIR band from our main analyses. In future work, we will explore the feasibility of time series mapping for liveability. However, we study the effect of adding the NIR band to our liveability prediction model in the results and discussion section, where we report the numerical results for a model trained on all four bands.

As some LBM variables are summed over a radius of $200m$ around the grid cells, the square patch size should cover at least 500 by 500 metres such that it approximates the extent of the LBM grid cell centres. As such, we extract patches of 500 by 500 pixels centred on each grid cell. As a result of the image patch being larger than the 100 by 100-metre LBM grid cells, there is an overlap with the 24 neighbouring aerial image patches.

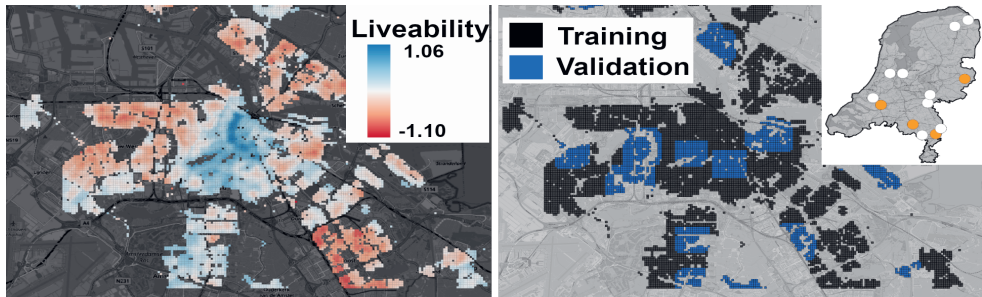


Figure 2.2: **Left:** Liveability scores over Amsterdam ranging from -1.10 (lowest, red) to 1.06 (highest, blue). **Right:** Example of data splits. Grid cells marked with dark grey are used during training. Blue grid cells are used for validation. In the top right, we show built-up areas that are considered. Areas marked with white points are used for training and validation, while areas marked with orange points are used only during inference.

Data splitting

We use data from nine built-up areas for training and validation. Within each area, we create square blocks of patches for validation, and we assign the rest to the training set. Through the overlap with neighbouring patches, some of the validation set is seen during training. However, this was not found to result in issues with generalisation during testing. We use the remaining four built-up areas as an independent test set. The four cities were chosen for their geographic diversity and size. Dordrecht is proximate to Rotterdam, and it is part of the *Randstad* area, which is the largest conurbation of the Netherlands. As Amsterdam and Rotterdam are part of the training dataset, Dordrecht is therefore the most similar city in the test set. Eindhoven and Hengelo are both cities that follow a different development pattern compared to those in our training split. Both cities began to develop significantly as a result of industrialization, which makes them developmentally distinct from the cities in our training split. This difference in developmental context allows us to study how well our model adapts to unseen developmental layouts. Lastly, Beesel is a small village along the German border, which tests the model’s ability to transfer to smaller settlements (as Beesel is the only village in the training dataset), and to remote regions. We show an example of our training and validation set stratification for the municipality of Amsterdam in Figure 2.2. We show the number of samples per split in Table 2.1.

2.2.2 Bottleneck CNN for liveability prediction

In this section, we present the interpretable bottleneck model used to predict liveability from overhead aerial images. We use a two-step approach to predicting the overall liveability score of an area. To obtain a transparent and interpretable prediction of liveability that is concordant with the design of the LBM scores, we use a semantic bottleneck design

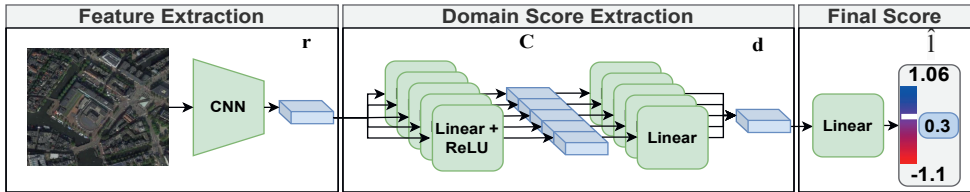


Figure 2.3: Architecture of our model. Using a CNN, we first extract a vector of features \mathbf{r} . We then construct the rows of the feature matrix \mathbf{C} , where each row is a feature vector \mathbf{C}_i that is specific to one domain score. Each feature vector is then used to compute a domain score d_i . Finally, the domain score vector \mathbf{d} is used to compute the overall patch liveability score \hat{l} .

(Marcos et al., 2021; Levering et al., 2020). A semantic bottleneck forces the prediction of a final layer to be interpretable by first predicting a set of semantic concepts, which are then linearly re-combined to predict the target variable. We chose this type of architecture because the LBM is, by design, a combination of the five domain scores. We can therefore use the semantic bottleneck to enforce the prediction of the liveability to be a linear combination of the domain scores, and this mimics the logic of the original LBM model. As such, our model is tasked with predicting concepts as a vector of domain-specific sub-scores, which we denote with \mathbf{d} . These domain scores are then used in a linear layer with a bias term to regress the predicted patch liveability score l . Our architecture is shown in Figure 2.3.

Our model is first tasked with extracting relevant features for the prediction of liveability. The feature extractor takes the aerial image patch as input and produces a global feature vector \mathbf{r} . We use a standard convolutional neural network feature extractor for this purpose. Using this global feature vector \mathbf{r} , we then predict a liveability domain score for each of the $i \in \{1 \dots D\} \in \mathbb{N}$ domains being considered. These liveability domain scores describe the contribution of different domains to the overall liveability of a place in explainable aspects, such as *amenities* and *safety*. The domain scores correspond to the domain scores presented in the middle columns of Figure 2.1. To predict the domain scores, we use a two-layer Multi-Layer Perceptron (MLP) to create each row of the feature matrix \mathbf{C} . The first linear layer recombines the extracted features into a 250-dimensional vector, which is activated by a ReLU non-linearity. Notice that this feature vector \mathbf{C}_i represents a summary of the features as they are relevant for each domain, which we leverage when interpreting the model’s propositions in Section 2.3.2. The second layer of each MLP uses the feature vector \mathbf{C}_i to regress the domain-specific liveability sub-score, which are the scores in the middle column of Figure 2.1. We then concatenate all of the liveability domain scores to form the domain score vector \mathbf{d} . From the liveability domain score vector \mathbf{d} (plus a bias term), we then directly regress the overall liveability score \hat{l} . In doing so, we enforce that the overall scenicness is only predicted by the linear combination of domain scores rather than spurious correlations that the model may pick up from the aerial images.

As the domain scores are predicted as an intermediate task in our model, we can assess their accuracy to determine how well liveability domain scores can be predicted (research question 1).

Our model is trained using a combination loss of the domain score losses and the liveability score loss. The domain score loss is given as the sum of the mean squared errors over all of the domain scores w.r.t. their reference score \hat{d}_i :

$$\mathcal{L}_{domain} = \sum_{i=1}^D (\mathbf{d}_i - \hat{\mathbf{d}}_i)^2 \quad (2.1)$$

The loss of the liveability score is the mean squared error w.r.t. the reference score \hat{l} :

$$\mathcal{L}_{final} = (l - \hat{l})^2 \quad (2.2)$$

Finally, we combine both scores to create the overall loss to propagate:

$$\lambda \mathcal{L}_{domain} + \mathcal{L}_{final} \quad (2.3)$$

where λ is a weighting term set empirically to regulate the importance of the domain scores compared to the liveability score prediction.

2.2.3 Set-up

Our feature extraction model is a ResNet-50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) from which we remove the final fully-connected layer. Our model is trained on a single NVIDIA TitanX GPU with a batch size of 20. We optimise our models using the AdamW optimiser (Loshchilov and Hutter, 2019) with an initial learning rate of $5e^{-5}$ and weight decay rate of $1e^{-4}$. We train our model for 15 epochs. To prevent that the model learns a set of features unrelated to the domain scores in feature matrix C , we set the weighting term λ of Eq. (3.8) to 100 such that the model favours the correct prediction of the intermediate task over the predictions of the final linear layer.

We assess the quality of our results with three metrics for each of the five domain scores of the LBM dataset, as well as for the overall liveability score. Firstly, we calculate the root mean squared error as a measure of error for all scores. Secondly, we calculate the coefficient of determination (R^2) to determine the quality of the fit for each score. Lastly, we compute Kendall’s τ (Kendall, 1938), which measures the ranking of neighbourhood patches. This is possible because the liveability scores in our dataset may also be interpreted as ordinal variables, in the sense that the quality of each neighbourhood can be compared to every

other neighbourhood, which represents a ranking. Kendall’s τ ranges from a perfectly inverse correlation given as -1 to a perfect correlation given as 1.

In order to assess how well a bottleneck model performs compared to an unconstrained model (research question 2), we also train an unmodified ResNet-50 model. This unmodified model is tasked with predicting the overall liveability score without the semantic bottleneck and serves as a baseline against which the bottleneck model is compared.

Lastly, we train a model on the aerial image patches with the NIR band included to assess how this affects liveability prediction. We use the same hyperparameter selection, and we initialise the network using pre-trained ImageNet weights (Deng et al., 2009). We use the weights from the red band for the NIR input channel.

2.2.4 Feature vector analyses

In order to understand our model’s perceptions of the five domain scores and the liveability score, we designed a series of feature vector analysis experiments that help explain how the model observes the different domains of liveability. We use t-SNE embeddings and neighbourhood typology data to further understand how our model observes urban spaces.

t-SNE embeddings

We assess the model’s visual perception of the different neighbourhood typologies. To do so, we perform t-SNE dimensionality reduction (Maaten and Hinton, 2008) to visualise the latent space of the feature vectors of our model. t-SNE iteratively projects the high-dimensional space into a lower number of dimensions while preserving their neighbourhood structure in the original high-dimensional space. By doing so, we can reduce the feature vectors to just two dimensions while respecting the non-linear relationships learned by the model in the original high-dimensional space. This allows us to visualise which patches are considered visually similar by the model. We perform t-SNE dimensionality reduction on the *buildings* row of the domain feature matrix, which is $C_{building}$, and the global feature vector r . We use a perplexity (balance between global and local patterns) of 100, a learning rate of 500, an early exaggeration (tendency for clusters to become compact) of 150, and we run our model for 1’000 iterations. We consider all patches in the dataset, rather than just the test set patches, in order to analyse data structures across the 13 built-up areas. We can then overlay the neighbourhood typologies of each patch for each point in the reprojected 2-dimensional space, allowing us to infer the visual homogeneity of neighbourhood typologies for that particular score.

Table 2.2: Neighbourhood typologies considered for our feature vector analyses, as defined by Kleerekoper (2016)

Typology	Period	Characteristics
Historical inner city	< 1900	3-5 layers, much concrete
Pre-war block	1900-1940	3-4 layers, moderate amount of greenery
Working-class district	1910-1940	2-3 layers, single-family houses, little to no greenery
Post-war district	1945-1990	2-3 layers, gardens, diversity in housing styles
Cauliflower district	1970-1990	Single-family housing with gardens, winding streets, lots of green
Sub-urban expansion (Vinex)	1990-present	Large diversity in housing styles
Renovated low-rise	1990-present	Neighbourhoods which have undergone renovation
Villas	All	Spacious, single houses

2.2.5 Neighbourhood typologies

The Netherlands has a long history of spatial planning and zoning, which has been extensively described and documented in official policy and literature (Ministry of Infrastructure and the Environment, 2012). Over the years, there have been many different planning philosophies intended to address the housing needs at the time. The LBM project did not explicitly take into account the neighbourhood planning styles but rather used decade-spanning building age groups. As such, the neighbourhood typologies can be considered a more complete description of the neighbourhood style compared to the age brackets of the LBM. We perform two experiments using the neighbourhood typologies. Firstly, we assess how well our model is able to perceive the liveability of neighbourhood typologies through scatterplots, which compare the predicted liveability to its reference value for each patch with a significant amount of a given typology. Secondly, as part of our feature vector analyses, we can assess how our model perceives the homogeneity of different typology styles as well as the links between certain planning styles as defined by Dutch planners. It is expected that patches with the same neighbourhood topologies would group together, as they share similar visual characteristics.

Our typology reference dataset is formally defined by Kleerekoper (2016). Here, we use a subset of 8 neighbourhood typologies, T (see Table 2.2). In our selection, we consider a variety of different design styles, the number of building layers, and construction periods. The typologies are digitised by the climate atlas of the Netherlands initiative (Kleerekoper et al., 2018). This dataset consists of district-level polygons, listing the relative presence (%) of each typology in each district. Since they cover districts, the polygons are only available at a coarser resolution than the grid cells of the LBM. To match the typological

Table 2.3: Performance difference on the test set between a model trained with only RGB information, and a model with the NIR band included.

Score	RGB-only			RGB+NIR		
	RMSE	R ²	τ	RMSE	R ²	τ
Population	0.045	0.61	0.46	0.051	0.55	0.41
Phys. env	0.049	0.61	0.41	0.05	0.69	0.51
Safety	0.089	0.61	0.50	0.078	0.68	0.47
Amenities	0.043	0.55	0.37	0.041	0.62	0.42
Buildings	0.064	0.70	0.51	0.058	0.73	0.54
Liveability	0.155	0.70	0.52	0.145	0.74	0.54

Table 2.4: RMSE scores achieved by the model within each built-up area of the test set. (Pop.=population, P.env=physical environment, Amen.=Amenities)

Region	Pop.	P.env	Safety	Amen.	Buildings	Liveability
Dordrecht	0.052	0.048	0.082	0.037	0.067	0.150
Eindhoven	0.044	0.051	0.098	0.038	0.063	0.166
Beesel	0.031	0.046	0.072	0.080	0.047	0.100
Hengelo	0.042	0.048	0.077	0.050	0.063	0.141

presence of the district level to the grid level, we use the proportion of overlap between the grid cell and each district polygon. For a given typology $t \in T$, a grid cell $g \in G$, and a set of polygons overlapping the grid cell defined as P , we calculate the proportion of each topology present as follows:

$$g_t = \sum_{p=1}^P \left(\frac{p_{area}}{g_{area}} \right) p_t. \quad (2.4)$$

2.3 Results

2.3.1 Liveability prediction

In table 2.3, we show the R^2 and Kendall’s τ metrics of both the RGB-only model and the RGB+NIR model on the test set. We show both the five domain scores and the final liveability score, which is regressed directly from the domain scores. The RGB+NIR model is shown to outperform the model with just the RGB bands on most scores, with the notable exception of the population score, where a decrease in accuracy occurs. The results show that the addition of NIR information is useful when it is available, as it may result in a better-performing model. However, historical aerial images in the Netherlands do not have NIR information. The rest of the results and discussion sections are therefore based on the RGB-only model to maintain compatibility of our analyses with future work.

Table 2.5: R^2 scores achieved by the model for each built-up area of the test set. (Pop.=population, P.env=physical environment, Amen.=amenities)

Region	Pop.	P.env	Safety	Amen.	Buildings	Liveability
Dordrecht	0.65	0.47	0.65	0.71	0.76	0.70
Eindhoven	0.66	0.62	0.66	0.57	0.76	0.75
Beesel	0.24	0.31	0.54	0.03	0.60	0.45
Hengelo	0.42	0.56	0.62	0.47	0.65	0.63

Table 2.6: Kendall’s τ scores achieved by the model within each built-up area of the test set. (Pop.=population, P.env=physical environment, Amen.=Amenities)

Region	Pop.	P.env	Safety	Amen.	Buildings	Liveability
Dordrecht	0.45	0.40	0.41	0.40	0.56	0.51
Eindhoven	0.49	0.50	0.48	0.38	0.55	0.57
Beesel	0.28	0.32	0.37	-0.02	0.39	0.39
Hengelo	0.39	0.46	0.43	0.36	0.48	0.47

In Tables 2.4, 2.5, and 2.6, respectively, we show the RMSE, R^2 , and Kendall’s τ metrics obtained by the RGB-only model for each built-up area in our test dataset. Across all regions, our model is able to infer the general trend of all scores, with some noticeable exceptions. Firstly, the achieved metrics can vary strongly per region and domain score. For instance, the model generalises far less well to Beesel, which is far smaller than the other test sites. However, the decrease in performance is dependent on the domain scores, with some scores being more affected than others.

In Table 2.7, we show the metrics for the validation validation set. We also show the difference with the test set to show the capacity of each domain score to generalise to unseen regions. Based on the decrease in metrics between the validation and the test set, our results suggest that *direct* domain scores (*physical environment* and *buildings*) are easier to generalise than *indirect* domain scores. This is mostly the case for buildings (a minor decrease in R^2 and even an increase in τ), and to a lesser extent for *physical environment*.

We show a direct comparison between our model with a semantic bottleneck and a model that is directly trained to predict liveability in Table 2.8. Our results show that the use of a bottleneck model mostly improves performance on this task. While an unconstrained model has a marginally better R^2 score, the bottleneck model outperforms an unconstrained model when considering Kendall’s τ .

Lastly, we show the spatial prediction patterns for both the overall liveability score for each test set region as well as the *buildings* domain score. In Figure 2.4, we show the predictions for the *buildings* domain score for all regions in our test set compared to the LBM labels. The patterns for the four test regions show that our model provides smooth and consistent predictions, and it is able to accurately capture the majority of the fine-grained trends. It

Table 2.7: Metrics achieved by the model on the validation set and their relative difference to metrics computed over the entire test set.

Score	R^2	% Change	Kendall’s τ	% Change
Population	0.84	−26.5%	0.66	−27.4%
Phys. env	0.87	−29.8%	0.64	−21.3%
Safety	0.84	−26.6%	0.65	−36.8%
Amenities	0.95	−41.9%	0.71	−48.9%
Buildings	0.85	−16.4%	0.68	−24.4%
Liveability	0.86	−18.3%	0.67	−22.2%

Table 2.8: Comparison of our model’s overall metrics for the liveability score to an unmodified model tasked with directly predicting liveability from aerial images. The bottleneck model matches an unmodified model in terms of R^2 and surpasses it in Kendall’s τ .

Configuration	Val R^2	Test R^2	Val τ	Test τ
Bottleneck	0.861	0.670	0.670	0.521
Baseline	0.801	0.674	0.606	0.484

is, however, frequently unable to predict very positive or very negative building quality scores. In Figure 2.5, we show the predicted liveability scores for each patch in the test regions. Again, the model predicts the general trend correctly but struggles to predict values towards either end of the distribution.

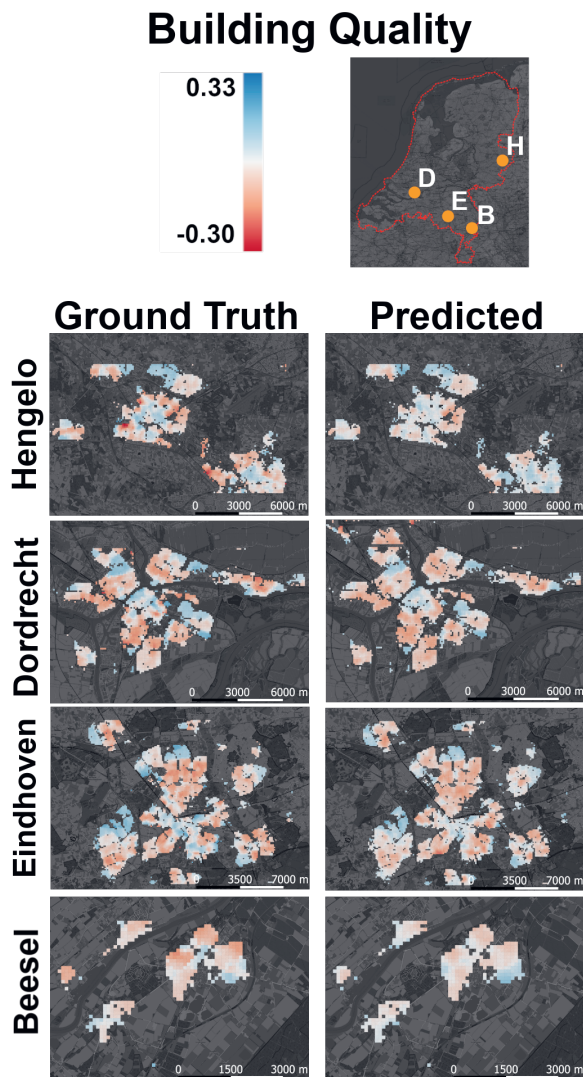


Figure 2.4: Predictions for the *buildings* domain score for all regions in the test set. Deeper shades of red represent a low building quality score, while deeper shades of blue denote high building quality. The letters on the left-hand side are the first letters of each of our test regions.

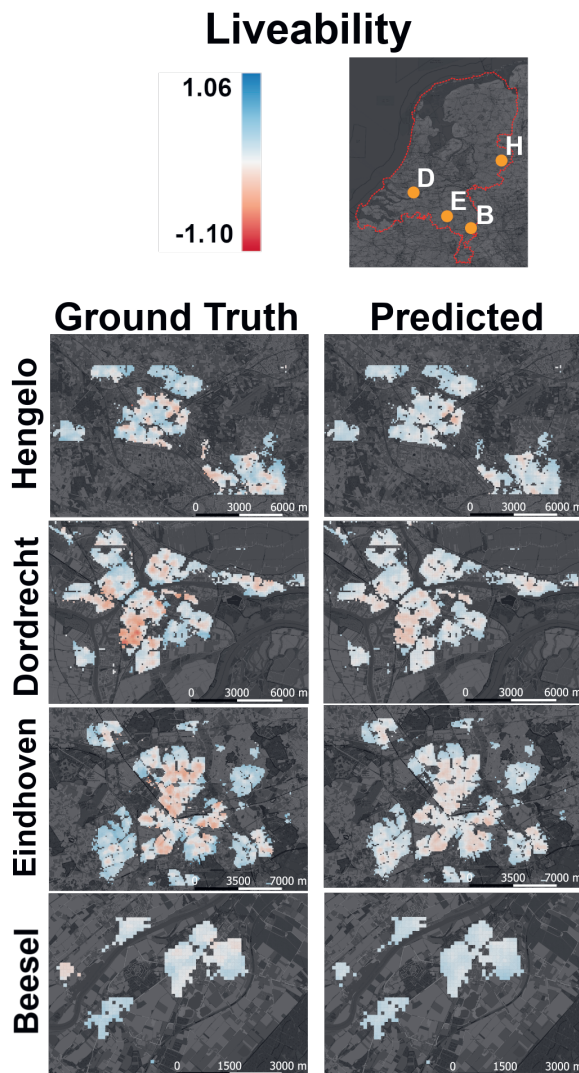


Figure 2.5: Predictions for the final liveability score for all regions in the test set. Deeper shades of red represent a lower liveability score, while deeper shades of blue denote a higher patch liveability score.

2.3.2 Feature vector analyses

Neighbourhood typologies

In Figure 2.6, we show the predicted distribution of scores for each of the neighbourhood types. For each of the selected typologies, we show the building quality prediction

distributions over the test set. Patches are included when there is 20% or more of the given typology present within the neighbourhood. From these graphs, we show that our model approximates the trend well for most typologies in our unseen test regions and with similar accuracy.

In Figure 2.7, we show the same plot for the overall liveability score. Trends emerge when comparing the scatterplots for the building quality score to the plots of the overall liveability score. A notable difference is that the model is able to better predict the overall liveability trend of the working-class districts, while it struggles to predict the housing quality of these neighbourhoods in the unseen test regions.

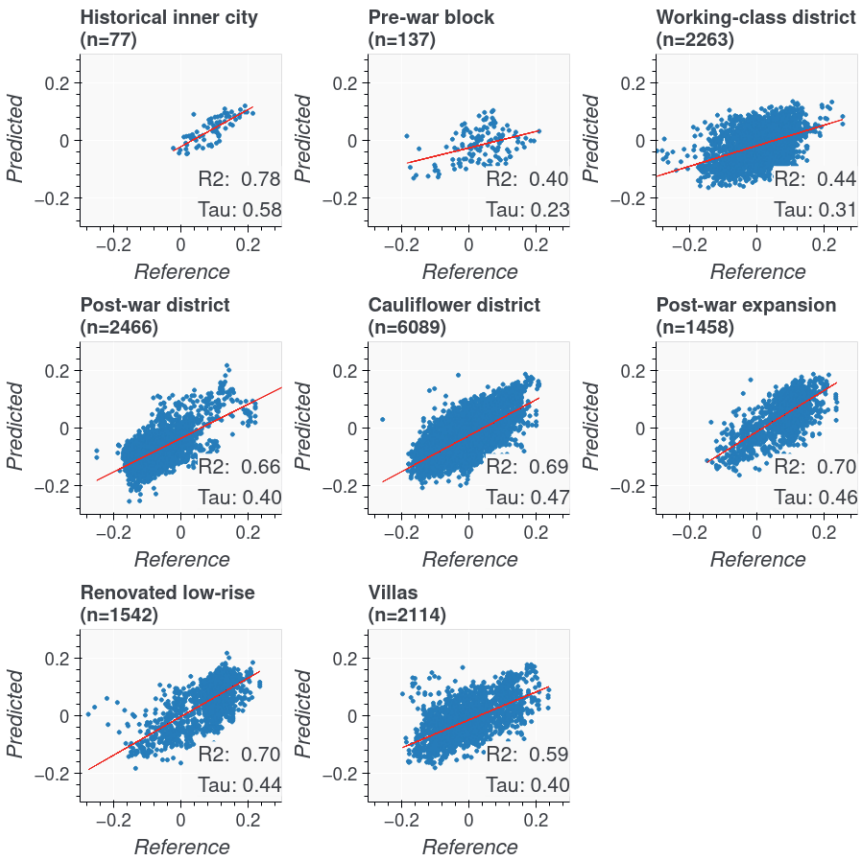


Figure 2.6: Scatterplots of the *buildings* domain score for all patches in the test set for each of the neighbourhood typologies considered in this research. Patches are included in a scatterplot when there is 20% or more of the given typology present. We show the reference value of each point on the x-axis and the predicted value on the y-axis.

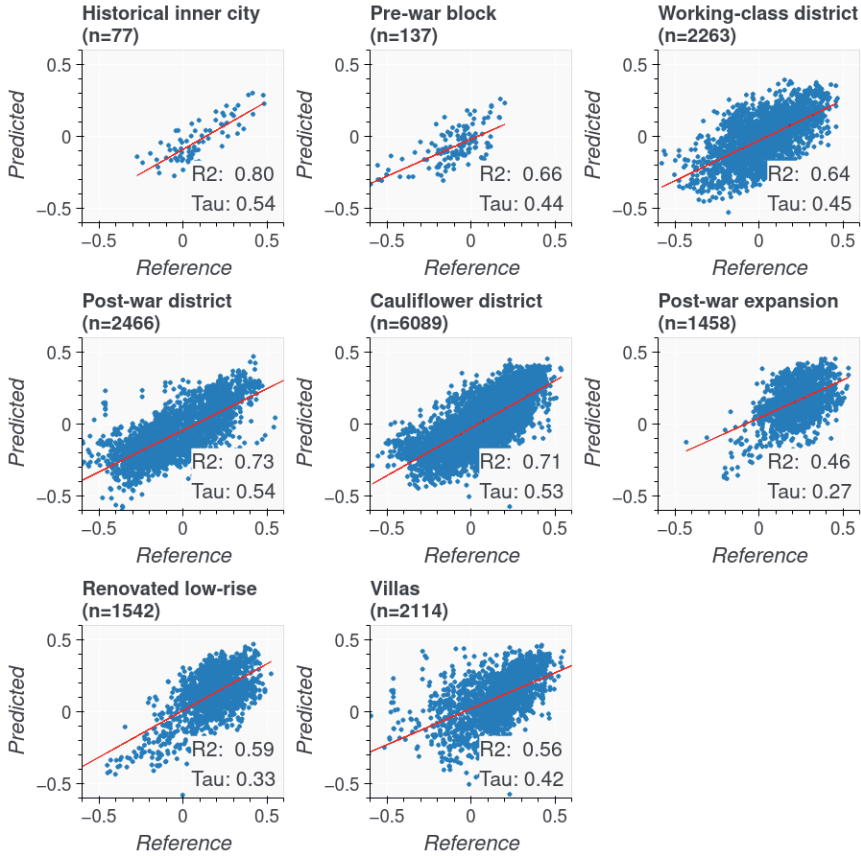


Figure 2.7: Scatterplots of the overall liveability score for all patches in the test set for each of the neighbourhood typologies considered in this research. Patches are included in a scatterplot when there is 20% or more of the given typology present. We show the reference value of each point on the x-axis and the predicted value on the y-axis.

t-SNE embeddings

In Figure 2.8, we show a *t-SNE* plot of the 8 neighbourhood typologies for the global feature vector \mathbf{r} of Figure 2.3, which is the feature vector from which the domain feature matrix C is then derived. The global feature vector \mathbf{r} therefore represents an aggregate summary of all 5 scores at once. As such, this plot represents which neighbourhood typologies are similar across all domain scores. From the graphs, we can conclude that most typologies contribute to domain scores in differing ways, resulting in a heterogeneous spread across the plots, from which it can be deduced that only a neighbourhood typology as a descriptive variable can not explain the variety of all domain scores. However, it becomes more interesting when we consider the *buildings* domain score using $C_{buildings}$.

In Figure 2.9, we show a t-SNE plot of the 8 neighbourhood typologies for the *buildings* domain score. This feature vector reflects only how the model perceives the building quality of patches. The plots for the *buildings* domain score reveal that the selected typologies have varying degrees of visual homogeneity, i.e., they occupy different regions of the t-SNE space with different degrees of spread. *Sub-urban expansion* neighbourhoods, *renovated* neighbourhoods, and *historical inner city* neighbourhoods are considered the most visually homogeneous as perceived by our model.

In particular, the *sub-urban expansion* and *renovated district* neighbourhoods form a single cluster of modern building styles (near example 3 of Figure 2.10), as both of these typologies only appear after the 1990s. This period saw a paradigm shift towards sub-urban construction, though this cluster does not fully encapsulate sub-urban trends, as, for instance, villas are still predominantly present outside of it. The top-most cluster in the t-SNE diagram (near example 1 of Figure 2.10) shows the dense inner city patterns that are present predominantly in Amsterdam and Rotterdam, both historically and pre-war districts. The visual dissimilarity of these areas from any other building style is particularly striking, as they form a small but visually distinct cluster while much of the feature space tends to clump together. It shows that these areas have exceptional properties when it comes to building quality. And indeed, when compared to the other cities in the dataset, Amsterdam and Rotterdam are the two most metropolitan areas within the dataset with certain unique features, such as the canal houses in Amsterdam.

2.4 Discussion

2.4.1 Predicting the liveability of dutch cities with aerial images and semantic intermediate concepts

The capability of the model to predict various domains varies strongly, as evidenced by Table 2.7. Between the metrics that have been evaluated, the model is best able to generalise the *direct* domain scores. The *buildings* domain score especially retains good performance for both metrics in the unseen regions. It is followed by the *physical environment* domain score, which sees a greater reduction in the R^2 metric but retains a high Kendall’s τ score. Of the *indirect* domain scores, only the *population* domain score generalises well to unseen regions. While the *safety* domain score only sees a more drastic reduction in Kendall’s τ , the *amenities* domain score sees a dramatic reduction in both R^2 and Kendall’s τ on the test set. It has the best performance on the validation set, but the strong decrease in performance suggests that amenities are not suitable to predict from aerial images. It should be noted that there are better methods to determine access to amenities compared to prediction from overhead imagery, such as using openly available geodata registries (Sapena et al., 2021). However, in this research, we wanted to study the consequences of predicting proxy variables without the use of auxiliary information in order

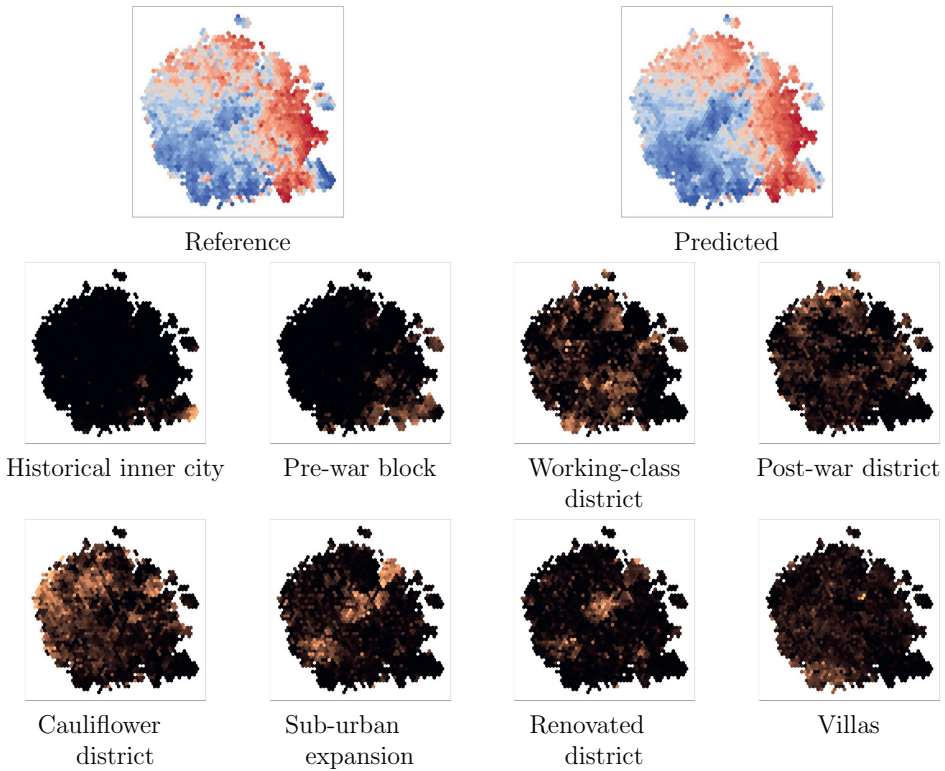


Figure 2.8: t-SNE representation of the features from which the domain-specific feature vectors are derived (vector r in Figure 2.3), overlaid with the percentage of each typology that is present within a patch. Brighter colours represent a higher percentage of the typology present.

to study the ensemble of domain scores and their link to liveability in a comprehensive way. Compared to previous literature, our results lead to several observations. Firstly, we corroborate the findings of Arribas-Bel et al. (2017) and Suel et al. (2019) that high-resolution imagery can be used to predict indirect domain information. Secondly, building on Scepanovic et al. (2021), our results also further prove that directly visible domains may be predicted from remote sensing images. Thirdly, we demonstrate that an end-to-end learned regression pipeline from components to liveability (e.g., the two-step regression experiment of Scepanovic et al. (2021) for urban vitality) does not have to come at the cost of performance on the final task. Lastly, our experiments for the first time raise the proposition that domains relating to liveability that are directly predictable from aerial images are easier to generalise to unseen regions than indirect domain scores.

Our results show that the use of an end-to-end trained bottleneck model generally improves model performance for the final task of predicting liveability. Our bottleneck model matches

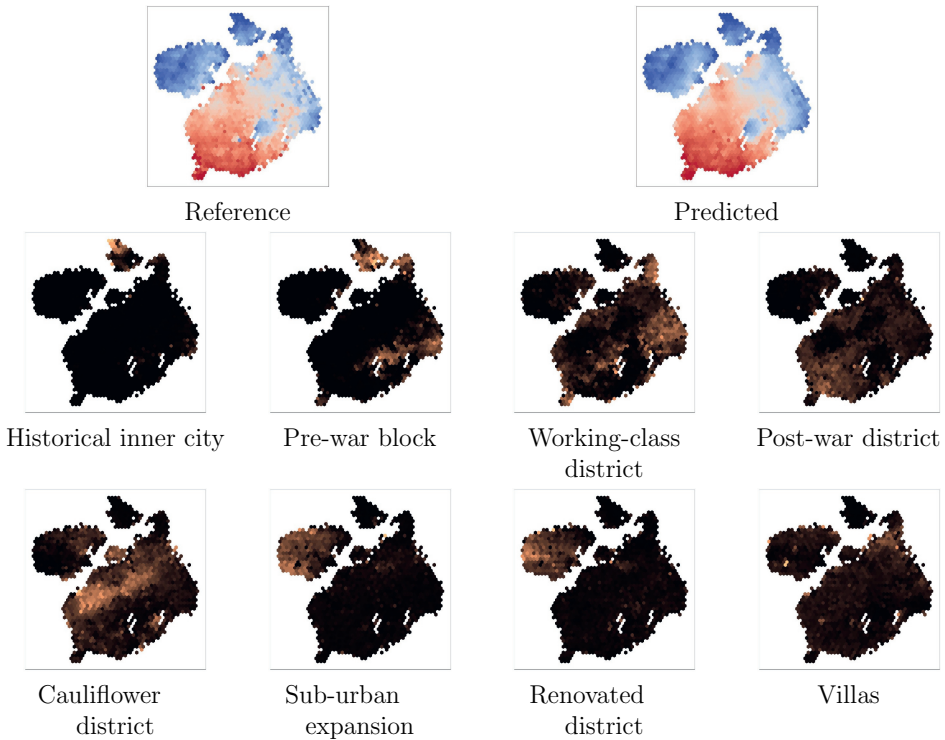


Figure 2.9: t-SNE representation of the features used to predict the *buildings* domain score, overlaid with the percentage of each typology that is present within a patch. Brighter colours represent a higher percentage of the typology present.

the R^2 metric of the unconstrained baseline model and slightly surpasses it on Kendall’s τ . This shows that a linear mapping from the domain scores (which are a decomposition of the overall liveability score) is sufficient for reconstructing the overall liveability score. The reported metrics corroborate earlier findings that the intermediate prediction of a semantic layer can increase the model’s performance on the final task (Levering et al., 2020).

As evidenced by the results, models trained on aerial imagery can transfer fairly well to unseen regions, even across developmental contexts. The cities in our training dataset have a longer history than two of the cities in our test set, namely Hengelo and Eindhoven. Both of these cities started growing as a result of industrialization. As such, their urban form is partially different from the cities with a longer history. Despite this contrast, our model does not have a decrease in performance compared to Dordrecht, which is a city close to the Rotterdam metropolitan area with a longer history of growth. These results suggest that the learned features are robust across developmental contexts. As a result,

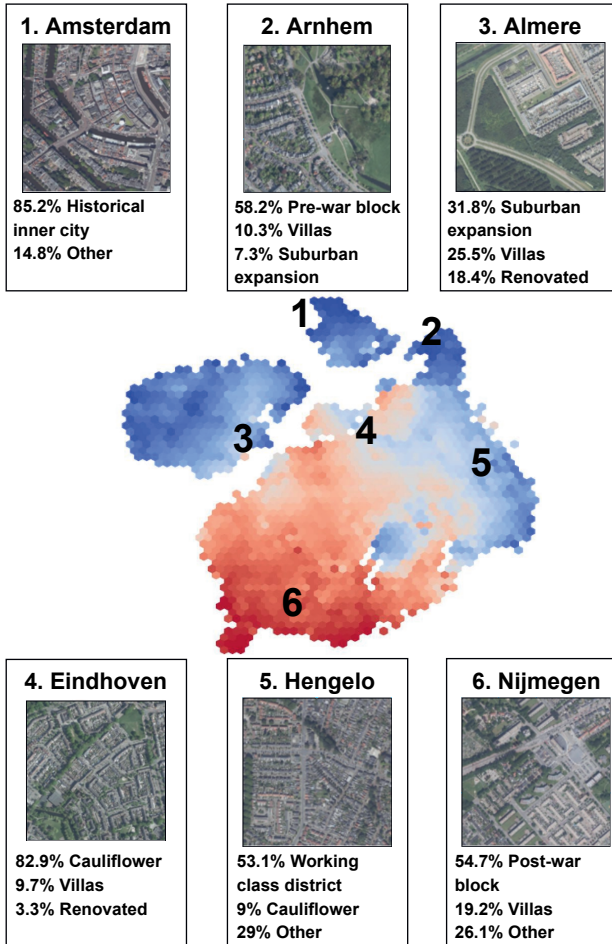


Figure 2.10: Example aerial image patches with their corresponding neighbourhood-level typology labels plotted over the *buildings* domain score embedding. Note that the neighbourhood typology information is often only available at a coarser spatial scale and therefore they may not fully represent the individual patch content.

our findings suggest that extrapolation of liveability factors to unseen regions is a plausible objective, even when generalising across developmental contexts. For amenities, the proxy correlations from overhead images are especially tenuous. In the LBM project, the score is originally predicted from distance variables that exceed the size of our 500-metre resolution patches, for instance, the number of bars within a 2-kilometre radius of the neighbourhood. As such, in a city environment, the model can accurately guess that most amenities are close to a neighbourhood. The inclusion of amenities as a dimension score therefore allows us to study proxy variables with an extreme example. The amenities predictions for Beesel

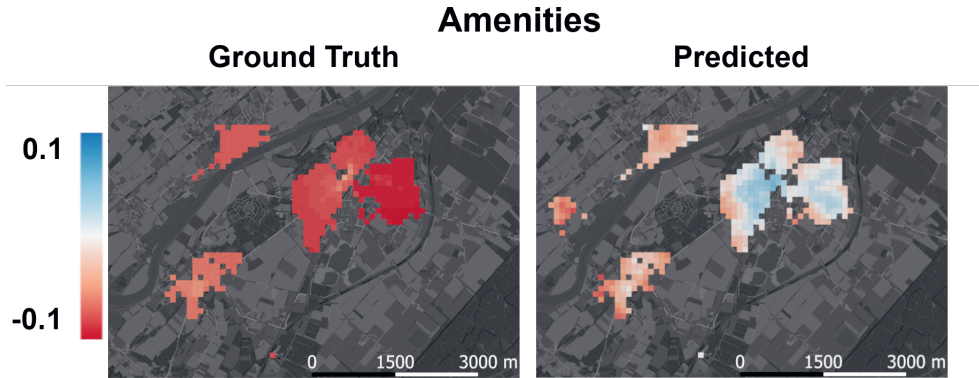


Figure 2.11: Predictions for the Amenities domain score over the region of Beesel. The maps highlight that the model fails to predict the trends present in the villages, leading to negative performance metrics for this test region.

in Figure 2.11 showcase how transferability becomes a problem with proxy variables that rely on urban context, as the model predictions are not at all correlated with the reference labels. For the other three testing sites, this domain score generalises better, as they are medium- to large-sized cities. However, as Beesel is a small village, the model loses geographical context, as the proximity to important amenities such as hospitals is far less certain.

In Section 2.2.1, we presented the variables used for the domain scores of the LBM project. For the *Population* domain score, the variables included in this domain score can be particularly stigmatising. The use of ethnicity data has especially drawn criticism from researchers, as the inclusion of ethnicity without accounting for confounding variables may lead to false stereotypes. While not accounting for confounding variables, the first version of LBM was already used to justify policy decisions. The main concern for the research was to maximise the R^2 coefficient, and as a result, the researchers did not take into account the importance of mitigating stigmatisation (Uitermark et al., 2017). The second version of the LBM has attempted to mitigate the stigmatising effects of including ethnicity variables by only using the residuals after accounting for income. However, it was still widely criticised (Baggerman, 2020; Teeffelen, 2021). In version 3.0 of the LBM, stigmatising variables such as the ones used by population score have been phased out in favour of a more generalised domain, namely *social cohesion* (Leidelmeijer and Mandemakers, 2020). However, during our analysis, this improved version was not yet available. In our research, we have decided to use the population score as it represents a generalised score, which allows us to determine how well socio-economic and socio-demographic data can be predicted and how well this domain will generalise to unseen regions. However, we refrain from analysing prediction patterns for this domain score so as to not perpetuate or justify the use of these stigmatising variables.

2.4.2 Perspectives for Liveability Monitoring with EO

In this section, we discuss how our research can help provide deeper perspectives for liveability monitoring from remotely sensed imagery. There are several possible approaches for modelling liveability using remote sensing. When liveability reference data is not available, deterministic intermediate variables may be used as a proxy, where it is assumed that each variable is an indicator for liveability. For instance, the United Kingdom uses an *index of multiple deprivation*, which measures the relative deprivation of *Lower Layer Super Output* areas with a mean population of 1'500 residents. The index is measured at a fine-grained neighbourhood scale and considers seven domains (income, employment, education, health, crime, barriers to housing and services, and living environment) (Penney, 2019). It combines these dimensions into a final deprivation index using different weights for each dimension using guidance from liveability theory. Some of the individual deprivation factors of this index of multiple deprivations have been predicted through remote sensing (Arribas-Bel et al., 2017; Suel et al., 2021). In such a setup, the role of remote sensing would be to interpolate and extrapolate intermediate variables in support of liveability modelling. Liveability can also be modelled through remote sensing in an end-to-end manner by first predicting intermediate factors and then recombined into a liveability score by using expert opinions (Huang and Liu, 2022). Such a method allows for the acquisition of large-scale inventories of liveability measurements without needing any reference data. The downside to this deterministic measuring process is that the importance of intermediate variables to liveability is not calibrated empirically through resident opinions. As such, the expert opinions on which intermediate variables matter most may be different from the liveability experienced by residents.

As a compromise between deterministic and empirical liveability modelling, hedonic pricing assumes that housing prices are in part indicative of the liveability of a neighbourhood, as people are willing to pay more for houses in liveable areas. This is a simple and scalable assumption, which makes it attractive for large-scale modelling, as the definition of the reference data is constant no matter the location. Therefore, if house sales information is available, it may serve as a proxy for the liveability of an area (Bency et al., 2017; Yao et al., 2018). However, the downside of hedonic pricing is that the assumed contribution of liveability is not tested against resident opinions either, meaning that it may still be off from the liveability experienced by residents. Moreover, a model may need more information to infer how much signal can be attributed to the desirability of a location. For instance, an area may have a poor quality built environment but very attractive surroundings. As a result, it may trend towards the average.

The most informative type of reference data is based on surveyed resident opinions. Such a data source does not assume that there is a relationship between proxy factors and experienced liveability but provides the evidence to directly test such a hypothesis in practice. However, fine-grained liveability reference data based on residents' opinions is

only scarcely available. Surveying efforts are expensive, hard to perform on a large scale, and sensitive to a variety of biases, such as response bias (the tendency for respondents to give inaccurate answers) and participation bias (the inability or unwillingness of certain groups of residents to respond). A well-performed study at scale is therefore a labour-intensive process. The privacy of respondents also needs to be respected, which further complicates the spatial scale at which information is typically reported. This makes the LBM a remarkable project, as it is on a fine spatial scale and is partially modelled on the subjective opinions of residents. To our knowledge, it provides the first large-scale yet fine-grained dataset of liveability that incorporates resident opinions, thus opening possibilities for understanding liveability in the Dutch context, but with limitations that we discuss in the next section.

2.4.3 Limitations

In our research, we use $1m$ resolution aerial image patches from the nationally available aerial image, which is open data, while the liveability reference data is of fine spatial resolution and nation-wide available as well. The unique availability of both data sources allows us to observe liveability with unprecedented geographical scale, resolution, and fidelity. While this work does allow us to pursue the limits of what may be measured at scale from remote sensing imagery, it restrains the methodology to regions with similar data availability. Despite this restriction, our results may be replicated in other countries with liveability labels through commercial satellite services. In that sense, our results are scalable to any region, but they are most strongly applicable to regions with high data availability. Our results indicate that domains that can be observed directly through aerial imagery are easier to generalise than domains that need proxies in order to be predicted. This has practical implications for using remote sensing to fill gaps in data availability for the purposes of predicting liveability. For instance, where possible, amenity data should be derived from sources other than remote sensing imagery, as open geodata registries provide coverage of the most important amenities for most countries. However, if building types are largely homogeneous between two areas but building quality data is only available in one area, then it may be worthwhile to gap-fill this data through remote sensing.

While liveability monitoring from Earth observation has been proven to work for several different research cases and for each of the different types of reference datasets that are available, the subjectivity of the topic continues to hamper comparisons across studies. First, there is no standard definition for liveability, which plays a role in determining a common ground for liveability studies (Paul and Sen, 2020). Second, the way it is measured varies for each study, as do the variables and methods used to measure liveability. We therefore consider liveability prediction to only be valid within the cultural context in which it is measured, with very limited generalisation beyond this context. In other words, the values that make a place liveable are culture- and location-specific. As such, we do not believe that liveability prediction models could be applied out of the box in a

completely different cultural context. While our models retain sufficient performance in unseen cities for the extrapolation of liveability surveys, the entirety of our dataset falls within the same cultural context, which is the Netherlands as a country. As such, our dataset has a largely homogeneous cultural and policy context. Attempting to extrapolate outside of the Netherlands, e.g., attempting to predict liveability in Belgium or Germany with our model, will most likely be less successful due to a difference in cultural and policy context.

The LBM project is an ongoing project that is still being updated. While the input variables and the domains are updated between versions, the reference data upon which the liveability scores are calculated remains unchanged. Meanwhile, the aerial image data will be updated annually for the foreseeable future. As such, the data used in our study can theoretically be used to test whether the relationship between the spatial configuration of settlements and their liveability is persistent over time. As the temporal extent of the datasets increases throughout the years, this option will become more salient as significant changes to the liveability of a neighbourhood, such as gentrification and impoverishment, will take years to manifest.

2.5 Conclusions

In this paper, we study the prediction of liveability from aerial images at the neighbourhood level for 13 built-up areas in the Netherlands. To do so, we test the applicability of remote sensing to predict five domain scores relating to liveability. We assess how well domains that can be learned directly from the image content itself (*physical environment* and *buildings*) can be predicted, as well as domains that require proxy correlations (*population*, *safety*, and *amenities*). Our results indicate that liveability domain scores generalise fairly well to unseen regions, even in regions with a different developmental context. Furthermore, our results indicate that domains that can be directly predicted from the image pixels generalise better than domains that rely on proxy correlations, as the reduction in performance between the validation and the test set is lower for these predicted domain scores. We also study how our model perceives the liveability of different neighbourhood typologies. Our results indicate that our model is proficient at recognising the liveability of different urban typologies, though with varying accuracy. Secondly, through t-SNE dimensionality reduction, we inferred how our model observes homogeneity within neighbourhood typologies. Our results show that our model considers certain neighbourhood typologies to be visually distinct for the purposes of recognising building quality, but less so for overall liveability. Our research suggests that remote sensing can be used to extrapolate liveability surveys to new and unseen regions within the same cultural and policy context. Finally, our study may enable longitudinal studies across time series of aerial images in order to monitor liveability. The code for our project is available at <https://github.com/ahlevering/liveability-rs>.

Chapter 3

On the relation between landscape beauty and land cover: A case study in the U.K. at Sentinel-2 resolution with interpretable AI

This chapter is based on:

Levering, A., Marcos, D., Tuia, D., 2021. On the relation between landscape beauty and land cover: A case study in the U.K. at Sentinel-2 resolution with interpretable AI. *ISPRS Journal of Photogrammetry and Remote Sensing* 177, 194–203. <https://doi.org/10.1016/j.isprsjprs.2021.04.020>

Abstract

The environment where we live and recreate can have a significant effect on our well-being. More beautiful landscapes have considerable benefits for both health and quality of life. When we choose where to live or our next holiday destination, we do so according to our perception of the environment around us. In a way, we value nature and assign an ecosystem service to it. Landscape aesthetics, or scenicness, is one such service, which we consider in this paper as a collectively perceived quality. We present a deep learning model called ScenicNet for the large-scale inventorization of landscape scenicness from satellite imagery. We model scenicness with an interpretable deep learning model and learn a landscape beauty estimator based on crowdsourced scores derived from more than two hundred thousand landscape images in the United Kingdom. Our ScenicNet model learns the relationship between land cover types and scenicness by using land cover prediction as an interpretable intermediate task to scenicness regression. It predicts landscape scenicness and land cover from the Corine Land Cover product concurrently without compromising the accuracy of either task. In addition, our proposed model is interpretable in the sense that it learns to express preferences for certain types of land covers in a manner that is easily understandable by an end-user. Our *semantic bottleneck* also allows us to further our understanding of crowd preferences for landscape aesthetics.

3.1 Introduction

In a time where increasing urbanisation is a constant factor across the world, we sometimes need a break from the busy and tiring reality of the modern city to enjoy greener and more relaxing landscapes. Landscape beauty, also referred to as scenicness, is indeed a driver for tourism (Krippendorf, 1984), while it is also a driver for the creation of cultural value (Daniel et al., 2012; Havinga et al., 2020). Beyond providing tourists and artists with a place to seek out, landscape scenicness has also been found to improve people’s quality of life. Velarde et al. (2007) reviewed literature covering the relationship between health and landscape beauty and found that observing scenic landscapes is associated with a reduction in stress, improved attention capacity, better recovery from illnesses, a feeling of general well-being, and positive improvements to one’s mood. Grinde and Patil (2009) conducted a literature study on the relationship between plants and quality of life and found that the absence of plants is associated with a lower quality of life and health. Seresinhe et al. (2015) quantified the relationship between scenicness and self-reported health and found that scenic environments are associated with an increase in self-reported health. In a later study, they also considered the relationship between self-reported happiness and landscape beauty and found that people are happier in scenic environments (Seresinhe et al., 2019). As such, there is a significant incentive to know where scenic landscapes are located, as well as to understand the factors that contribute to landscape scenicness.

Much research has been devoted to determining landscape scenicness. Theoretical research on the topic stems back to the 1960s through the 1980s, when major theories about human-landscape interactions were formed, as summarised by Schroeder and Daniel (Schroeder and Daniel, 1981). A popular measure for landscape beauty at the time was the *Scenic Beauty Estimate*, which depended on crowdsourced ratings based on images of the landscape (Daniel, 1976). As scenicness is a subjective quality (since '*beauty is in the eye of the beholder*'), accessing such information directly from the observer was (and still is) the only possible way, in the hope that the individual subjective views would then converge to a set of collective rules of perceived beauty. The practice of estimating landscape beauty then adopted digital means by the time that computers and geo-information systems became widely available, such as relating crowdsourced scenic beauty estimates to land cover types through geo-information systems (Palmer, 2004). Recent efforts in data collection (Seresinhe et al., 2017) led to the distillation of the first large-scale crowd-sourced dataset of landscape preferences, called ScenicOrNot¹, consisting of 217,000 ground-level images with scenicness scores from three or more annotators. This dataset is of sufficient size and diversity to allow for the emergence of machine learning research aimed at the automatic estimation of landscape scenicness, which was mostly tackled by means of convolutional neural networks (Marcos et al., 2019; Seresinhe et al., 2017; Workman et al., 2017). However, it may be difficult to acquire ground-based images of remote regions, such as those from the Geograph project², on which the ScenicOrNot dataset is based. For such regions, it could be beneficial to use remote sensing imagery, which is available globally and is frequently updated, to provide the scenicness assessment. Furthermore, remote sensing imagery is not affected by ground-based image biases such as weather patterns such as cloudy versus blue skies or the presence of rainbows, or photographers' biases on which scenes or objects to photograph. In this respect, remote sensing imagery could be considered more objective than ground-based images. The question remains: is it possible to predict scenicness directly from remote sensing images? In other words, we formulate the hypothesis that the characteristics visible in satellite images (e.g., land cover) allow us to estimate the beauty of the landscape. To verify this hypothesis, we resort to a deep learning approach.

In recent years, Convolutional Neural Networks (CNNs) have become a popular tool for image analysis in the remote sensing domain (Zhu et al., 2017). CNNs are commonly applied to typical remote sensing tasks such as land classification (Sumbul et al., 2019; Demir et al., 2018), or precise object delineation at very high resolution (Campos-Taberner et al., 2016; Maggiori et al., 2017; Volpi and Tuia, 2017). While they are traditionally applied to RGB and multispectral data, there nowadays exists a wide corpus of literature about the use of deep learning for other modalities, such as hyperspectral remote sensing (Audebert et al., 2019). As a result, deep learning is becoming increasingly popular

¹<http://scenicornot.datasciencelab.co.uk/>

²<https://m.geograph.org.uk>

in the geosciences community, where the technology is used to tackle a wide range of problems, such as weather prediction, snow pack modelling, or climate change monitoring (Camps-Valls et al., 2021).

However, their superior performance on a variety of tasks comes at the price of interpretability, since CNNs offer less transparency in their predictions compared to other machine learning models. Researchers in machine learning are therefore increasingly stressing the importance of interpretability in deep learning systems (Samek and Müller, 2019; Miller, 2019) in order to be able to challenge the assumptions of deep neural networks and to assess whether a model is trustworthy. Additionally, interpretability can be used to discover meaningful patterns to further our understanding of which learned patterns matter most (Lapuschkin et al., 2019).

While interpretability as a means of improving trust in deep learning models has picked up considerably in computer vision, it is still in its infancy for remote sensing tasks, and traditional machine learning methods have proven to be easier to interpret (Huysmans et al., 2011). In particular, understanding how variables contribute to predictions has been heavily studied with tree-based and kernel methods. Tree-based methods allow for interpretability by ranking input variables according to their influence on the final prediction, such as mode impurity and mean decrease in accuracy for Random Forest models (Biau and Scornet, 2016). Gaussian Processes allow for model inversion and parameter retrieval through their confidence intervals (Svendsen et al., 2020). Linear combinations of multiple kernels can be used to obtain variable importance estimates for kernel methods (Tuia et al., 2010). But when it comes to deep learning methods, the ranking of input importance is less straightforward, and one of the inner features needs extra engineering steps. Instead, post-hoc input attribution methods such as Class Attention Mapping (Zhou et al., 2016) are frequently considered as a solution to the interpretation problem for deep neural networks trained on remote sensing imagery. These methods are used to highlight which regions of the image contribute the most to the output of the model. They are commonly used in various object retrieval tasks, such as locating solar panels (Imamoglu et al., 2017), structures of interest (Vasu et al., 2018), or aeroplanes (Fu et al., 2019). Attribution methods such as Class Activation Maps (Zhou et al., 2016) work well when there is a clear right or wrong answer visible in the image. For instance, an aeroplane can be clearly identified by a human in a very high-resolution satellite image, making the correctness of a pixel attribution method easy to verify. However, attribution methods are less effective when a task is subjective or when it depends on the coalescence of multiple patterns, which cannot easily be highlighted in the image. Scenicness is one such task, as landscape beauty can be the result of the interplay between visible elements of the landscape, and such interplays cannot easily be highlighted in the input images. We therefore have to consider alternative interpretation methods to explain our predictions.

To help us understand the drivers of landscape scenicness using deep learning, we adopt semantic bottlenecks (Marcos et al., 2021; Marcos et al., 2019), which use the prediction results of an intermediate task, ideally objective and made of human-understandable concepts, to predict the target task while still allowing models to be trained in an end-to-end fashion. Such models have previously been applied for scenicness estimation from ground-based images. As proposed in Marcos et al. (2019), the prediction of image scenicness may depend on its content, such as the presence of snow, clouds, or roads. The presence of each object or concept may then be used to create a scoring vector for the prediction of scenicness. In that case, the semantic bottleneck was therefore made of a series of scene-class objects, and to each object, a positive (this object impacts scenicness positively) or negative (this object impacts scenicness negatively) weight was assigned. The final score was made up of a bias (average scenicness) plus the combination of the single detected object scores. We build on this concept for ground-based images and adapt it to the task of scenicness prediction from remote sensing imagery while using land cover as an interpretable intermediate task. In doing so, we improve on our preliminary study (Levering et al., 2020) by adapting our model to accommodate differing scenicness scores within the same land cover class, since depending on the context, one land cover type can impact positively, negatively, or not at all the beauty score. In addition to estimating the scenicness of landscapes, our model therefore also allows us to study the relationship between landscape scenicness and land cover types.

In this paper, we conceptualise an interpretable deep learning model for remote sensing imagery that uses land cover prediction as an intermediate task for landscape scenicness regression (Section 3.2). We train our model to reproduce the average ScenicOrNot beauty score at the level of single patches extracted from Sentinel-2 images over the United Kingdom. We implement a semantic bottleneck, forcing predictions to be explicit in the land cover classes that the model is observing and explicitly using to predict the scenicness. To do so, we use the Copernicus CORINE land cover inventory (EU Copernicus Program, 2018) and predict intermediate land cover multilabel maps. Our results (Section 3.4) show that we can extend existing scenicness prediction models with an interpretable bottleneck without experiencing any loss of accuracy, neither in the scenicness nor land cover prediction task. In return, our model provides explanations about what it is observing and what leads it to make a certain beauty prediction. As such, it becomes simple to challenge the decisions of the model and analyse errors.

3.2 Methods

We propose an interpretable model for landscape scenicness estimation that uses a semantic bottleneck (Marcos et al., 2019). We design the semantic bottleneck such that it uses the outputs of a land cover prediction task to estimate the scenicness of a given satellite image. We refer to our model as **ScenicNet**.

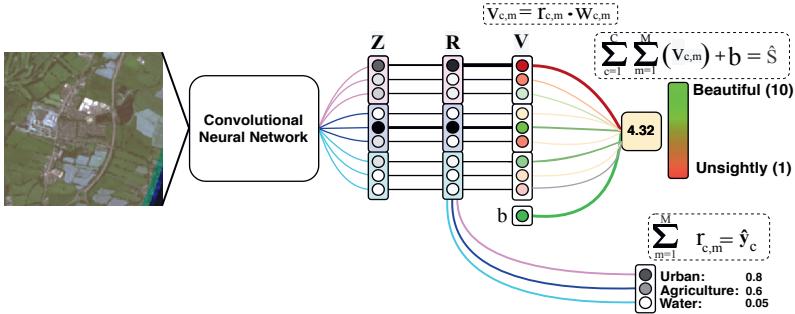


Figure 3.1: Architecture overview of our semantic bottleneck. The model first extracts a matrix of \mathbf{Z} features from a satellite image. Over these \mathbf{Z} features, it then multiplies a classwise softmax with a sigmoid non-linearity (Eq. (3.2)) to extract mode presence scores \mathbf{R} . Land cover presence is predicted from these features by summing the resulting matrix (Eq.(3.3)). We multiply this presence matrix with one learned weight per mode to derive their scenicness contribution for a given sample (Eq.(3.5)). The sum of all modes is added together with a bias term to create the final scenicness prediction (Eq.(3.6)).

Our model is summarised in Fig. 3.1. It uses a standard CNN backbone tasked with feature extraction, a multi-label land cover classifier intermediate head, and a scenicness regressor, which depends linearly on the output of the land cover classifier. Since it comprises two separate prediction heads considering different tasks learned from different datasets (see Section 3.3), it can be seen as a multitask model, such as in Marmanis et al. (2018) and Volpi and Tuia (2018).

Our main contribution is a method to disambiguate intra-class scenicness differences by allowing the model to discover sub-classes with different scenicness values associated with them. We call these sub-classes *modes*. Each mode corresponds to a neuron within a group of neurons associated with the same land cover class. Each mode is also connected to the scenicness head (via the weights w described below). Each mode therefore contributes to both the detection of land cover and the estimation of beauty. The number of modes per class is defined by a hyperparameter, M , manually set. Summing up, for each land cover class $c \in C$, the model has M outputs, each with an associated learned scenicness weight. This means that a land cover class can influence scenicness positively when in a given association of classes and then negatively when associated with others. Depending on the specific association, one or another mode of the class will be activated.

3.2.1 Land cover head

Our model first has to predict C land cover classes from the feature extractor. The feature extractor produces $C \times M$ scores $\mathbf{Z} \in \mathbb{R}^{C \times M}$ for each mode input $m \in \{1, \dots, M\}$ belonging to a given class $c \in \{1, \dots, C\}$, where $z_{c,m}$ corresponds to the features of mode m

in class c . These scores are then normalised (Eq. 3.2) and summed for each class (Eq. 3.3), to obtain the C land-cover class scores as a vector $\hat{\mathbf{y}} \in \mathbb{R}^C$. As depicted in Fig. 3.1, the land cover prediction problem is cast as multi-label, i.e., every class is considered separately and can be detected simultaneously with others. We use a binary cross entropy loss for every land cover class $c \in \{1 \dots C\}$ and compare predictions $\hat{\mathbf{y}}$ with the ground truth $\mathbf{y} \in \{0, 1\}^C$.

For the purposes of scenicness prediction, we want to force the model to choose which mode to use for a given sample to reduce ambiguity on which modes contributed to each prediction. In order to keep the scenicness prediction layer interpretable, we also want the model to only keep the modes that have a meaningful contribution to the prediction process active. To do so, we first calculate a Softmax non-linearity for each mode input $m \in \{1, \dots, M\}$ belonging to a given class $c \in \{1, \dots, C\}$:

$$\text{softmax}(z_{c,m}) = \frac{e^{z_{c,m}}}{\sum_{j=1}^M e^{z_{c,j}}} \quad (3.1)$$

For each element $z_{c,m}$ we then multiply their respective softmax scores with a sigmoid over the mode input to compute the mode presence probability for a given mode $r_{c,m}$ of matrix \mathbf{R} :

$$r_{c,m} = \text{sigmoid}(z_{c,m}) \cdot \text{softmax}(z_{c,m}) \quad (3.2)$$

The softmax ensures that only one mode is dominantly active, as all class-specific contributions add up to one. Through direct multiplication with the sigmoid non-linearity, we allow the model to indicate which modes are active, if any. We can then use this mode presence matrix \mathbf{R} to obtain class presence scores by summing all mode presence scores $r_{c,m}$ belonging to a given class c :

$$\hat{y}_c = \sum_{m=1}^M r_{c,m} \quad (3.3)$$

We can use these class-wise land cover presence scores in the following sum over c binary cross-entropy functions (one per land use class) (Fig. 3.2a):

$$\mathcal{L}_{CLC}(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_c \hat{y}_c \log(y_c) + (1 - y_c) \log(1 - \hat{y}_c) \quad (3.4)$$

Where y is the ground truth for a single sample from the land cover dataset.

The gradients learned from the land cover prediction (in pink to purple colors in Fig. 3.1) are then backpropagated into the main body of the CNN through the class-specific multi-

mode land cover bottleneck. The updated mode presence scores \mathbf{R} will therefore impact the scenicness prediction described in the next section.

3.2.2 Scenicness prediction head

The second head of our model is responsible for predicting landscape scenicness as a regression problem. In order to regress a scenicness value, our model multiplies a learnable weighted matrix $\mathbf{W} \in \mathbb{R}^{C \times M}$ elementwise with the mode presence scores matrix \mathbf{R} to create a matrix \mathbf{V} with mode-specific scenicness contributions, where $v_{c,m}$ represents the contributions of a single mode:

$$v_{c,m} = r_{c,m} \cdot w_{c,m} \quad (3.5)$$

The sum of all mode contributions is then added together with a bias term $b \in \mathbb{R}$ in order to compute the predicted scenicness value:

$$\hat{s} = \left(\sum_{c=1}^C \sum_{m=1}^M v_{c,m} \right) + b \quad (3.6)$$

We then use this predicted scenicness score to compute the following squared error loss function:

$$\mathcal{L}_{SoN}(s, \hat{s}) = (s - \hat{s})^2 \quad (3.7)$$

Where s is the crowdsourced scenicness score for a single sample. During training, we backpropagate the mean squared error of each batch.

With a choice of $M > 1$, our model can learn more than one representation for each $c \in \{1 \dots C\}$ classes. However, we want to encourage the model to use the minimum number of modes needed for the prediction task to stop the model from forming complex non-linear interactions between multiple modes. We encourage this through the softmax in Eq. (3.2), through which we limit the activation budget of the model. The softmax rescales the contributions of each mode relative to all mode activations within a class. Therefore, the model cannot activate all modes equally, forcing it to make deliberate choices on which modes to use for each training example.

3.2.3 Combined loss function

Each one of the two processing heads of the model backpropagates gradients related to a loss specific either to the land cover task (\mathcal{L}_{CLC} , Eq. (3.4)) or to the scenicness estimation task (\mathcal{L}_{SoN} , Eq. (3.7)). The final loss of our explainable model is obtained by a weighted combination of the two terms:

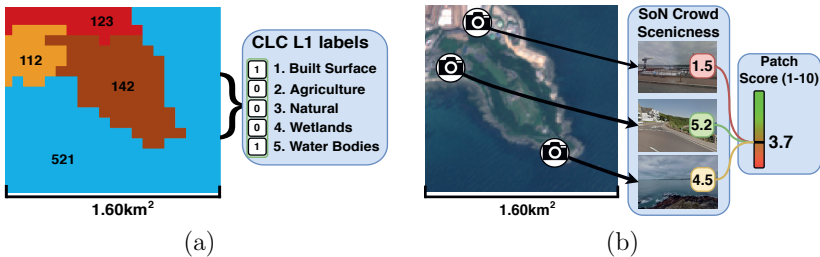


Figure 3.2: Ground truth creation; (a) CORINE values are aggregated to their 1st digit, then assigned a binary present/not-present label. (b) SoN image scores within the patch boundary are averaged, which gives us the patch scenicness score.

$$\mathcal{L} = \mathcal{L}_{SoN} + \lambda \mathcal{L}_{CLC} \quad (3.8)$$

where λ is a weighting term set empirically.

3.3 Data and setup

3.3.1 Data

Our model is concurrently trained on two tasks, namely land cover prediction and scenicness regression. In order to generate the training data for both tasks, we lay out a regular grid of 1.60km by 1.60km across the entirety of Great Britain as a common prediction grid. For each grid cell, we then collect three data sources (Fig. 3.2): 1) A land cover inventory, 2) a landscape scenicness dataset with location information, and 3) satellite imagery with a maximum of 1% cloud coverage across Great Britain.

- *Land cover.* For the land cover prediction, we make use of the CORINE land cover inventory of 2018 EU Copernicus Program, 2018. The CORINE Land Cover (CLC) is a pan-European dataset created from a combination of Sentinel-2 imagery and national land cover products. It consists of a hierarchy of three levels. CLC Level 1 consists of five land cover classes: 1) Urban, 2) Agriculture, 3) Forests and natural areas, 4) Wetlands, and 5) Water. CLC level 3 contains fine-grained land cover classes, such as 111) Continuous Urban Fabric, and 421) Salt Marshes. For our experiments, we use CLC Level 1 as training labels, and we use the L3 labels for a qualitative assessment of the modes of our model in the discussion section. We opt for a more simplistic land cover classification task to ensure that the model is able to learn an accurate representation of land cover classes. For each grid cell, we create a binary vector where 0 and 1 denote absence and presence for each class. We show this process as well as the land cover classes of the first-level hierarchy of CLC in figure 3.2a.

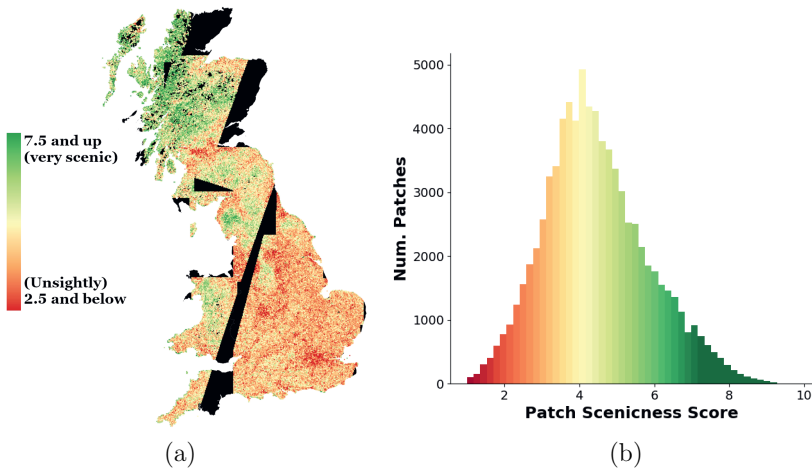


Figure 3.3: Ground truth creation; (a) Map of the ground truth scores of every patch in our dataset. (b) Histogram of all patch scores.

- *Landscape scenicness.* We derive our landscape aesthetics score from ground-based image evaluations from the ScenicOrNot dataset. ScenicOrNot (SoN) is a crowd-sourced dataset consisting of 215,000 ground-level images across Great Britain obtained from the Geograph UK project. Each image is rated with a score between 1 (not scenic) and 10 (most scenic) for its landscape aesthetic beauty by one or more volunteers on an openly accessible online platform. Moreover, each image is stored with its geolocation, and as such, they can be analysed spatially. For each grid cell in our regular grid, we assign the average scenicness score of the geotagged images within its bounds. We display this process in figure 3.2b. Figure 3.3 illustrates the final ground truth as well as the histogram of its distribution across the U.K.
- *Remote sensing data.* As input to our model, we use Sentinel-2 satellite imagery. We download atmospherically corrected (L2A) satellite tiles with at most 1% cloud coverage across Great Britain, which have been taken between 2018 and 2019. We retain the 10- and 20-metre resolution bands of each satellite tile. We upsample the 20m-resolution bands to 10m using nearest neighbour interpolation. We remove any image patches that have all-zero values in the red, green, or blue colour bands. In total, we collect 121,067 patches of size 160×160 pixels, corresponding to an extent of 1.60×1.60 kilometres each. Land cover information is available for all of these patches, while scenicness scores are available for 83,374 patches. We randomly sample splits of 75/15/10% for training, validation, and testing. We sample without geographical stratification to maximise the opportunities for the model to learn meaningful scenicness differences for each class.

The scripts for creating our ground truth dataset can be found in the following Zenodo repository: <https://sandbox.zenodo.org/record/747445>. This repository also contains a PyTorch implementation of our architecture.

3.3.2 Set-up

As a feature extractor, we use a ResNet-50 (He et al., 2016), which has not been pre-trained as we use multi-spectral imagery. We set the number of class-specific modes M to 3, and we initialise the weights in W for the class-specific modes to 0.5, 0.01, and -0.5, respectively, so that the model is encouraged to develop non-symmetrical scenicness contributions for each mode. The scenicness prediction of our model is dependent on the land cover prediction task, but during training, both tasks compete for signal. To avoid that the model learns a bottleneck optimised for scenicness and that is not aligned with the CLC semantics, we set a larger weight λ for the land cover loss in Eq. (3.8), to a factor 10.

We explore the benefit of having multiple modes by running a baseline experiment with M , the number of sub-class nodes per class, set to 1, which makes it functionally equivalent to a linear regression dependent on the class prediction score. We train both models for 15 epochs with the ADAM optimizer (Kingma and Ba, 2014). We set the initial learning rate to 0.0005, and we add a weight decay factor of 0.0001. We use 16 samples per batch. During training, we weight each loss by the inverse square root of its frequency so that we can train on a balanced number of samples.

For every training iteration, we sample one batch to compute Eq.(3.4) and Eq.(3.7). If both labels (CLC and SoN) are available for a given patch, then we compute both losses for the sample. When processing a sample only having land cover information (and no ScenicOrNot label), we set the loss of Eq.(3.7) to 0. We combine and backpropagate the losses according to Eq.(3.8). We repeat this procedure until the smallest dataset (SoN) is exhausted, at which point the epoch ends.

We compare our models against unconstrained ResNet-50 models trained on each task separately. For the land cover prediction task, we set the number of outputs of the final fully connected layer to 5 to equal the number of CORINE classes in the level-1 hierarchy. For the task of scenicness regression, we set the number of outputs of the fully connected layer to 1 such that the model regresses one single scenicness, as in (Workman et al., 2017; Levering et al., 2020). We also test the performance of our model with $M = [2, 5]$ using the same training settings, but with a random initialization of W . We evaluate the land cover prediction performance of our model using the average F1-score (Rijsbergen, 1979) for each class. The F1-score gives the harmonic mean between the precision and the recall of a given class. A value of 1 indicates perfect precision and recall. To assess the scenicness prediction performance of our model, we use the root mean squared error (RMSE) across all examples. We also calculate Kendall’s Tau (Kendall, 1938) over the predicted scenicness scores, which is a ranking correlation coefficient that tests whether two

arrays have similarly ranked values. For Kendall’s Tau , 1 indicates a perfect relationship between the predicted scores and the ground truth, and -1 indicates the inverse.

Finally, we compare the results of our scenicness regression to models that directly regress the scenicness score from the CORINE ground truth labels. We train a linear model using the level-1 hierarchy of CORINE to compare it to our 1-mode linear bottleneck. We then train a random forest regressor (Breiman, 2001) with 50 trees, a maximum depth of 25, and a minimum of 5 samples per split on the L1 and L3 CORINE ground truth labels to test the performance of our multi-mode models against.

3.4 Results and discussion

3.4.1 Numerical scores

In Table 3.1, we display the numerical performances of the four considered models. Each of our ScenicNet models outperforms an unconstrained network on the land cover prediction task. Our 3-mode and 5-mode ScenicNet models also match the scenicness regression baseline on Kendall’s τ . Our results show that our ScenicNet model is able to leverage its modes to learn complex land cover class representations that relate to scenicness in varying ways, rather than the single learnable pattern for the 1-mode model. The numerical improvements of our multi-mode ScenicNet models on the land cover F1-score also indicate that the land cover prediction task seems to benefit from the scenicness prediction task, which is an underlying assumption of multi-task learning (Caruana, 1997).

Table 3.1: F1-score, RMSE, and Kendall’s τ of each model on the test set.

	land cover F1-score	Scenicness RMSE	Scenicness τ
Only CORINE	0.846	-	-
Only SON	-	1.027	0.452
ScenicNet (1 mode)	0.859	1.080	0.435
ScenicNet (2 modes)	0.867	1.053	0.441
ScenicNet (3 modes)	0.872	1.038	0.456
ScenicNet (5 modes)	0.872	1.036	0.457

For the baseline and the 3-mode ScenicNet model, we also present the precision, recall, and F1-score for each land cover class, which can be found in Table 3.2. Our 3-mode ScenicNet model improves on the baseline for land cover prediction on all land cover classes. In the cases of urban and wetlands, our model particularly improves the number of recalled samples.

To test the relationship between land cover and scenicness, we compare our models against a linear regressor and a random forest regressor, which use the land cover ground truth labels

Table 3.2: Class-wise performance metrics of the CORINE baseline and ScenicNet with 3 modes. In each column we display the performance of the baseline on the left, and our model on the right.

	Precision		Recall		F1	
	Base-line	Ours	Base-line	Ours	Base-line	Ours
Urban	0.859	0.865	0.701	0.740	0.772	0.798
Agriculture	0.971	0.974	0.936	0.946	0.954	0.960
Forests and Natural	0.848	0.974	0.821	0.946	0.835	0.960
Wetlands	0.805	0.781	0.617	0.775	0.699	0.778
Water	0.973	0.965	0.968	0.979	0.970	0.972

Table 3.3: RMSE, and Kendall’s τ of models trained to regress scenicness from the land cover ground truth labels.

	Scenicness RMSE	Scenicness τ
Linear (L1)	1.150	0.417
Random Forest (L1)	1.081	0.425
Random Forest (L3)	1.061	0.444

to directly regress the scenicness score. We show our results in Table 3.3. Remarkably, our linear 1-mode model outperforms the score-to-score regression models. We hypothesise that our model is able to provide better performance in predicting scenicness from LC classes by allowing for subtle modifications to the LC probability maps that help with scenicness regression. Our results also show that these subtle modifications not only do not degrade the LC prediction performance but actually provide a substantial boost due to the synergy between the two tasks. By contrast, both a linear model and a random forest regressor use only the binary label present in the ground truth, without the possibility of tweaking it to improve the scenicness prediction performance.

3.4.2 Mode activity

While our model is initialised with M modes, the Softmax function of Eq.(3.2) lets the model spend an activation budget across its modes. Through this activation budget, the model develops the tendency to allocate the vast majority of the signal to a single mode. By doing so, we encourage the model to learn a specific mode only if it needs to account for classes with contrasting scenicness values, such as forests near a city compared to forests in a scenic highland. As a result, it can occur that modes for some classes become inactive (i.e., the sigmoid+softmax combination never activates above 0.5), as there are too few intra-class contradictions to account for. In the case of $M = 3$, we found that the model eventually converges to use 2 modes per class at most, while the inactive modes can be pruned without affecting the performance of the model. Setting M to 2

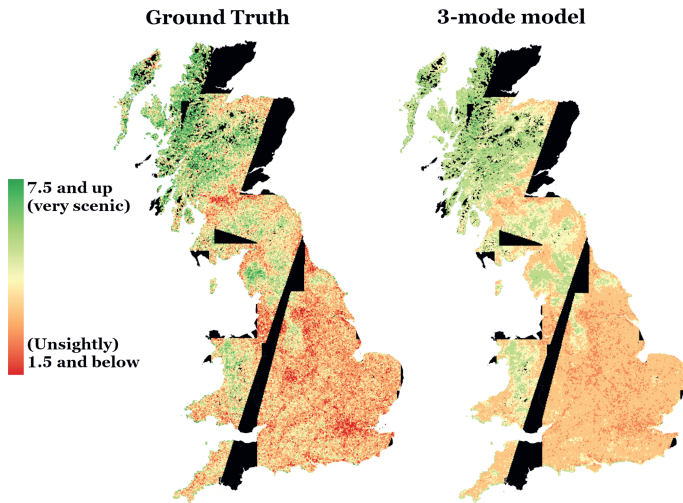


Figure 3.4: Left: Geotagged scenicness scores from the ScenicOrNot project. Right: Scenicness values predicted by 3-mode ScenicNet model. Scenicness is clipped between 1.5 and 7.5 to show more variation in the 3-mode model.

resulted in a solution that is slightly worse than $M = 3$, while $M = 5$ resulted in a model with similar performance. As $M = 3$ gives a model with similar performance but less complexity, we chose this model for our experiments and discussion. We list the active modes of our 3-mode model for each class in Table 3.4. The choice of M should therefore be determined through experimentation, as it should capture the latent dynamics between the two tasks.

3.4.3 Visual evaluations

In this section, we evaluate the performance of our 3-mode model, as well as its activation patterns and the behaviour of its mode. Figure 3.4 illustrates the scenicness predictions of our 3-mode ScenicNet model alongside the ground truth. As can be seen, the model picks up on the major patterns of the ground truth scenicness labels. It is clear that major cities such as London and Manchester are considered unsightly, while Scotland and Wales are considerably more scenic than England. Our model also captures the relationship between elevated areas and scenicness, where higher areas typically correlate with greater scenicness. However, it is also apparent that the model is unable to approximate extreme values, such as those found in downtown London and the Scottish highlands, which we suspect to be caused by an under-representation of these values in the ground truth, as suggested by the histogram in Figure 3.4.

We explore the latent space of our 3-mode ScenicNet model to understand which patches our model considers visually similar. Our main interest with this experiment is to discover whether visually related areas and concepts are similar in the high-dimensional latent space of the CNN model. We reduce the 2'048 outputs of the feature extractor to 100 principal components using a Principal Component Analysis (Pearson, 1901), which are then reduced to 2 dimensions using t-SNE dimensionality reduction (Maaten and Hinton, 2008). t-SNE is a non-linear visualisation technique that performs dimensionality reduction by learning an embedding that preserves neighbourhood structures, i.e., samples that are neighbours in the high-dimensional space must remain neighbours after projection. For the t-SNE hyperparameters, we use a perplexity of 300, a learning rate of 200, and we set the number of iterations to 1'000. We show the resulting plots for the predicted scenicness and class labels in Figure 3.5. We find that the latent space of our model is organised by the predictions made through the class-specific modes. Predictions routed through each mode relate to strongly differing land cover archetypes, which are grouped by their relative scenicness. This organises the latent space into an arrangement where both similarity in land cover visuals (e.g., "bare rocks" and "sparsely vegetated") as well as their relative scenicness are important. An example of this behaviour can be seen in the overlap between modes 3+ and 4+: activations of both of these modes are neighbours in the latent space, while they both have a considerably high learned scenicness score. From Table 3.4, we can infer that these modes are activated by a similar set of fine-grained land cover concepts, namely highland and plains environments. These findings are encouraging as they indicate that the model is consistent in the concepts it considers to be scenic between different but related land cover classes. The plots of the modes also reveal a gradual transition in visual similarity from man-made land cover classes to natural areas. The large cluster in the centre is dominated by un-scenic agriculture and urban land covers, which correspond to England's countryside. To the right, it is connected with and slowly transitions into a cluster dominated by mixed agriculture and woodland environments typically found in Wales, the north of England, and Scotland. From this transition, we infer that the model considers natural areas to be more scenic. This pattern is reflected in the gradient of the top-left cluster. It sees urban areas on the far left of the cluster transition into very scenic natural areas and wetlands at the other edge of the cluster, which suggests that there is a similar transition of scenicness from man-made to natural areas for coastal environments.

Effect of multiple modes to the final prediction The learned bias of our model is 4.65, which corresponds to an average value of scenicness for the whole region. Deviations from this value are related to the land cover-related weights. We further assess these deviations by analysing the most-recalled level-3 CORINE classes per mode in Table 3.4, as well as their weights. We find that each class has at most two active modes, with a large difference in scenicness scores between both modes. Each mode tends to recall different thematic clusters, such as mode 4-(the minus sign represents here the negative influence

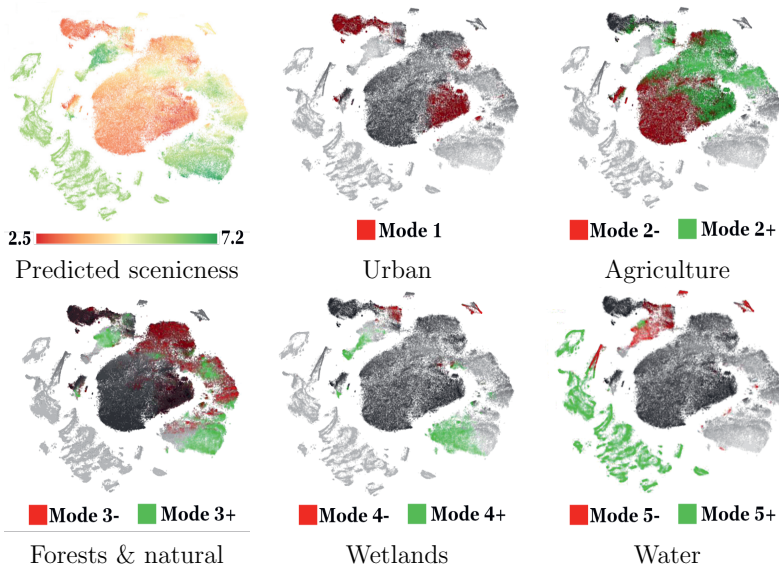

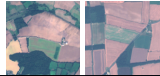




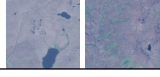




Figure 3.5: Low-dimensional representation of the prediction outputs of our 3-mode ScenicNet model, visualised using t-SNE. We display the predicted scenicness scores for each datapoint as well as the predictions of the active modes of each class. The red colours of each mode refer to modes with the most negative weight within each class, while green is used for the mode with the most positive weight within each class.

this mode has on scenicness) recalling flat coastal wetland environments, while mode 4+ tends to recall elevated boglands, Scottish highlands, and loch environments, which impact landscape beauty positively. This spatial binning effect of the positive and negative modes can be seen in figure 5.3 for all classes, except for the *Urban* and *Agriculture* classes. The *Urban* class defaults to one single un-scenic mode, while the *Agriculture* class experiences strong mixing between its two modes, as both semantic clusters tend to be widespread throughout the country. The presence of human influences on the landscape can be seen in the water and wetland classes. While the coastline of England is predicted to be very scenic (mode 5+), its rivers and estuaries are only mildly positively associated with scenicness (mode 4-). This pattern is visible in all of the major estuaries in England. However, the inlets and open waters connected to the ocean in Scotland are considered strongly positive. While inland waters are only mildly positively associated with scenicness, as in England, their presence strongly correlates with natural areas (mode 3+) and wetlands (mode 4+). The weights of these modes indicate that our model considers these land cover classes to be very scenic in the Scottish Highlands. This indicates that people value inland water environments, but mostly for their nature and wetland environments. The validation of such observations, for example, via interviews, could be the topic of further studies.

Table 3.4: Modes for each class with their learned scenicness score and their most-recalled level-3 CORINE labels. We renamed modes according to their scenicness score and removed inactive modes from the table. While our model is trained with the coarse 5-class first-level hierarchy ground truth of CORINE, the two modes of each class (except urban) are associated with differing fine-grained land cover concepts.

Mode	Weight	Top L3 class by recall	Most activating
Bias	4.65	-	-
1	-0.938	111 cont. urban fabric (0.966) 141 green urban areas (0.948) 121 industrial/commercial (0.688)	
2 -	-1.080	244 agro-forestry (1.0) 222 fruit trees (0.611) 211 non-irrigated (0.561)	
2 +	0.068	313 mixed forests (0.621) 243 agriculture with nature (0.594) 311 broad-leaved forests (0.562)	
3 -	-0.172	312 coniferous forests (0.586) 324 woodland-scrub transition (0.576) 313 mixed forests (0.523)	
3 +	1.391	332 bare rock (0.757) 333 sparsely vegetated (0.756) 334 burnt areas (0.667)	
4 -	-0.678	421 inland marshes (0.512) 423 intertidal flats (0.405) 522 estuaries (0.404)	
4 +	1.178	412 peat bogs (0.620) 333 sparsely vegetated (0.496) 332 bare rock (0.314)	
5 -	0.105	331 beaches, dunes, sands (0.614) 522 estuaries (0.590) 123 ports (0.55)	
5 +	1.193	523 sea/ocean (0.818) 521 coastal lagoons (0.5) 331 beaches, dunes, sands (0.181)	

The learned weights of our modes can be related to three previously quantified observations. Firstly, our model supports the notion that the presence of human influences and structures in a landscape reduces its beauty (Vries et al., 2012; Lindemann-Matthies et al., 2010; Palmer, 2004; Hodgson and Thayer, 1980), as visible by the scenicness weights of modes 1, 2- and 5-. However, not all classes with human influences are considered un-scenic, such as estuaries and beaches in mode 5-. Secondly, landscape beauty is greater in natural areas where there is an open canopy (Schirpke et al., 2013; Hill and Daniel, 2007), which can be inferred from the differences in scenicness values of modes 2+, 3-, 3+, and 5+. These results are corroborated by the spatial patterns of modes 3+ and 4+, which can be

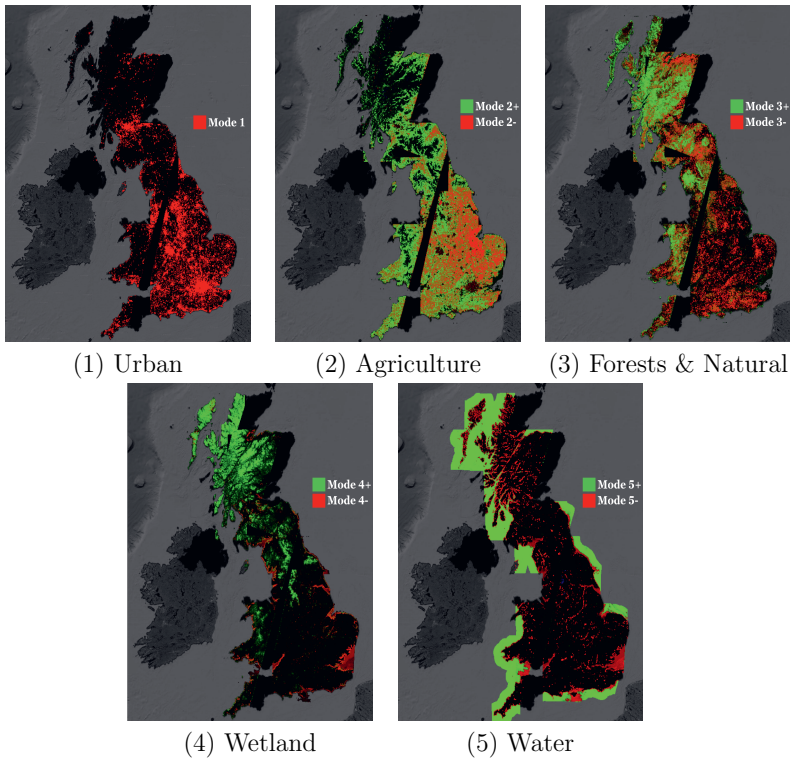


Figure 3.6: Plots of predictions for each patch made by our 3-mode ScenicNet model. Areas coloured red are predicted by the negative mode of a given class, while green areas represent the positive mode. Areas with blended colours have activations in both modes, resulting in a scenicness score that is in between both weights.

seen in the forest map in figure 5.3. These modes are considered very scenic, while their recalled geo-located patches often correspond with hilly and mountainous regions. Lastly, our learned weights for the agriculture class do not directly support survey data, which indicates that the British public enjoys the British countryside for its landscape beauty (Hall et al., 2004). It should be noted that these quantified patterns are difficult to relate to our research as they cover different countries or regions and use different measurement techniques. Further research may attempt to learn patterns on a local scale to see whether local patterns in the United Kingdom extend across regions.

3.5 Conclusions

In this paper, we present and test a novel method for large-scale inventorization of landscape scenicness that uses land cover prediction as an interpretable intermediate task. Our model is able to learn scenic and un-scenic representations of the same land cover type by being able to choose which of several land cover-specific weights to use for the scenicness regression task. Our model outperforms an unconstrained model on the task of land cover prediction while matching an unconstrained model on scenicness regression. Furthermore, our model is able to express preferences for fine-grained land cover types while being trained on just five coarse land cover concepts, which allows us to study the relationship between landscape beauty and land cover types. Our work also opens up possibilities for knowledge and sub-class discovery. We note that our findings are still subject to the fact that all data come from the U.K., only apply to landscape preferences in the U.K., and are most probably provided by British citizens. Expanding these findings to global measures of landscape aesthetics would require a larger corpus of crowdsourced data as well as images coming from all over the world. Creating such a dataset would open the possibility for cultural and global studies about human preferences and appreciations of nature.

Chapter 4

Cross-modal learning of housing quality in amsterdam

This chapter is based on:

Levering, A., Marcos, D., Havinga, I., Tuia, D., 2021. Cross-Modal Learning of Housing Quality in Amsterdam, in: *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GeoAI 2021*, pp. 1–4. <https://doi.org/10.1145/3486635.3491067>

Abstract

In our research we test data and models for the recognition of housing quality in the city of Amsterdam from ground-level and aerial imagery. For ground-level images we compare Google StreetView (GSV) to Flickr images. Our results show that GSV predicts the most accurate building quality scores, approximately 30% better than using only aerial images. However, we find that through careful filtering and by using the right pre-trained model, Flickr image features combined with aerial image features are able to halve the performance gap to GSV features from 30% to 15%. Our results indicate that there are viable alternatives to GSV for liveability factor prediction, which is encouraging as GSV images are more difficult to acquire and not always available.

Modern-day urbanisation has led to large increases in the number of people living in cities. It is expected that more than half of the global population will live in cities by 2050 (United Nations, Department of Economic and Social Affairs, Population Division, 2019). While cities are increasingly important for finding work, it is frequently the case that cities show very disparate access to service and quality of infrastructure, and therefore a mixed quality of life in urban dwellings. Not taking into account people's social and physical housing needs can have detrimental effects on their well-being. For instance, the physical quality of housing is an indicator of one's mental well-being (Evans, 2003). In a broader sense, the quality of a neighbourhood may affect the residents' dietary and physical activity patterns (Thompson and Kent, 2014), as well as their morbidity (Barber et al., 2016). Evidently, monitoring that neighbourhoods are liveable and of adequate quality could support policymakers and urban planners in designing more liveable cities. Liveability is typically measured using surveys. However, surveyed data is expensive to acquire and infrequently available, and their results may be hard to scale beyond the original survey area. Ideally, quality of life data gathered through such surveys would be available at large scales, on a frequent basis, and at a low cost to monitor the liveability of urban areas and identify areas for improvement.

Image data such as ground-level photography can offer a solution to this problem, as it is easier to acquire and scale. Prior research has shown that ground-level images can reliably pick up attributes relating to urban sentiments (Dubey et al., 2016; Naik et al., 2014). A potential drawback is that large-scale collection of this data is often not trivial, and images may be affected by biases such as lighting and weather effects. Aerial images are another source of image data that can be considered. Their main advantage over ground-level images and surveys is that they can be used to survey large areas in a single data collection effort. For aerial images, it has also been proven that they can be used to survey factors relating to quality of life (Scepanovic et al., 2021; Levering et al., 2021a).

In this research, we focus on building quality scores surveyed in Amsterdam on a hectometer scale. We are interested in determining if a combination of ground-level images and aerial

images can improve the prediction of liveability factors. For the aerial image modality, we train models using high-resolution aerial image data. For the ground-level model, we compare two pre-trained feature extractors to determine if models tuned towards liveability make a noticeable difference in performance on our dataset of housing quality. Furthermore, we also train models to combine both modalities to test whether or not they can improve the overall prediction accuracy of housing quality.

4.1 Data

We use three data sources in our study: housing quality scores of the city of Amsterdam, Aerial imagery, and ground-level imagery.

Housing quality scores

For our liveability ground truth labels, we use housing quality scores over the city of Amsterdam. This score quantifies how housing contributes to liveability. This data is available as a grid with cells covering 100m² each. It is derived from various statistics such as building age, ownership situation, and consumption of utilities such as electricity. The statistics used to create the building score are averaged from buildings within 200m² from the patch center. The grid with scores is published by the Leefbaarometer project ¹.

Aerial imagery

For the aerial images, we use aerial image patches of 500x500 pixels with a spatial resolution of 1 metre derived from the 2017 national aerial image dataset (PDOK, 2017). For each patch, we consider the housing quality score attributable to the 100-metre patch center. We do this to ensure that the model has the context needed to recreate the housing score, as the scores were created by using data from a 200m² square-metre radius from the cell center. As such, there is a 200m² overlap for each patch with its neighbouring patches.

Ground-level images

We test two data types for our ground-level images. Firstly, we use Google Street View (GSV) panorama images, which are widely used for urban attribute prediction (Naik et al., 2014; Dubey et al., 2016). We use the dataset of GSV panorama images in Amsterdam from (Srivastava et al., 2019). The dataset is designed for building function classification, and as such, each panorama image in this dataset is oriented to directly face a building in the city by using the location and orientation data of the panorama images. After filtering images to the extents of the aerial image patches, we retained 90/256 images. Our second dataset consists of Flickr images. Flickr is freely available and consists of crowdsourced images, making it more flexible and easier to acquire than GSV panorama

¹<https://www.leefbaarometer.nl/>

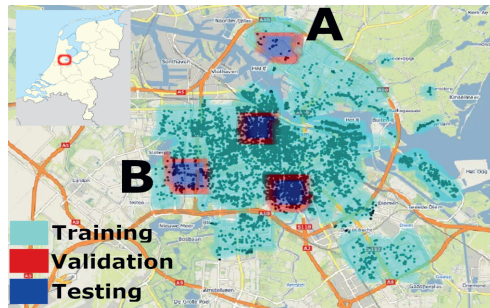


Figure 4.1: Data splits of our experiments over the city of Amsterdam. Testing squares are padded by validation cells to ensure that no test data is seen during training. Cyan squares are for training, red for validation, and blue for testing. Black points represent geotagged photos of the Flickr buildings subset.

images. It is a source of data that is increasingly used to study the environmental factors contributing to individuals’ well-being from a first-person perspective (Havinga et al., 2020). We gathered Flickr images taken between 2004 and 2020 with a geotag located in the city of Amsterdam, which resulted in 54’250 images.

Data splits

We split our dataset into training, validation, and test sets by selecting square regions within the dataset. The edges of these squares partly overlap with the training set as a result of the patch size. We therefore assign the edges to the validation set. The region centres are assigned to the test set to avoid correlation between the sets due to spatial co-location. Our splits are shown in Figure 4.1. For both sources of ground-level images, if no images intersect with an aerial image patch, then we leave the patch out of the subset.

4.2 Methods

Our model is tasked with predicting a housing quality score \hat{s} from data within a patch. It consists of an aerial feature branch and a ground-level feature branch, as shown in Figure 4.2. The ResNet-50 (He et al., 2016) feature extractor of the aerial feature branch is initialized with weights for housing quality prediction over The Netherlands from our preliminary study (Levering et al., 2021a). For ground-level features, we use a ResNet-50 pre-trained ImageNet model, as well as a pre-trained ResNet-50 Place Pulse 2 (PP2) (Dubey et al., 2016) model. Place Pulse 2 is a dataset for urban sentiment analysis consisting of GSV images. The aerial feature branch extracts a 2048-dimensional vector \mathbf{a} from an input aerial image. The ground-level feature branch produces one 2048-dimensional feature vector for each of the N geotagged ground-level images within the patch. We

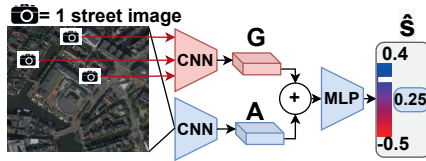


Figure 4.2: Multimodal model predicting housing quality scores. Only the aerial branch and the merging layer shown in blue are trained. Features extracted from the ground-level images are fixed. Depending on the subset, the ground-level image branch uses features extracted from either Google Street View, or Flickr images.

average-pool these vectors to form the ground-level feature vector \mathbf{g} , following the same design as (Srivastava et al., 2020). To merge the feature vectors \mathbf{a} and \mathbf{g} into the feature vector \mathbf{m} , we perform pairwise addition:

$$\mathbf{m} = \mathbf{a} + \frac{1}{N} \sum_{n=1}^N \mathbf{g}_n \quad (4.1)$$

The vector \mathbf{m} is then passed to a two-layer perceptron to extract joint features over the merged vector. We first apply batch normalisation to the features, which are then passed to the first fully connected layer, which produces a 100-dimensional vector. These features are subsequently passed to the final fully connected layer to regress the building score \hat{s} . We train our model using a Mean Squared Error loss calculated over the predicted patch housing score \hat{s} w.r.t. the ground truth patch housing score s :

$$\mathcal{L}_{score} = (s - \hat{s})^2 \quad (4.2)$$

During the first three epochs, only the fully connected layers are trained. Starting at epoch four, the aerial feature extractor is also optimised to fine-tune it to Amsterdam. The ground-level feature extractor is not modified at any stage since it has been pre-trained with images of a similar nature.

4.3 Experimental setup

Beyond reporting the results on the full method using either GSV or one of the Flickr subsets for the ground-level branch, we also perform ablation studies to test the performance of each branch individually. To test the aerial branch, we set the ground-level features \mathbf{g} to be a vector containing zeroes. We do the same to the aerial features \mathbf{a} to test the ground-level features.

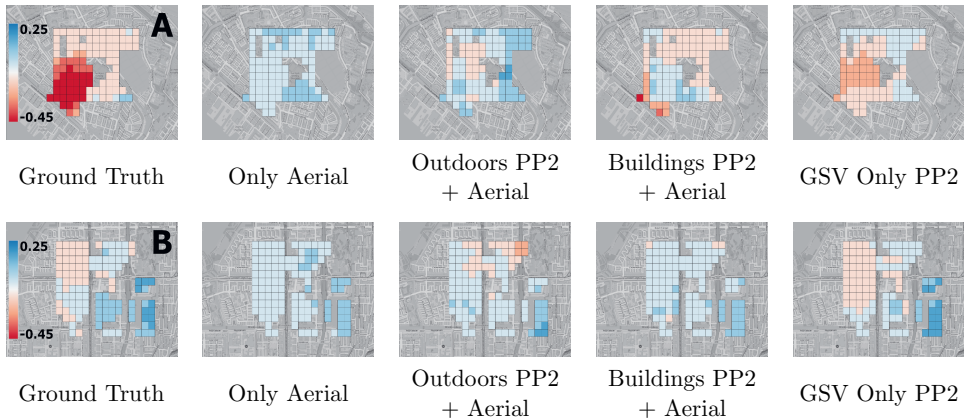


Figure 4.3: Plots of predictions of building quality score for the best model of each data subset on the two most spatially diverse tiles. Their locations are displayed in Figure 4.1. Colours range from red (low-quality) to blue (high-quality).

We train all models with the Adam optimizer for 25 epochs, which we initialise with a learning rate of 0.001 and a weight decay of 0.001. We report the root mean squared error (RMSE) of the housing quality score as well as Kendall’s τ (Kendall, 1938), which is a ranking coefficient between -1 and 1, which indicates whether or not samples are correctly placed in the right order in terms of increasing housing quality score. A value of -1 indicates a perfectly inverse ranking, while a value of 1 represents a perfect ranking.

Filtering Flickr images

As Flickr consists of social media photos that are less organised than GSV images, filtering is necessary to retain only images that are beneficial for building score regression. We apply a pre-trained Places365 (Zhou et al., 2018) model for scene classification to test two filtering methods. As a weak filtering method, we retain only images where 9 out of the 10 most activated classes are marked as outdoors in the dataset. We refer to this subset as **Flickr Outdoors**. In total, the Flickr Outdoors subset contains 34’222 images. Secondly, we select images that have at least one scene strongly related to buildings above a given threshold, for which we consider 24 building-related classes. This threshold was empirically tested and set to an activation of 0.05. This will filter the dataset more aggressively to focus on buildings in favour of the housing quality score. We refer to this subset as **Flickr Buildings**. The Flickr Buildings subset contains 11’774 images. Filtering out images also resulted in some patches having no ground-based images, which had to be excluded. We show our patch distribution per subset in Table 4.1.

Table 4.1: Patches per split for each subset

Subset	Train	Validation	Test	Coverage
Aerial	4'300	570	491	100%
GSV	4'294	570	491	99.88%
Flickr Outdoors	4'255	570	491	98.97%
Flickr Buildings	4'027	538	458	93.69%

4.4 Results and discussion

Using GSV images as ground modality

We show our results when using GSV images for the ground branch in Table 4.2. Overall, we find that using only features extracted using a PP2 model provides the best results, slightly edging out ImageNet features in terms of Kendall's τ , which reaches a value of 0.778. When using GSV for ground-level features, merging with overhead aerial imagery does not provide performance improvements, regardless of the feature extractor. In addition, pre-training the feature extractor on ImageNet or PP2 results in similar performances. Compared to a Kendall's τ of 0.5810 obtained from using only aerial images, there is a 30% performance gap.

Using Flickr images as ground modality

Table 4.3 shows the results obtained from the Flickr ground images. Our results show substantial differences between the three modalities. The least competitive result in terms of Kendall's τ occurs when merging the aerial image features with Outdoor Flickr images using ImageNet features, 0.576. The best unimodal setting with Flickr images consists of using PP2 features from the Flickr Buildings subset, reaching a Kendall's τ of 0.649. For unimodal prediction, it is important to simultaneously use the Flickr Buildings images along with PP2 pre-training, since using either the Flickr Outdoors subset or only ImageNet pre-training results in a loss of performance, down to 0.596 and 0.602, respectively. Adding the aerial branch to the PP2 pre-trained Flickr Buildings model results in another increase in performance, up to 0.686. By using a combination of Flickr Buildings PP2 features and aerial features, the performance gap compared to the best GSV model is halved.

By comparing the metrics of the three subsets, we can assess the suitability of alternatives to be used instead of GSV. While GSV image features prove to be most suitable for building quality at the city scale, it is encouraging that Flickr and aerial images are able to close the performance gap. Furthermore, while GSV images are more suitable for urban analyses, the data is often difficult to acquire for larger areas or even entirely unavailable. In contrast, Flickr images are easy to acquire and widely available. The success of using Flickr shows that general-purpose social media data sources can also be

Table 4.2: Metrics when using GSV as the ground image source

Modality	Ground Features	Kendall's τ
Aerial	n.a.	0.5818
GSV	ImageNet	0.7651
GSV	PP2	0.7780
Aerial & GSV	ImageNet	0.7699
Aerial & GSV	PP2	0.7656

Table 4.3: Metrics when using Flickr as the ground image source

Modality	Ground Features	RMSE	Kendall's τ
Aerial	n.a	0.1314	0.5810
Outdoor	ImageNet	0.1100	0.5755
Outdoors	PP2	0.1155	0.5962
Buildings	ImageNet	0.1179	0.6024
Buildings	PP2	0.1031	0.6487
Aerial & Outdoors	ImageNet	0.1427	0.5729
Aerial & Outdoors	PP2	0.1306	0.6215
Aerial & Buildings	ImageNet	0.1142	0.6243
Aerial & Buildings	PP2	0.1104	0.6862

used, as they are easier to scale over larger areas, for instance, through crowdsourcing efforts.

Spatial predictions

In Figure 5.3, we show the spatial predictions for the best model of each subset for the two most spatially diverse testing tiles. The maps show that the PP2-only GSV model is able to approximate the ground truth most accurately in both tiles. Both Flickr models struggle with predicting the extent of the low-quality housing in tile A. This may be caused by a lack of images in the north of Amsterdam, which can be seen in the black points of Figure 4.1, which represent images of the Buildings subset. This area is less popular with tourists, which may explain the lack of Flickr photos. The GSV dataset has much better coverage in this area, which is reflected by the better prediction quality for this tile.

4.5 Conclusions

In this paper, we use a combination of features extracted from ground-level and aerial images to predict the quality of houses in Amsterdam. For the ground-level images, we tested two pre-trained feature extractors, one on ImageNet and one on Place Pulse 2, a dataset for subjective perception of urban ground-level images. We collected and refined three ground-level image datasets: Google Streetview (GSV), Flickr Outdoors, and Flickr Buildings. The latter two were obtained by filtering geotagged Flickr images. Using only GSV images resulted in the best overall performance, providing a 30% increase in Kendall's τ with respect to using only aerial imagery. This suggests that the nature of GSV imagery is well-suited, as it captures 360° panoramas of most of the city's streets at regular intervals. However, this type of imagery is costly to obtain and not always available. Our results show that using less curated but more easily available social media images such as Flickr can still provide a 15% increase in performance w.r.t. the aerial imagery if both the images and the feature extractor are carefully selected for the task.

Chapter 5

Prompt-guided and multimodal landscape scenicness assessments with vision-language models

This chapter is based on the following submitted article:

Levering, A., Marcos, D., Tuia, D., Jacobs, N., 2023. Prompt-guided and multimodal landscape scenicness assessments with vision-language models, *PLoS One*.

Abstract

Recent advances in deep learning and Vision-Language Models (VLM) have enabled efficient transfer to downstream tasks even when limited labelled training data is available, as well as for text to be directly compared to image content. These properties of VLMs enable new opportunities for the annotation and analysis of images. We test the potential of VLMs for landscape scenicness prediction, i.e., the aesthetic quality of a landscape, using zero- and few-shot methods. We experiment with few-shot learning by fine-tuning a single linear layer on a pre-trained VLM representation. We find that a model fitted to just a few hundred samples performs favourably compared to a model trained on hundreds of thousands of examples in a fully supervised way. We also explore the zero-shot prediction potential of contrastive prompting using positive and negative landscape aesthetic concepts. Our results show that this method outperforms a linear probe with few-shot learning when using a small number of samples to tune the prompt configuration. We introduce Landscape Prompt Ensembling (LPE), which is an annotation method for acquiring landscape scenicness ratings through rated text descriptions without needing an image dataset during annotation. We demonstrate that LPE can provide landscape scenicness assessments that are concordant with a dataset of image ratings. The success of zero- and few-shot methods combined with their ability to use text-based annotations highlights the potential for VLMs to provide efficient landscape scenicness assessments with greater flexibility.

5.1 Introduction

In these times where urban expansion is prevalent, maintaining the quality of our landscapes is increasingly important, as it plays a significant role in our overall well-being. Beyond their visual appeal, scenic landscapes offer a multitude of benefits, both tangible and intangible. They provide us with a sense of tranquility, an escape from the stress of our daily lives, and a connection to the natural world. Moreover, research has shown that exposure to scenic environments is associated with many positive effects. Exposure to natural environments is shown to be beneficial to our attention span (Velarde et al., 2007; Berman et al., 2008), our stress management (Velarde et al., 2007; Roe et al., 2013), and our overall happiness (Seresinhe et al., 2019). Scenic landscapes also instill a sense of comfort, tranquility, and safety (Galindo and Rodriguez, 2000). Beyond personal health benefits, scenic landscapes are a driver for tourism (Krippendorf, 1984), as well as cultural ecosystem services (Daniel et al., 2012; Havinga et al., 2021).

One way to protect natural environments is to document their presence and evaluate their aesthetic appreciation by humans, or *scenicness*. Improvements in computer vision methods in the past decade have made it possible to predict scenicness directly from images. Meanwhile, increased internet connectivity across the globe has enabled crowdsourcing at

unprecedented scales. These developments combined have resulted in the exploration of landscape aesthetic preferences directly from images in a data-driven setting (Seresinhe et al., 2015; Dubey et al., 2016; Biljecki and Ito, 2021). However, such research efforts in turn suffer from the large amounts of annotated images needed to train deep-learning models. As a result, only a few studies have attempted to study aesthetic preferences on a local scale using deep learning and data-driven methods.

Recent research developments have seen the convergence of natural language processing and computer vision into *Vision-Language Models* (VLM), such as the CLIP model (Radford et al., 2021). Trained on images gathered from the internet with their corresponding text captions, these models are able to relate the content of images to textual descriptions to determine their relatedness. CLIP and similar models have since closed the gap between data-efficient learning strategies such as few-shot learning and fully supervised training on large datasets. Rather than needing tens of thousands of annotated examples, VLMs are able to use a fraction of data and still perform competitively on many tasks (Radford et al., 2021; Zhou et al., 2022b; Song et al., 2022). Therefore, VLMs hold the potential to enable accurate data-driven analyses of many tasks using small-scale datasets.

In this study, we evaluate the effectiveness of VLMs for quantifying and mapping landscape scenicness at scale, using both rated image datasets and a new text-based annotation approach. Firstly, we explore the potential of VLMs for landscape scenicness assessments using data-efficient learning regimes using a dataset of rated images. We explore a few-shot prediction setting with linear probes and a zero-shot setting using prompts of opposing landscape scenicness concepts. Secondly, we introduce a new annotation method that leverages the ability of VLMs to associate text with images, which we refer to as *Landscape Prompt Ensembling* (LPE). We demonstrate that ensembles of rated text descriptions provided by volunteers can provide landscape scenicness assessments without the need for an image dataset while annotating.

5.2 Related works

5.2.1 Scenicness prediction with machine learning

The prediction of landscape aesthetic qualities from images became possible with the rapid improvements of machine- and deep learning models in the last decade.

Dubey et al. (Dubey et al., 2016) introduced the Place Pulse 2.0 dataset, where labelers were shown two images of urban streetscapes and asked to vote on their preferred image with respect to six different qualities. Among the adjectives that volunteers were asked to rate was *beautiful*. Subsequent research has explored a variety of topics, such as relating adjective predictions to objects in urban environments to determine their influence (Zhang et al., 2018), determining their spatial patterns through land use classes (Wei et al., 2022),

and using them to synthesise ideal neighbourhoods (Wijnands et al., 2019). In further research concerning the aesthetic quality of urban spaces, Verma et al. (Verma et al., 2018) crowdsourced scenicness ratings on a local scale and used these ratings to measure the effects of changing conditions on scenicness. Christman et al. (Christman et al., 2020) relate objects that are implicitly assumed to be scenic or unsightly (e.g., flowers or trash bags) to the walkability of neighborhoods. Chen and Biljecki (Chen and Biljecki, 2023) predict the design and aesthetic quality of urban areas in Singapore and compare their importance to features of the landscape.

Directly related to our research is the research performed on the estimation of scenicness as a quantitative score. The dataset used for this purpose is the ScenicOrNot (SON) dataset (Seresinhe et al., 2015), a crowdsourcing effort to rate the scenicness of images across the entirety of Great Britain and the Isle of Man. Early works trained convolutional neural networks to regress the scenicness score directly (Seresinhe et al., 2017). Subsequent research has largely focused on understanding how scenicness relates to its environment. Havinga et al. (Havinga et al., 2021) first extracted landscape features such as their scene class, then related them to scenicness through a Random Forest approach. Further research has attempted to use CNNs for explicit relations between intermediate concepts such as scenes or land cover to relate them to scenicness in an interpretable manner (Marcos et al., 2019; Marcos et al., 2021; Levering et al., 2021a). Finally, Arendsen et al. (Arendsen et al., 2020) attempted to discover how concepts that have not been trained on relate to scenicness. SON has also been used to study scenicness from new perspectives, such as a hybrid ground-and-overhead imagery perspective (Workman et al., 2017), as well as through satellite imagery (Levering et al., 2021b).

5.2.2 Vision-language models

While much attention has been devoted to classification tasks using VLMs and data-efficient learning methods (Radford et al., 2021; Zhou et al., 2022a; Gabeff et al., 2023), comparatively little has been done to develop methods compatible with regression tasks. Li et al. introduced OrdinalCLIP, which uses an ordinal output space in order to utilise the classification-based few-shot methods (Li et al., 2022). Hentschel et al. (Hentschel et al., 2022) trained a linear probe on CLIP image features on the regression task of image photographic aesthetics understanding in a few-shot setting, with competitive results compared to a fully-trained baseline. Ke et al. (Ke et al., 2023) introduced VILA, a VLM fine-tuning and zero-shot prediction pipeline for image aesthetics. The authors fine-tuned a pre-trained VLM on image-text captions, where the text captions provide feedback on the given image about its aesthetic properties or qualities. The authors then test the zero-shot regression performance of their model by using the contrast between two prompts (e.g., "good photo" versus "bad photo"), where the model's confidence in the positive prompt is assumed to be correlated with photographic aesthetic beauty. Their method delivered

competitive results compared to fully-trained baselines. The authors also experiment with other tasks, such as photo-aesthetic captioning.

5.3 Data

5.3.1 ScenicOrNot

For the prediction of landscape scenicness, we require a reference dataset with image ratings. For this purpose, we use the ScenicOrNot (SON) dataset (Seresinhe et al., 2015) as our reference dataset. SON consists of a collection of approximately 217 000 images with ratings provided by anonymous volunteers on a scale between 1 (most unsightly) and 10 (most beautiful). The images used by SON are obtained by the Geograph UK project¹. The images are stored with geolocation information and can therefore be used for mapping. To acquire scenicness ratings for the images, the authors then used a crowdsourcing website², which loads a random image and asks the visitor to rate it. Images were only included in the dataset if they had at least 3 ratings. The SON website provides a file that contains the ratings of each image, along with the image path on the Geograph website. After removing images that had been taken offline from the ratings file, we are able to use a total of 212 104 images. We show the spatial distribution of the rated images and corresponding scores in Figure 5.1.

5.3.2 Landscape prompt annotations

For the SON dataset, volunteers were asked to provide ratings of images directly, which creates a dataset with image ratings. Such an approach provides accurate labels for individual images, but implicitly, such ratings are not informative about individual voter aesthetic preferences for certain landscape types. Instead, we propose to annotate a dataset that captures the explicit landscape preferences of each volunteer by leveraging the ability of VLMs to relate text descriptions to image content. We refer to our annotation method as *Landscape Prompt Ensembling* (LPE). To create a prototype dataset, we involved a group of non-expert volunteers and asked them to provide prompt-rating pairs that describe the landscapes of the United Kingdom through an anonymous survey. In this survey, we showed four example images of scenes with descriptions and ratings that voters could imagine. We then asked them to imagine and write out their own landscape impressions of the United Kingdom. As such, in our annotation process, we do not need an extensive image dataset during annotation, instead relying on the volunteers' imagination. We also surveyed the confidence that voters have in their ratings and whether or not they have visited the United Kingdom before. In total, we received 45 responses. We removed empty responses and responses that did not provide prompts in the requested format. Of

¹<https://www.geograph.org.uk/>

²<https://scenicornot.datasciencelab.co.uk/>

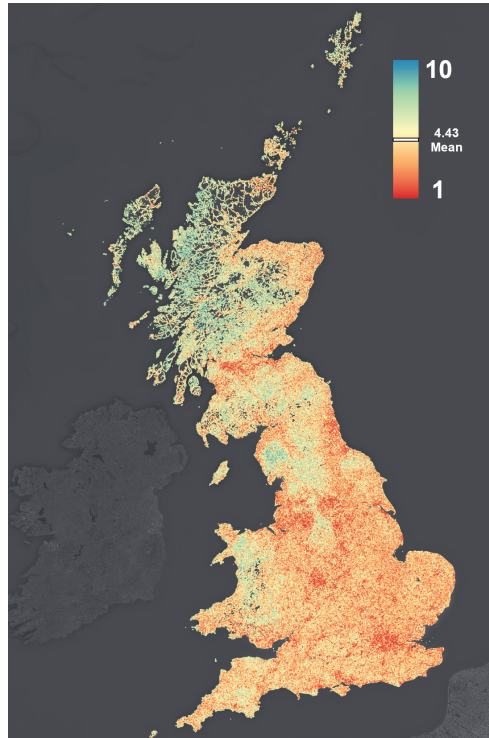


Figure 5.1: ScenicOrNot image ratings plotted at their georeferenced coordinates for the entirety of Great Britain and the Isle of Man. Values range from 1 to 10, where 10 is the most scenic, with an average scenicness rating of 4.43.

these 27 remaining responses, the median number of prompts provided was 4. In total, voters provided 136 prompts. 18 voters provided more than 1 prompt at once. Of the 19 respondents for which confidence information is available, the average voter confidence was 3.36 out of 5, and 57% of the labelers had visited the United Kingdom before. We give an overview of our annotation process in Figure 5.2. We show how our LPE annotations can be used to generate image ratings, such as in the SON dataset, through early ensembling (Section 5.4.4) and late ensembling (Section 5.4.4).

5.4 Methods

In our research we explore how well CLIP (Section 5.4.1) can be used in efficient data regimes for the task of scenicness prediction. For this purpose, we use recent advances in model pre-training and we test the utility of the text encoder of our VLM. After testing the robustness of CLIP features in a few-shot setting (Section 5.4.2), we propose a method for

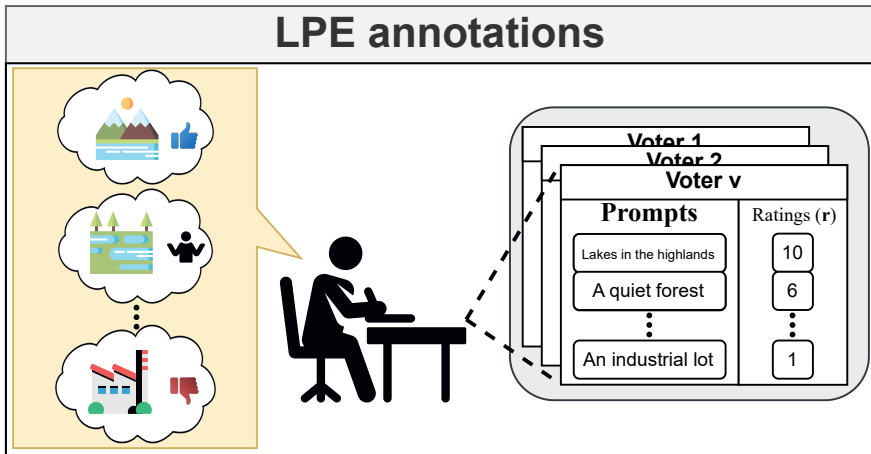


Figure 5.2: Rated landscape prompt annotation workflow. Voters are asked to imagine landscapes that they like or dislike. They are then asked to write a description of these landscapes and to give it a rating between 1 and 10. The resulting dataset is a collection of landscape descriptions and their associated ratings from every voter.

zero-shot prediction based on contrastive prompting in section 5.4.3). We then demonstrate the power of VLMs in assessing landscape scenicness through text descriptions by using LPE (Section 5.4.4), where we generate image ratings from our LPE annotations.

5.4.1 CLIP

The feature extractor we use in our experiments is a CLIP-pre-trained VLM vision transformer (Radford et al., 2021). This model consists of an image feature extractor and a text feature extractor, which have been jointly optimised during pre-training so that they share the same embedding space. The image encoder encodes each image to a vector \mathbf{x} , and the text encoder encodes each textual prompt to a vector \mathbf{t} . The vision/text embedding space learned by CLIP is multimodal and aligned, in the sense that a prompt and a corresponding image would be mapped in the same location of the embedding space. We use the image and text encoders as provided in the checkpoints of OpenAI (details in Section 5.5).

5.4.2 Few-shot learning of CLIP features

In this setting, we study how well VLM models can predict scenicness using only the image encoder through few-shot learning. In other words, we only consider the image feature vector \mathbf{x} extracted by the image feature extractor of CLIP without using any text features. We optimise a single linear layer, which maps each image feature vector to a scenicness

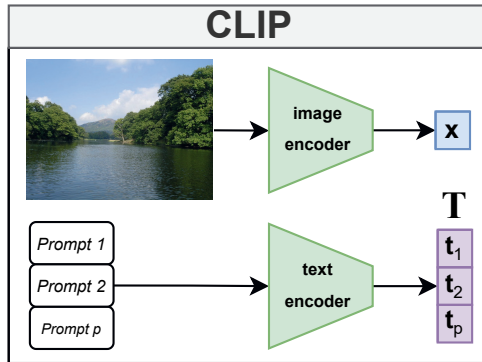


Figure 5.3: Overview of the CLIP model. CLIP uses two separate encoders, which map images and text to the same latent space.

score prediction \hat{s} :

$$\hat{s} = \mathbf{x}^T \mathbf{w} + b, \quad (5.1)$$

where \mathbf{w} is a vector of weights and b is a bias term. The linear layer is then optimized using a squared error loss:

$$L_{scenic} = (s - \hat{s})^2, \quad (5.2)$$

where s is the SON rating matching the input image.

5.4.3 Contrastive prompting

In our second experiment we test the zero-shot capabilities of the text encoder. In this setting we explore which prompt formulations are suitable for predicting landscape scenicness. Our method uses a shared prompt context with a pair of antonyms to derive the model’s preference for the positive concept of the antonym pair. The prompt context is a general sentence such as "A photo of a landscape that is [...]", where the text in brackets is replaced by either synonyms or antonyms of scenicness. The use of this prompt is intended to provide good discriminative text features for the task, and its importance was first demonstrated by Zhou et al. (2022b). For a given set of two prompts comprised of one scenicness synonym and one scenicness antonym, we first calculate the text feature activation matrix \mathbf{T} of size $(d_t \times 2)$, where d_t is the number of features of the text encoder of the VLM. For a given image activation vector \mathbf{x} of size $d_i = d_t$, we can then calculate the logits and use a Softmax activation function to determine the activation of the prompts for the image under consideration:

$$\mathbf{a} = \text{Softmax}((\mathbf{x}^T \mathbf{T}) * Z), \quad (5.3)$$

where the scaling factor Z has been estimated empirically during the pre-training of CLIP. Note that the activations of \mathbf{a} sum to 1. We can assume that the model’s confidence in the positive sceniness synonym prompt at index 1 is linearly related to the sceniness of a given image. We rescale the model’s confidence in the positive prompt at index 1 to the 1 to 10 range of the sceniness reference data to derive the predicted sceniness score \hat{s} :

$$\hat{s} = (a_1 * 9) + 1 \quad (5.4)$$

The resulting predictions can then be compared to the reference sceniness scores to determine the performance of the given prompt construction. As this is a zero-shot method, no parameter updates are performed. We give a graphic overview of our approach in Figure 5.4.

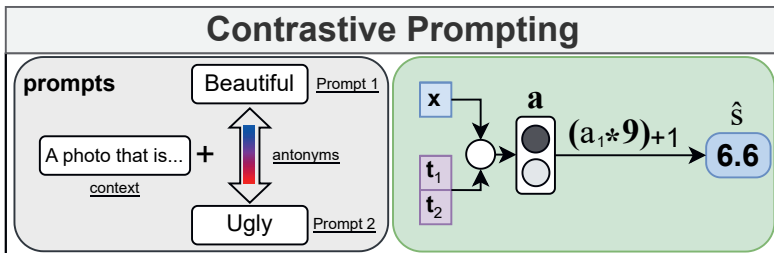


Figure 5.4: Prediction pipeline for the contrastive prompting method. We first define a positive and a negative prompt with a shared prompt context. Then, we use the model’s confidence in the positive prompt and rescale it between 1 and 10 as the sceniness prediction \hat{s} for a given image.

5.4.4 Landscape prompt ensembles

In this experiment, we test two ways of deriving image sceniness ratings from the rated text prompt annotations described in Section 5.3.2. In this annotation format, each voter $v \in V$ provides a list of landscape prompts with matching ratings. For each voter v , we can extract a matrix of text features \mathbf{T}_v from the prompts with CLIP. For each voter, we also have a vector of ratings, \mathbf{r}_v . We test how well our prompt-based annotations can be used to generate image ratings through two types of ensembling methods, *late ensembling* and *early ensembling*, which calculate the prompt activations in differing ways.

Early ensembling

With early ensembling, we hypothesise that if many voters provide many prompts, then there will be a few prompts that most accurately describe each image. For instance, a photo that depicts a river in a forest benefits from having the river mentioned, as it provides a more detailed description than just a prompt about forests in general. For this purpose, we consider all prompts provided by all voters at once. First, we concatenate the

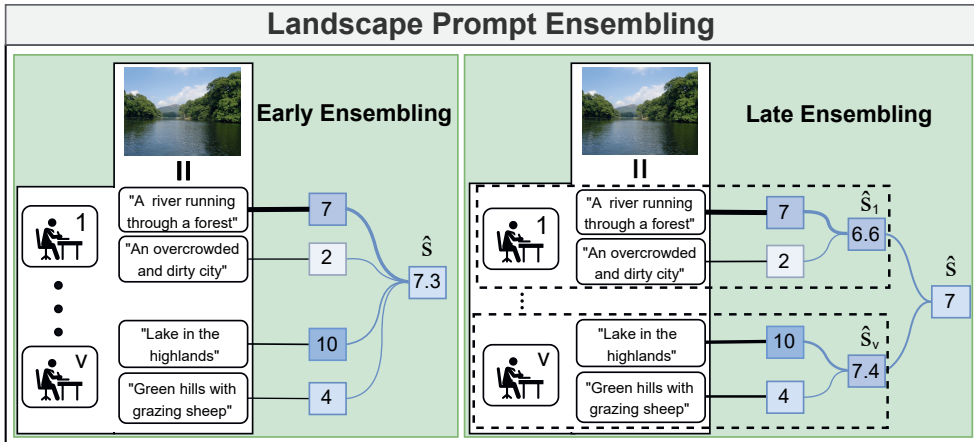


Figure 5.5: Methods of ensembling for generating image ratings from landscape prompts. In early ensembling, we use the likelihood that any given prompt matches an image. The likelihood of each prompt is then multiplied by its voter-provided scenicness rating to determine the scenicness score of a given image. In late ensembling, we consider this weighted likelihood for the prompts of each voter individually to calculate a voter-specific scenicness score. We then average across all voter scenicness scores to calculate the scenicness score for the image.

encoded text feature matrices of the prompts of all voters into a single text feature matrix \mathbf{T} :

$$\mathbf{T} = [\mathbf{T}_1 \quad \mathbf{T}_2 \quad \dots \quad \mathbf{T}_V]. \quad (5.5)$$

We can then calculate the activations of each prompt with Eq. (5.3) to derive the prompt activation vector \mathbf{a} . In this setting, \mathbf{a} contains the activations of all prompts provided by all raters for the given image through a single Softmax. As such, it represents the probability that each prompt best matches the image. We then concatenate the ratings of all voters into a single ratings vector \mathbf{r} such that it matches the shape of the activation vector \mathbf{a} :

$$\mathbf{r} = [\mathbf{r}_1 \quad \mathbf{r}_2 \quad \dots \quad \mathbf{r}_V]. \quad (5.6)$$

By multiplying the probability that each prompt best matches the given image with its provided scenicness rating, we can then compute the predicted scenicness score of the image from all contributions:

$$\hat{s} = \mathbf{a}^\top \mathbf{r}. \quad (5.7)$$

Late ensembling

In the late ensembling case, we hypothesise that having many voters with less detailed prompts will capture the variance of landscape preferences on a macro-scale, similar to

the variance observed for the ratings of individual images in SON. For each voter $v \in V$, we first extract the activations of their provided prompts:

$$\mathbf{a}_v = \text{Softmax}((\mathbf{x}^\top \mathbf{T}_v) * Z). \quad (5.8)$$

To calculate the scenicness score for a given image, we can then multiply the activation vector of each voter with the ratings of the voter and take the average across the total number of voters to calculate the image scenicness score:

$$\hat{s} = \frac{\sum_{v=1}^V \mathbf{a}_v^\top \mathbf{r}_v}{V} \quad (5.9)$$

5.5 Experimental set-up

For our VLM, we use the ViT-L/14 variant of the CLIP pre-trained models (Radford et al., 2021). We freeze the feature extractors and do not optimise them during any of our experiments. We evaluate our methods using the root mean squared error (RMSE), the coefficient of determinant R^2 and Kendall’s τ , which is a ranking coefficient ranging from -1 (all values are inversely ranked) to 1 (all values are ranked perfectly in order) (Kendall, 1938). We release the code for our experiments on GitHub³.

5.5.1 Few-shot learning

We run experiments with a total number of samples of $n \in (25, 50, 100, 250, 500)$. We cluster all landscape photos in SON, run k-Means ($k = 25$), and pick n/k samples of each cluster. We optimise the linear layer using stochastic gradient descent. We use 5-fold cross-validation over the learning rates $5e-3$, $2.5e-3$, and $1e-3$, and we use the model with the best training R^2 during testing, as the R^2 is more stable than the RMSE on this task. We compare the performance of the ViT-L/14 model to an ImageNet-pretrained ConvNeXt-Large model (Liu et al., 2022) to determine the effect of web-scale pre-training. We use the same training regime for this model. Lastly, we compare the performance of both few-shot trained models to a ConvNeXt-Large model trained on the entire dataset. We train this baseline model with a learning rate of $1e-3$ with the Adam optimizer (Kingma and Ba, 2014). We randomly sample 10% of the dataset for testing. Of the remaining 90% of the dataset, we use 85% for training and 15% for validation.

5.5.2 Contrastive prompting

We test different prompt configurations in our contrastive prompting experiments. We consider six prompt contexts: *"A photo that is"*, *"A photo that is extremely"*, *"A photo of an area that is"*, *"A photo of an area that is extremely"*, *"A photo of a landscape*

³https://github.com/ahlevering/prompt_guided_scenicness

that is", "A photo of a landscape that is extremely". We test these prompt contexts to evaluate two aspects, namely 1) the importance of emphasis on landscapes ("area" and "landscape"), and 2) the effect of adding a superlative ("extremely"). For the positive and negative concepts, we use the synonyms and antonyms of "scenic" as listed on a thesaurus website⁴. The synonyms that we use are "breathtaking", "grand", "spectacular", "striking", "dramatic", "panoramic", "impressive", "beautiful", and "picturesque". For the antonyms, we use "normal", "usual", "dreary", "ugly", "ordinary", "despicable", and "gloomy". We test all possible combinations of context, positive, and negative concept choices, which results in 378 unique prompt configurations. To determine the stability of prompt configurations we compute the metrics on the full dataset as well as the samples used in the 25-sample few-shot learning case. In doing so, we test if a single representative sample from each type of landscape in the SON dataset will result in similar highly-performing prompt configurations compared to the full dataset.

5.5.3 Landscape prompt ensembles

Both the early and late ensembling methods do not require parameterization and can be run as-is. We instead analyse the trade-off between the number of voters and the number of prompts they provide in the case of late ensembling. We test the effect of using ensembles with a minimum of 2, 5, 8, or 10 prompts. In doing so, we can test if it is better to have many voters with a small number of prompt suggestions or to have a few voters who provide more accurate prompt suggestions.

5.6 Results and discussion

5.6.1 Few-shot experiments

Figure 5.6 illustrates the performance of the few-shot linear probes. Our results suggest that it is possible to accurately predict scenicness with far fewer labelled examples than has previously been attempted. Compared to the fully-trained baseline model, which was pre-trained on ImageNet, we find that the linear probe based on the ViT-14/L CLIP pre-trained transformer is nearly as accurate when using $n = 500$ labelled samples. However, despite requiring approximately 342 times fewer training samples than the fully-trained baseline, this can still be considered a non-trivial labelling effort as all of the images in SON are rated by at least 3 voters. Reducing n further, we observe that a model that uses just $n = 100$ samples has approximately 10% lower accuracy but uses 5 times fewer labels. Our findings suggest that a few-shot linear probe can provide an adequate estimation of landscape scenicness even with a limited number of samples.

⁴<https://www.thesaurus.com>

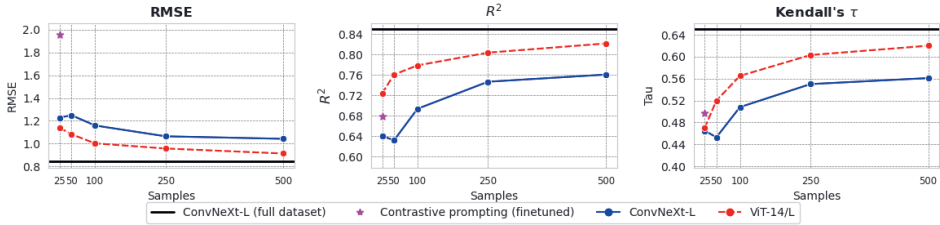


Figure 5.6: Results for the few-shot linear probes. The black line shows the performance of a ConvNeXt-Large model initialised with ImageNet weights trained on the complete SON dataset. The blue line represents the same model, where only the linear probe is fine-tuned in a few-shot setting. The red line shows the performance of the ViT-14/L vision transformer pre-trained using CLIP. While both the ConvNeXt-Large and ViT-14/L models provide adequate few-shot performance, the transformer model with web-scale pre-training and more parameters performs substantially better. When using 25 samples to estimate the best prompt combination, our zero-shot contrastive prompting method (shown in magenta) shows superior ranking performance compared to the few-shot models, although predictions are further off from the reference values as evidenced by the high RMSE.

5.6.2 Contrastive prompting

In Figure 5.7, we show the performance distribution for each choice of prompt context. From the distributions over the full dataset, we observe that the method is sensitive to the exact formulation of the prompt and that both the context and the choice of synonyms and antonyms are of importance. The best-performing prompt context on our dataset ("*A photo of an area that is extremely*") still has outliers, which produce a poor fit. In the best-case scenario, the contrastive prompting method outperforms the few-shot probe on the task of ranking samples when only a limited number of samples are available to train the probe.

In the pure zero-shot setting, it is not possible to know a priori which prompt combination is optimal. In Figure 5.7, we therefore also show the performance of each prompt context when applied to the samples of the $n = 25$ case of the few-shot setting, which we refer to as the *calibration set*. The resulting distributions for each prompt context are highly similar, which suggests that a few labelled samples may be used to tune the prompt configuration. We show the metric performance of the best prompts on the fine-tuning set in Table 5.1. We take the best-performing prompt configuration and apply it to the full dataset. Figure 5.6 highlights that it outperforms a linear probe on the task of ranking samples, but it also has a very high RMSE and a marginally worse R^2 . These results suggest that the optimal prompt configuration for this task is maximally discriminating between the positive and negative concepts.

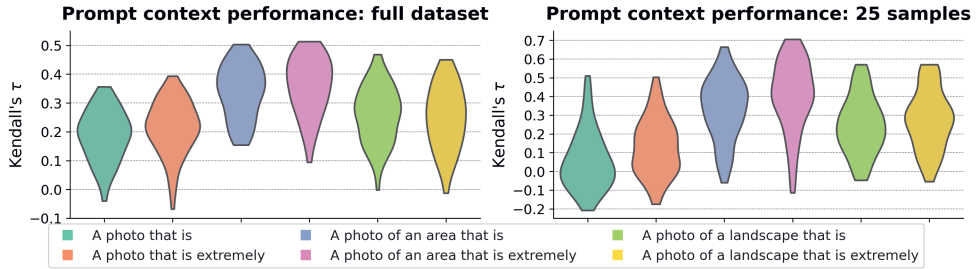


Figure 5.7: Distribution of Kendall’s τ of each of the six prompt contexts of the contrastive prompting method when evaluated on the entire dataset (left) and when applied to only the $n = 25$ samples from the few-shot learning case (right).

Table 5.1: Metric performance of the top-five contrastive prompting configurations evaluated on the calibration set. The highly-performing prompts in this setting are similar to those observed for the full dataset.

Context	Positive concept	Negative concept	RMSE	R^2	τ
area that is extremely	breathtaking	usual	1.52	0.841	0.705
area that is	panoramic	normal	2.31	0.801	0.664
area that is extremely	panoramic	ugly	2.64	0.813	0.658
area that is extremely	breathtaking	normal	3.43	0.744	0.644
area that is extremely	panoramic	usual	3.50	0.729	0.644

5.6.3 Prompt ensembles

In this section, we compare the computed image scenicness ratings of our LPE ensembling methods to the SON image ratings. In Table 5.2, we show the numerical metrics of both of our prompt ensembling methods when compared to SON labels. The results demonstrate that late prompt ensembling is a more effective ensembling method than early prompt ensembling by a substantial margin. We hypothesise that this is the result of the implicit variance in the appreciation of landscapes between voters, which is not accounted for during early ensembling. Even if the VLM model retrieves the most accurate prompt for a specific image with high confidence, the rating assigned to this image could be strongly influenced by the voter’s personal preferences. Therefore, late ensembling appears to be the more reliable method for generating image ratings from the rated prompt annotations. The numerical comparison with SON also suggests that it is better to have many voters provide prompts as opposed to a small subset of voters who provide many prompts. Interestingly, the quality or diversity of each ensemble appears to matter less

Table 5.2: comparison of our LPE method with SON image scores. Late fusion results in image scenicness ratings that are closer to the SON image ratings, and including more voters results in a higher degree of agreement with SON, even if these voters provide fewer prompts per person.

Method	voters	Total Prompts	RMSE	R^2	τ
Early	27	137	3.28	0.535	0.377
Late					
>= 2 prompts	18	129	2.49	0.684	0.475
>= 5 prompts	10	105	2.29	0.657	0.453
>= 8 prompts	5	74	3.46	0.631	0.456
>= 10 prompts	3	49	3.54	0.620	0.437

than the number of individual ensembles. Therefore, a higher number of respondents is more important than ensuring that all respondents provide many prompts at once. Using our small dataset, we even find that late prompt ensembling can be more effective at acquiring scenicness rankings of images than fine-tuning a linear probe using 25 or fewer image labels. Our results demonstrate that LPE shows potential for landscape scenicness assessments without the need for an image dataset during annotation.

5.6.4 Geographical prediction patterns

In Figure 5.3, we show the spatial prediction results for all methods used in our study. We observe a few notable differences between methods. In the few-shot learning setting, the main difference between using $n = 25$ samples (1 examples per cluster centroid) and $n = 500$ samples (20 examples per cluster centroid) is an increase in the model’s ability to predict values at both ends of the distribution, such as the very low scenicness in cities and the very high scenicness of the Scottish highlands. When considering the zero-shot contrastive prompting approach, we observe that the most effective prompt combinations are all highly discriminative between the scenicness synonyms and antonyms. The resulting predictions do follow the distribution of scenicness of the reference data, but with hardly any predictions in the middle of the scenicness distribution. This explains the high RMSE that can be seen in Figure 5.6, despite both its R^2 and Kendall’s τ being competitive compared to the few-show methods. It is therefore likely better to use quantile maps to display scenicness predictions for contrastive prompting. At the country level, all of the presented methods show the most important scenicness patterns, such as low scenicness in cities and high scenicness in elevated areas and shrublands.

We further study the image scenicness ratings obtained by our LPE late ensembling method by analysing the ranking of land cover classes. We sample the level-2 land cover class of the 2018 CORINE inventory (EU Copernicus Program, 2018) at the geolocation of each image within SON. In Figure 5.8, we show a comparison between the SON image

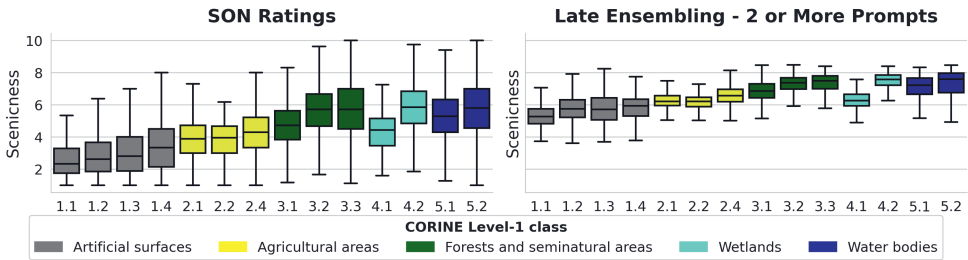


Figure 5.8: Comparison of ground truth average image ratings as plotted for each class in SON (left) with the image ratings generated by the LPE method that most closely matched SON in ranking performance (right). The relative rankings for each land cover class are highly similar, though the LPE mean rating is far higher than in SON.

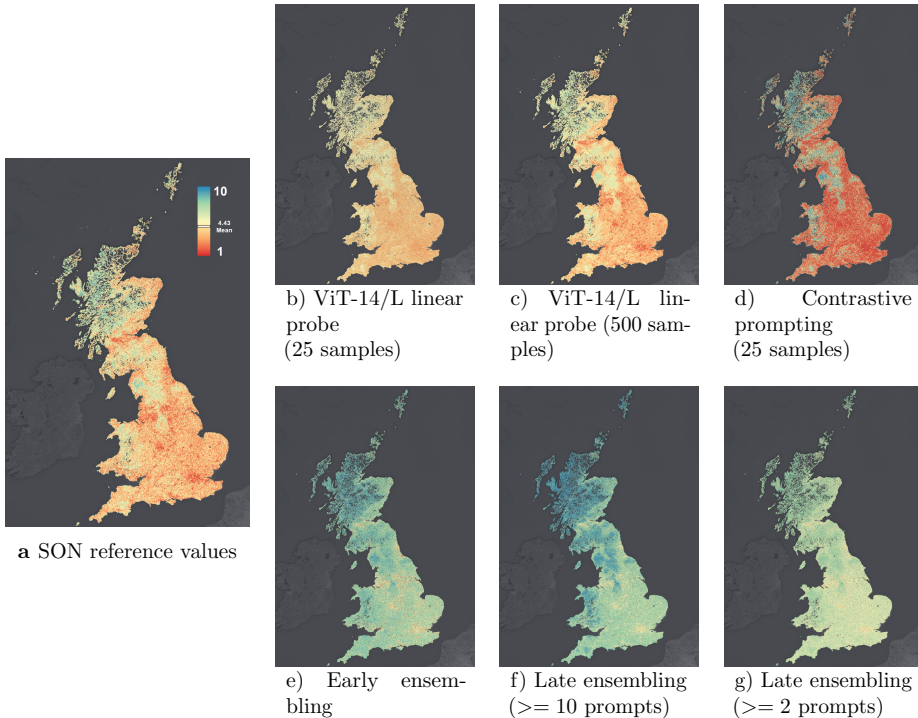
ratings (left) and the ratings calculated by LPE (right) for each land cover class. The plot indicates that there is strong agreement in the ranking of the scenicness of most land cover classes between SON and LPE. One notable exception is that class 1.2 “*Industrial, commercial, and transport units*” is considered to be more scenic by our LPE method when compared to the SON image ratings. However, we do not observe an obvious reason for this increase, such as voters consistently rating industrial elements higher. Further research will be needed to discover the cause.

5.7 Conclusions

In this paper, we studied the potential for VLMs to reduce the labelling dependency of data-driven scenicness prediction. Firstly, we studied the potential of the image encoder of a CLIP pre-trained transformer model to provide good features for scenicness prediction when fine-tuned with 25 to 500 samples from the ScenicOrNot dataset. Our findings prove that a linear probe fine-tuned on 500 samples is just 5% less accurate than an ImageNet pre-trained model trained on 342 times more samples (Kendall’s τ of 0.625 compared to 0.651).

Secondly, we explored the potential of zero-shot contrastive prompting, which uses a shared prompt context and a synonym and antonym for scenicness (e.g., “beautiful” and “ugly”) to find the model’s preference for the synonym, which provides a continuous scenicness score. We determined that the choice of prompt context is important for the performance of the method and that a suitable prompt configuration can be discovered using representative samples from the final dataset. The best-performing contrastive prompt configuration tuned to 25 samples outperforms a linear probe when ranking samples. However, it has the tendency to predict extreme values, and as a result, it has a higher standard error.

Table 5.3: Maps of all methods tested in our research compared to the SON reference data (shown in panel a)). The first row (panels b-d-f) showcases methods based on the image ratings of SON, while the second row (panels c-e-g) showcases methods based on the descriptions provided by our volunteers. The predicted scenicness ratings of each method vary greatly, though the main trends between the models are similar, e.g., rugged wilderness being considered more beautiful than man-made areas such as cities.



Lastly, we introduced Landscape Prompt Ensembling (LPE), a new method for the annotation and prediction of landscape scenicness that uses rated textual descriptions of landscapes. Through a small-scale survey, we asked volunteers to provide text descriptions of landscapes they liked or disliked within the United Kingdom, along with a rating for the description. Our ensembling methods then use the confidence of the VLM that a prompt matches an image multiplied by its provided user rating to determine the scenicness contribution of that prompt. We tested if it is better to find the best prompts for a given image out of all prompts provided by all voters (early ensembling), or to calculate the scenicness score for each voter individually before averaging the scenicness scores of all voters (late ensembling). Our results indicate that the latter method provides image scenicness ratings that are more in agreement with ScenicOrNot ratings (Kendall’s τ of

0.475). We also demonstrated that the scenicness ranking of land cover classes of our LPE method is highly similar to the ranking observed for the ScenicOrNot ratings.

Our results show that VLMs have the potential to perform accurate data-driven landscape scenicness assessments on a smaller scale than previously possible and with greater flexibility. We also demonstrated that VLMs can open up new possibilities for landscape scenicness assessments beyond rated image datasets through LPE. We hope that our findings inspire future research experiments with VLMs on other landscape qualities, especially on regional and local scales.

Chapter 6

Synthesis

6.1 Main findings

The aim of this thesis is to address current shortcomings in Landscape Quality (LQ) prediction systems and to build accurate machine learning approaches that can be used and trusted by domain specialists, practitioners, and policymakers. The chapters of this thesis have addressed various lacking aspects of current DL models for LQ assessments. The results will be assessed from the perspective of the research questions formulated in Section 1.4. Each research question is discussed below, summarising the main findings and lessons learned.

6.1.1 Which patterns can be modelled and reproduced through DL-based LQ assessments?

Understanding the types of problems that DL-based LQ assessments may be used for will help to narrow down their potential application domains. This thesis has analysed several scenarios and perspectives through which LQ assessments can be performed. Chapter 2 considered the prediction of liveability from aerial overhead images. It also investigated contributing factors to liveability through its dimension scores. These scores covered dimensions that are visible in RS images (physical environment, building quality) as well as non-visible dimensions (population, safety, amenities). Results suggested that visible dimensions are easier to predict than non-visible ones. This chapter also studied the link between neighbourhood typologies and liveability and concluded that predictions were equally accurate across neighbourhood typologies. Regarding natural environments, Chapter 3 considered the possibility of predicting scenicness at Sentinel-2 resolution, where the scores of all ground-level natural images that fall within a given Sentinel-2 patch are aggregated. The aggregated landscape scenicness scores could be predicted with high accuracy, which demonstrated that it is possible to predict perceived LQs from lower-resolution satellite images. It also investigated the relation between land cover and

scenicness through a model that explicitly estimated how each land cover class contributes to the scenicness score prediction. The resulting model was found to be slightly better than an unconstrained one. These results proved that intermediate determinants of LQs can be predicted as part of the modelling process using intrinsically interpretable methods (also see RQ. 2 below).

6.1.2 How can LQ assessment workflows using DL be made more interpretable so that it is easier to acquire new knowledge?

A prevailing problem that is inherent to DL models is the difficulty in understanding the reasoning of the model, also known as the *black box* problem. This thesis considers several approaches and methods to address this shortcoming. Firstly, in Chapter 2, a semantic bottleneck was used to predict liveability from aerial images. The intermediate concepts used were domain scores, which were designed to be linearly related to urban liveability. Experiments proved that concept bottleneck models are a useful interpretability method for RS images and that intermediate concepts can be used to explore their relation to LQs. Chapter 3 extended bottleneck models to relax the 1-on-1 relations between intermediate concepts and the final score while remaining interpretable by design. It used land cover classes as intermediate concepts to be predicted from Sentinel-2 images, but instead of a single weight per concept, it allowed for three different weights. Using an attention mechanism, the model was regularised to use predominantly a single one among these weights. The resulting model was able to distinguish between positive-and negatively-contributing examples of the same land cover class and even hint at which fine-grained land cover classes are likely contributors to each option. These experiments proved that bottleneck models are useful not only for modalities beyond natural images but also for discovering more fine-grained sub-concepts by leveraging the downstream task.

Post-hoc analysis methods were frequently used in this research to explore the representations learned by the models, in particular through the use of t-SNE dimensionality reduction. This method was used to visualise the information of the high-dimensional feature vectors that precede the LQ calculation function into a 2-dimensional representation. The analyses of Chapter 2 related liveability to neighbourhood typologies. Through t-SNE visualisation, it could be determined which neighbourhood types are not homogeneous in visual appearance and housing quality. In Chapter 3, t-SNE was used to visualise the transitions of land cover types as perceived by the decision model. By doing so, it was made visible which land cover types are perceived as similar and how interactions between land cover classes relate to scenicness. Therefore, experiments have proven that t-SNE is a useful method for post-hoc knowledge extraction from the latent representation learned by a DL model.

6.1.3 What are the benefits and challenges of multimodal DL approaches for LQ assessments?

Recent research in ML and DL has shown that multimodal methods can improve performance or even result in new ways to train and utilise models, such as the CLIP VLM discussed in Section 5.4.1. As such, this thesis has sought to discover the benefits and challenges of multimodal learning for LQ assessments. Chapter 4 considered the typical case for multimodal learning, where complementary modalities are used during training for performance benefits. Experiments were performed to predict housing quality on a regular geolocated grid by using natural images and overhead aerial images. For natural images, street photos from Google Streetview (GSV) and photos from Flickr, a photography social media platform, were used. Data fusion of aerial images with both of these data sources was attempted, as well as using each modality individually. Results suggested that using Google Streetview images alone yields the most accurate results. Using only aerial images was approximately 30% less accurate when considering Kendall's τ ranking coefficient. Using only curated natural images from Flickr was approximately 20% less accurate than using GSV images. However, a data fusion approach using Flickr images and aerial images reduced the performance deficit to 15%. The results from these experiments suggested that natural-aerial fusion can be beneficial for performance gains, in particular when high-quality data such as GSV images are not available. As GSV images are a proprietary data source, this is often the case. Flickr images, on the other hand, are publicly available, so they are more reliable.

Experiments performed in Chapter 5 studied the possibilities and challenges of a different combination of modalities, namely through the interplay between text and images. The LPE method introduced in Section 5.4.4 considered multimodality in the dataset annotation process by using text examples for landscape scenicness rather than showing image examples to annotate. In doing so, LPE allows for the annotation process to be based only on text descriptions, which are converted into image ratings using the CLIP VLM. The resulting dataset of image scenicness ratings across the entirety of the UK was found to be strongly concordant with SON (R^2 of 0.68), a well-studied dataset of image scenicness ratings. These results could be attained despite having different volunteer cohorts. The results from these experiments proved that multimodal approaches are not limited to just prediction performance and that multimodal combinations can also be leveraged to develop new annotation processes for LQ assessments. In particular, this style of multimodal annotation can offer different perspectives and characteristics that cannot be acquired through image ratings.

6.1.4 Which approaches are effective at reducing the dependence on large datasets for LQ assessments using DL models?

The dependence of DL models on large datasets is limiting their potential to be applied to new LQs. Addressing this problem can make the dataset requirements for the study of new LQs less of a burden and make it easier to repurpose existing models for new tasks. Moreover, smaller datasets mean that it is possible to study the annotators, as only a small cohort is needed for annotation. Experiments performed in Chapter 5 address this research question for natural images by leveraging web-scale pre-trained models, which have been pre-trained on extremely large datasets. Section 5.4.2 firstly considered a few-shot training setting, with sub-sets of images that are significantly smaller (3–4 orders of magnitude). The performance of the linear models trained in this setting compared to a baseline model trained on the full dataset confirmed that accurate scenicness assessments are possible with as few as 25 samples. Going beyond the requirement of labelled examples, the contrastive prompting method introduced in Section 5.4.3 uses the difference between synonyms and antonyms to regress a scenicness score without any further model training. However, this setting was found to be highly sensitive to the wording of the contrastive text descriptions. In advance, it is also not possible to know which text formulations will perform well. Yet, when using 25 labelled calibration samples to find good prompt combinations, this method could perform as well as the 25-sample few-shot case without needing parameter tuning. As such, zero-shot learning with this approach could potentially yield good results, though this needs to be confirmed by further research. In summary, the experiments performed in Chapter 5 proved that web-scale pre-trained models can be re-purposed by using only a fraction of the samples that were previously required to attain good performance, as well as that LQ assessments with small-scale training datasets are a realistic objective. It should be noted that the models that enabled these breakthroughs have notable drawbacks, which are discussed in more detail in Section 6.3.

VLMs offer more benefits than just performance, in particular for annotation purposes. The LPE annotation method proposed in Section 5.3.2 presents an interactive approach to annotation for small datasets. It uses written landscape descriptions and ratings provided by volunteers, which can be ensembled through a VLM to provide image ratings based on how well each description fits a given image. As such, it relies on the quality and diversity of text descriptions to annotate. Through this approach, it is possible to rate landscapes without needing prototype image examples, which reflects a recent trend in computer vision methodologies (Sariyildiz et al., 2023). With these characteristics, LPE is an alternative annotation method with different properties than traditional image annotation pipelines, which can be used to test other hypotheses about landscape perceptions than purely the content of images. The results acquired by LPE are proof that multimodal methods can be used not just to improve performance but also to enhance other steps of the supervised learning pipeline.

6.2 Research outlook

This thesis has demonstrated that DL methods for LQ assessments can be improved to obtain deeper insights with fewer drawbacks. Through emerging developments such as VLMs, it will soon be possible to expand the scope of LQ assessments even further. Three particular research objectives can be identified as salient topics for future research. Firstly, the role of **biases and confounders** is infrequently studied in existing LQ research using DL models but will be increasingly important for future research. Secondly, recent advances in multimodal models may enable **human-machine interaction** (HMI) research, which will yield insights that conventional LQ assessments may not. Lastly, as DL-driven LQ assessments mature, the spatial, temporal, and cultural **generalisation** of LQs will increasingly become a topic of study, such that LQ assessment models can be operationalised to track the development of landscapes. This section discusses the potential of these emergent research topics.

6.2.1 Biases and confounders

When working with subjective LQ assessment datasets, there are invariably biases and confounders that can skew ratings and results. Biases affect all aspects of the annotation pipeline. When considering the content of images, it can be observed that ephemeral attributes affect how images are perceived. For instance, seasonal variables such as lushness affect scenicness in varying ways (Gong et al., 2015; Zhang et al., 2022; Hovee, 2023), and factors such as weather variables further affect the perceived scenicness of images (Seresinhe et al., 2019; Hovee, 2023). For that reason, seasonality is frequently used as an indicator variable in study designs (Gong et al., 2015; Zhang et al., 2022). Another potential source of variations in LQ ratings for images is the quality of the photographs. Expert photographers are able to leverage photographic techniques to create more interesting or more beautiful photos, which could result in skewed LQ ratings. Although photographic qualities have been identified and quantified in existing research (Bianco et al., 2016; Hosu et al., 2020; Ke et al., 2023), photographic biases are understudied in current literature (Hovee, 2023). Examples of biases that affect LQ assessments are given in Figure 6.1.

Seasonal and photographic factors cannot be directly accounted for when using large LQ assessment image datasets, such as SON (Seresinhe et al., 2015) or Place Pulse 2 (Dubey et al., 2016). Images in such datasets are not curated for photographic qualities, seasonality, or weather conditions. Instead, they rely on collecting and annotating images in large quantities in order to create datasets that are large enough for DL models to train on. Large datasets of fixed and uniform viewpoints, such as webcam footage, can capture seasonal dynamics, which can be used to smooth out seasonal biases (Jacobs et al., 2007). However, in the absence of such a source with LQ ratings, ephemeral and photographic attributes can create skewed and biased assessments, where the model gives confident responses based on ephemeral factors. Generative models may offer a practical solution

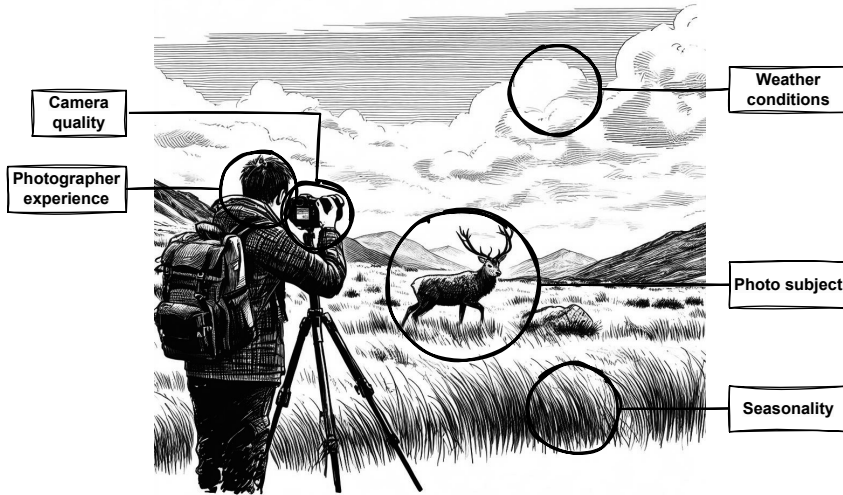


Figure 6.1: Examples of photographic and temporal biases that may affect LQ assessments. Image generated using Bing Image Creator (Microsoft, 2023).

to overcome ephemeral and photographic biases. As a possible solution, image-to-image synthesis models can be used to generate an average representation of a given image. Such methods have already been introduced for the augmentation of streetscapes (Wijnands et al., 2019; Dubey et al., 2023), so it is likely possible to synthesise corrected images with biases mitigated or removed.

6.2.2 Human-machine interaction

Human-machine interaction (HMI) has long been a fascination for computer science researchers. It is broadly defined as “*the interaction and communication between human users and a machine, a dynamic technical system, via a human-machine interface*” (Johannsen, 2009). Such interactive methods allow users to iteratively improve a model by letting users interact with predictions and outputs (Jiang et al., 2019). DL models are intrinsically not interactive, and as such, they require adaptations in order to become interactive. However, with adaptations made to the typical model training pipeline, approaches can be designed that have humans in the loop throughout the process, such as through interactive dataset generation for RS data (Hoeser and Kuenzer, 2022). Recent advances in vision-language models (VLM) and web-scale datasets for pre-training (i.e., datasets with hundreds of millions of samples gathered from the internet) have inspired many HMI initiatives, which have resulted in great academic and societal interest. Generative models built on these

advances, such as the text chat agent *ChatGPT*¹ and the image generation agent *Dall-E* (Ramesh et al., 2021), have demonstrated that DL models can be adapted to become excellent HMI agents. As a result, many interactive generative HMI models are being introduced to leverage these advances (Cao et al., 2023). However, the downsides of such models must not be ignored, such as intrinsic biases resulting from the web-scraped data sampling strategy. These downsides are discussed in more detail in Section 6.3.

The models typically used in LQ assessments are not designed with interactivity in mind, as most often they use a typical supervised learning approach. However, the annotation method presented in Section 5.3.2 has the potential to be extended into an interactive loop. The current set-up has a one-way information flow, as the user only provides rated text descriptions which are used to create an ensemble of many such descriptions. However, the process can be made personal and interactive by only considering the text descriptions of the user. After scoring all images and rendering the map, the user can then see which areas they would like to visit, look at photo ratings, and adjust their text descriptions to better suit their expectations. The practical applications of this approach to *personalized mapping* could be used to test people’s perceptions of landscapes or to find areas that a user may like to visit. These advances, enabled by sophisticated VLM approaches, can help advance our understanding of LQs in ways that were previously unachievable. Moreover, as proven in Chapter 5, it is possible to perform this type of research with small sample sizes and with a set-up that a non-technical user can use as well. These results demonstrate that HMI experiments for LQs and other aspects of geography are within reach of current models and that it may be time to test such approaches in applied research domains.

A natural extension of datasets generated by the LPE method is to study how well prompts imagined by volunteers match the calculated image ratings. By extending it with a HMI evaluation loop, it becomes possible to study the mind’s eye perceptions and inherent preferences of annotators. For instance, users can be shown image examples with ratings computed with LPE. For each prompt defined by the user, they can then be shown rated images where the prompt was maximally activated. Users can provide feedback on how fitting the image ratings are for the prompts and then adjust their prompts or include new ones. In such an iterative process, users gradually explore how well their mind’s eye view of the target region matches the landscape. This could be paired with post-hoc interviews or surveys to better understand if LPE helped raters better understand their perceptions of the target region. In a similar setting, users could also be shown the spatial distribution of LPE ratings to further explore if their perceptions of certain geographical regions match with their mind’s eye view. Through such experiments, methods such as LPE may enable the study of LQ preferences in settings that were previously not feasible due to manual labour requirements.

¹<https://chat.openai.com>

6.2.3 Generalisation

This thesis and similar research have shown that LQs can be accurately predicted from a variety of image modalities using DL models, and as such, they are mature enough for LQ assessments at scale. However, current research has not considered how well models generalise across regions, to new timesteps, and between study participant groups. This aspect of *generalisation* of a model determines how easily it may be used for out-of-distribution data (Tuia et al., 2016). For instance, domain shifts may be due to spatial, temporal, or cultural differences in data and annotator beliefs. For LQ assessments, generalisation is an important topic to consider, as it studies the limits of models in real-world settings. As such, generalisation can be considered an end goal of LQ assessment research, which is considered in more detail in this section.

For LQ assessments, generalisation starts with the dataset collection process. As it involves a subjective rating, the study participants responsible for annotating the data will give assessments based on their personal preferences. Inevitably, studies assess a subset of the population, and the characteristics of the study group will affect how well the model corresponds with ratings given by other study groups. For instance, the socio-economic background of respondents is known to affect landscape perceptions (Arthur, 1977; Abelló et al., 1986). It has also been observed that there are perceptual differences between demographic groups on how emotional places are perceived to be, and that negative emotions tend to be suppressed for online assessments when compared to in-person assessments (Huang et al., 2020). Lastly, with regards to respondents, the cultural background of respondents is a complex contributor to deviations from the human average (Abelló et al., 1986). Even the meaning of words is determined by cultural contexts (Thompson et al., 2020). In aggregate, the role of the individual is evidently important to the process of annotating LQ images. Testing the cultural generalisation of DL datasets for LQ assessments is currently not an objective that has been attempted. A primary problem is the lack of reference datasets with annotator information, as the barrier for handling, retaining, and releasing this type of information is higher since it concerns privacy-sensitive information that is protected by privacy laws such as the European GDPR (European Parliament and Council of the European Union, 2016). In order to advance DL-based LQ assessments towards a rigorous science, future research should strive for legal compliance to supply participant metadata. Through the release of annotator metadata, it will be easier to make studies, datasets, and models comparable.

Spatial generalisation concerns the ability for models to perform in unseen regions (Tuia et al., 2016). For instance, after training a model on images covering England, will it still work for images of Wales? Models with good spatial transferability therefore enable them to be used for similar problems in unseen areas. In this thesis, results were exclusively acquired for the United Kingdom and the Netherlands, with no attempt at spatial generalisation to other countries. First steps into the transferability of these models to unseen regions

can be made for the scenicness models. The SONimage dataset is based on Geograph data, which is also available in Germany. As such, a small-scale study could be performed similar to the LPE experiment to test if the LPE rating ensemble for the UK has similar predictions as the prompt ensemble created for Germany. Then, by rating German images on a large scale using LPE, the transferability of the Sentinel-2-based models of Chapter 3 could be attempted. While respecting the cultural differences discussed in the previous paragraph, such generalisation studies will yield valuable insights into how well existing models can be used to predict in countries with similar cultures and geographies.

Current research almost exclusively considers the assessment of LQs for single timesteps only. While there is much work to be done on the topic of inventorying, LQs are ephemeral in nature, and there is much that can be learned from their change processes. For instance, understanding which visual landscape elements correlate with the decay of the liveability of a neighbourhood can help devise early warning systems. Yet, it is precisely a lack of understanding about the possible veracity of such patterns that can cause fierce discussions, such as the hotly debated *broken windows theory* (Wilson and Kelling, 1982). While such theories may benefit planning purposes, more observational evidence is needed to support them. However, a downside is that longitudinal research using image modalities depends on the continued availability of images. As a result, such studies could not easily be conducted due to a lack of data. This problem is largely remedied in the current era through the increased availability of data. For natural images, Google Streetview images have already enabled the monitoring of urban processes such as gentrification (Ilic et al., 2019; Huang, 2022) and safety sentiments (Naik et al., 2017). While the use of natural images for these purposes may yield accurate, repeated assessments, the amount of data that is needed to cover a single year in a single city is enormous. As a result, analyses with natural images scale poorly both temporally and spatially. As such, it is more salient to perform such analyses using RS images. As follow-up research to the liveability prediction experiments in Chapter 2, the monitoring of liveability was attempted using time series data. A model was trained on data from 2016 and subsequently used to predict the liveability of 2012 and 2020. However, the results proved inconclusive, as the model was not able to pick up significant change patterns (Levering et al., 2023). This raises the question of over which time periods subtle changes manifest, and as a result, the timespan over which repeated monitoring using image modalities is a feasible objective. The model used for monitoring should also be adapted for the task, such as through domain adaptation methods. Additionally, having more sources of reference data may help to study the conditions necessary for LQ monitoring. Since the data collection process involves multiple time steps, it will involve long-term study set-ups and the long-term availability of images. At present, monitoring studies may be used to collect data during neighbourhood or landscape renovation processes. Images taken before and after present a redesigned view of the neighbourhood with immediately-visible changes. Before-and-after

views can be found on data sources such as Google Streetview, which may be a good starting point.

6.3 The role of web-scale pre-training for LQ assessments

Models that use web-scale pre-training using datasets gathered at the scale of the internet have been prominently discussed throughout this synthesis. They may be used to perform LQ assessments with less data (Chapter 5), re-imagine the annotation process (Section 5.3.2), generate de-biased images (Section 6.2.1), and understand LQs through HMI (Section 6.2.2). Evidently, web-scale and multimodal models such as CLIP may enable further breakthroughs for LQ assessments. Of particular interest is their potential to perform assessments without the need for re-training. The reduced technical know-how required to operate DL models will in turn enable more researchers to use them, thus making visual LQ assessments more broadly accessible for researchers and the public alike.

While web-scale datasets enable the training of powerful models, they are not without their fair share of criticisms and shortcomings, which have the potential to cause substantial societal harm. For instance, web scraping is commonly used to create web-scale datasets. However, the scraping methods and the datasets for seminal papers such as CLIP (Radford et al., 2021) are not publicly known. Publicly available datasets may use Wikipedia (Srinivasan et al., 2021) or web crawling repositories, resulting in datasets that are filtered but lack curation (Schuhmann et al., 2022). As discussed in Section 6.2.1, a lack of understanding about training datasets will introduce biases. For instance, countries may be overrepresented or underrepresented in web-scraped datasets by virtue of having a greater presence on the internet. Moreover, the copyright of data is also frequently not considered when constructing datasets. For instance, descriptive landscape photographs may have been used without the permission of photographers, and social media users may not want their opinions to be used to train models. Such issues call into question whether or not web-scale pre-trained models can be relied upon for research efforts.

As it stands, web-scale pre-trained models can provide many opportunities to bridge the gap between DL methods and applied research. However, their evident shortcomings call their reliability into question. Given both the benefits and the drawbacks of web-scale datasets, difficult decisions will have to be made. How can the performance and potential of web-scale datasets continue to contribute to the democratisation of DL models while mitigating societal harm? The outcome of this discussion will determine the role that web-scale pre-trained models can play in future LQ assessment research efforts.

References

- Abelló, R. P., F. G. Bernáldez, and E. F. Galiano (1986). “Consensus and contrast components in landscape preference”. *Environment and Behavior* 18.2. Place: US Publisher: Sage Publications, 155–178. DOI: 10.1177/0013916586182001.
- Alzubaidi, L., J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan (2021). “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions”. *Journal of Big Data* 8.1, 53. DOI: 10.1186/s40537-021-00444-8.
- Amir, S. and E. Gidalizon (1990). “Expert-based method for the evaluation of visual absorption capacity of the landscape”. *Journal of Environmental Management* 30.3, 251–263. DOI: 10.1016/0301-4797(90)90005-H.
- Arendsen, P., D. Marcos, and D. Tuia (2020). “Concept Discovery for The Interpretation of Landscape Scenicness”. *Machine Learning and Knowledge Extraction* 2.4. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, 397–413. DOI: 10.3390/make2040022.
- Arribas-Bel, D., J. E. Patino, and J. C. Duque (2017). “Remote sensing-based measurement of Living Environment Deprivation: Improving classical approaches with machine learning”. *PLOS ONE* 12.5. Publisher: Public Library of Science, e0176684. DOI: 10.1371/journal.pone.0176684.
- Arthur, L. M. (1977). “Predicting Scenic Beauty of Forest Environments: Some Empirical Tests”. *Forest Science* 23.2, 151–160. DOI: 10.1093/forestscience/23.2.151.
- Audebert, N., B. Le Saux, and S. Lefevre (2019). “Deep Learning for Classification of Hyperspectral Data: A Comparative Review”. *IEEE Geoscience and Remote Sensing Magazine* 7.2. Conference Name: IEEE Geoscience and Remote Sensing Magazine, 159–173. DOI: 10.1109/MGRS.2019.2912563.
- Baggerman, K. (2020). *Migratieachtergrond? Volgens de Leefbaarometer maak jij je wijk dan slechter*. Publication Title: Stadszaken.nl.
- Baltrusaitis, T., C. Ahuja, and L.-P. Morency (2019). “Multimodal Machine Learning: A Survey and Taxonomy”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2, 423–443. DOI: 10.1109/TPAMI.2018.2798607.

- Barber, S., D. A. Hickson, I. Kawachi, S. V. Subramanian, and F. Earls (2016). “Neighborhood Disadvantage and Cumulative Biological Risk Among a Socioeconomically Diverse Sample of African American Adults: An Examination in the Jackson Heart Study”. *Journal of Racial and Ethnic Health Disparities* 3.3, 444–456. DOI: 10.1007/s40615-015-0157-0.
- Bechtel, B., P. Alexander, J. Böhner, J. Ching, O. Conrad, J. Feddema, G. Mills, L. See, and I. Stewart (2015). “Mapping Local Climate Zones for a Worldwide Database of the Form and Function of Cities”. *ISPRS International Journal of Geo-Information* 4, 199–219. DOI: 10.3390/ijgi4010199.
- Bency, A. J., S. Rallapalli, R. K. Ganti, M. Srivatsa, and B. S. Manjunath (2017). “Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery”. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 320–329. DOI: 10.1109/WACV.2017.42.
- Berman, M. G., J. Jonides, and S. Kaplan (2008). “The Cognitive Benefits of Interacting With Nature”. *Psychological Science* 19.12. Publisher: SAGE Publications Inc, 1207–1212. DOI: 10.1111/j.1467-9280.2008.02225.x.
- Bianco, S., L. Celona, P. Napoletano, and R. Schettini (2016). “Predicting Image Aesthetics with Deep Learning”. In: *Advanced Concepts for Intelligent Vision Systems*. Ed. by J. Blanc-Talon, C. Distanto, W. Philips, D. Popescu, and P. Scheunders. Lecture Notes in Computer Science. Cham: Springer International Publishing, 117–125. DOI: 10.1007/978-3-319-48680-2_11.
- Biau, G. and E. Scornet (2016). “A random forest guided tour”. *TEST* 25.2, 197–227. DOI: 10.1007/s11749-016-0481-7.
- Biljecki, F. and K. Ito (2021). “Street view imagery in urban analytics and GIS: A review”. *Landscape and Urban Planning* 215, 104217. DOI: 10.1016/j.landurbplan.2021.104217.
- Biran, O. and C. V. Cotton (2017). “Explanation and justification in machine learning: A survey”. In: *IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8, 8–13.
- Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, et al. (2022). *On the Opportunities and Risks of Foundation Models*. DOI: 10.48550/arXiv.2108.07258. arXiv: 2108.07258[cs].
- Breiman, L. (2001). “Random Forests”. *Machine Learning* 45.1, 5–32. DOI: 10.1023/A:1010933404324.
- Bubalo, M., B. T. van Zanten, and P. H. Verburg (2019). “Crowdsourcing geo-information on landscape perceptions and preferences: A review”. *Landscape and Urban Planning* 184, 101–111. DOI: 10.1016/j.landurbplan.2019.01.001.

- Buhyoff, G. J. and M. F. Riesenman (1979). “Manipulation of dimensionality in landscape preference judgments: A quantitative validation”. *Leisure Sciences* 2.3. Publisher: Routledge eprint: <https://doi.org/10.1080/01490407909512917>, 221–238. DOI: 10.1080/01490407909512917.
- Campos-Taberner, M., A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. Le Saux, A. Beaupère, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia (2016). “Processing of Extremely High-Resolution LiDAR and RGB Data: Outcome of the 2015 IEEE GRSS Data Fusion Contest—Part A: 2-D Contest”. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9.12. Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5547–5559. DOI: 10.1109/JSTARS.2016.2569162.
- Camps-Valls, G., D. Tuia, X. X. Zhu, and M. Reichstein, eds. (2021). *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*. Wiley. 432 pp. DOI: 10.1002/9781119646181.
- Cao, Y., S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun (2023). “A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT”. Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2303.04226.
- Caruana, R. (1997). “Multitask Learning”. *Machine Learning* 28.1, 41–75. DOI: 10.1023/A:1007379606734.
- CBS (2016). *Inkomen per gemeente en wijk, 2016*. Publication Title: Centraal Bureau voor de Statistiek Type: webpagina.
- Chappuis, C., V. Mendez, E. Walt, S. Lobry, B. Le Saux, and D. Tuia (2022). “Language Transformers for Remote Sensing Visual Question Answering”. In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium. ISSN: 2153-7003, 4855–4858. DOI: 10.1109/IGARSS46834.2022.9884036.
- Chen, S. and F. Biljecki (2023). “Automatic assessment of public open spaces using street view imagery”. *Cities* 137, 104329. DOI: 10.1016/j.cities.2023.104329.
- Christman, Z. J., M. Wilson-Genderson, A. Heid, and R. Pruchno (2020). “The Effects of Neighborhood Built Environment on Walking for Leisure and for Purpose Among Older People”. *The Gerontologist* 60.4, 651–660. DOI: 10.1093/geront/gnz093.
- Copernicus Project (2023). *WMS browser*.
- Daniel, T. C. (2001). “Whither scenic beauty? Visual landscape quality assessment in the 21st century”. *Landscape and Urban Planning*. Our Visual Landscape: analysis, modeling, visualization and protection 54.1, 267–281. DOI: 10.1016/S0169-2046(01)00141-4.
- (1976). *Measuring landscape esthetics: The scenic beauty estimation method*. Rocky Mountain Forest and Range Experiment Station.

- Daniel, T. C., A. Muhar, A. Arnberger, O. Aznar, J. W. Boyd, K. M. A. Chan, R. Costanza, T. Elmqvist, C. G. Flint, P. H. Gobster, A. Gret-Regamey, R. Lave, S. Muhar, M. Penker, R. G. Ribe, T. Schauppenlehner, T. Sikor, I. Soloviy, M. Spierenburg, K. Taczanowska, J. Tam, and A. v. d. Dunk (2012). “Contributions of cultural services to the ecosystem services agenda”. *Proceedings of the National Academy of Sciences*. 109(23): 8812-8819. 109.23. Number: 23, 8812–8819. DOI: 10.1073/pnas.1114773109.
- Daniel, T. C. and J. Vining (1983). “Methodological issues in the assessment of landscape quality”. *Human Behavior & Environment: Advances in Theory & Research* 6. Place: US Publisher: Kluwer Academic/Plenum Publishers, 39–84.
- Demir, I., K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar (2018). “DeepGlobe 2018: A challenge to parse the earth through satellite images”.
- Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition. ISSN: 1063-6919, 248–255. DOI: 10.1109/CVPR.2009.5206848.
- Dubey, A., N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo (2016). “Deep Learning the City: Quantifying Urban Perception at a Global Scale”. In: *ECCV 2016*. Ed. by B. Leibe, J. Matas, N. Sebe, and M. Welling. Cham: Springer, 196–212. DOI: 10.1007/978-3-319-46448-0_12.
- Dubey, R., M. Hardy, T. Griffiths, and R. Bhui (2023). *AI-generated visuals of car-free American cities help increase support for sustainable transport policies*. DOI: 10.31234/osf.io/g9ptz.
- EU Copernicus Program (2018). *CLC 2018 — Copernicus Land Monitoring Service*. (Visited on 2019).
- European Parliament and Council of the European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*. Legislative Body: EP, CONSIL.
- Evans, G. W. (2003). “The built environment and mental health”. *Journal of Urban Health: Bulletin of the New York Academy of Medicine* 80.4, 536–555. DOI: 10.1093/jurban/jtg063.
- Fan, R., J. Li, W. Song, W. Han, J. Yan, and L. Wang (2022). “Urban informal settlements classification via a transformer-based spatial-temporal fusion network using multimodal remote sensing and time-series human activity data”. *International Journal of Applied Earth Observation and Geoinformation* 111, 102831. DOI: 10.1016/j.jag.2022.102831.
- Fu, K., W. Dai, Y. Zhang, Z. Wang, M. Yan, and X. Sun (2019). “MultiCAM: Multiple Class Activation Mapping for Aircraft Recognition in Remote Sensing Images”. *Remote*

- Sensing* 11.5. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, 544. DOI: 10.3390/rs11050544.
- Gabeff, V., M. Russwurm, A. Mathis, and D. Tuia (2023). “Scene and animal attributes retrieval from camera trap data with domain-adapted language-vision models”. In: *Computer Vision and Pattern Recognition Workshop cv4animals*.
- Galindo, M. P. G. and J. A. C. Rodriguez (2000). “Environmental aesthetics and psychological wellbeing: Relationships between preference judgements for urban landscapes and other relevant affective responses”. *Psychology in Spain* 4. Place: Spain Publisher: Colegio Oficial de Psicólogos, 13–27.
- Gebru, T., J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford (2021). “Datasheets for datasets”. *Communications of the ACM* 64.12, 86–92. DOI: 10.1145/3458723.
- Gevaert, C. M. (2022). “Explainable AI for earth observation: A review including societal and regulatory perspectives”. *International Journal of Applied Earth Observation and Geoinformation* 112, 102869. DOI: 10.1016/j.jag.2022.102869.
- Gómez-Chova, L., D. Tuia, G. Moser, and G. Camps-Valls (2015). “Multimodal Classification of Remote Sensing Images: A Review and Future Directions”. *Proceedings of the IEEE* 103.9. Conference Name: Proceedings of the IEEE, 1560–1584. DOI: 10.1109/JPROC.2015.2449668.
- Gong, L., Z. Zhang, and C. Xu (2015). “Developing a Quality Assessment Index System for Scenic Forest Management: A Case Study from Xishan Mountain, Suburban Beijing”. *Forests* 6.1. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, 225–243. DOI: 10.3390/f6010225.
- Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. Cambridge, Massachusetts. 800 pp.
- Grinde, B. and G. G. Patil (2009). “Biophilia: Does Visual Contact with Nature Impact on Health and Well-Being?” *International Journal of Environmental Research and Public Health* 6.9. Number: 9 Publisher: Molecular Diversity Preservation International, 2332–2343. DOI: 10.3390/ijerph6092332.
- Haan, M., G. A. Kaplan, and T. Camacho (1987). “Poverty and health. Prospective evidence from the Alameda County Study”. *American Journal of Epidemiology* 125.6, 989–998. DOI: 10.1093/oxfordjournals.aje.a114637.
- Hall, C., A. McVittie, and D. Moran (2004). “What does the public want from agriculture and the countryside? A review of evidence and methods”. *Journal of Rural Studies* 20.2, 211–225. DOI: 10.1016/j.jrurstud.2003.08.004.
- Halling, P. (2011). *Millenium Dome*. Geograph. (Visited on 2023).

- Havinga, I., P. W. Bogaart, L. Hein, and D. Tuia (2020). “Defining and spatially modelling cultural ecosystem services using crowdsourced data”. *Ecosystem Services* 43. Publisher: Elsevier, 101091.
- Havinga, I., D. Marcos, P. W. Bogaart, L. Hein, and D. Tuia (2021). “Social media and deep learning capture the aesthetic quality of the landscape”. *Scientific reports* 11.1. Publisher: Nature Publishing Group, 1–11.
- He, K., X. Zhang, S. Ren, and J. Sun (2016). “Deep Residual Learning for Image Recognition”. In: *CVPR*. CVPR, 770–778.
- Hentschel, S., K. Kobs, and A. Hotho (2022). “CLIP knows image aesthetics”. *Frontiers in Artificial Intelligence* 5.
- Hill, D. and T. C. Daniel (2007). “Foundations for an Ecological Aesthetic: Can Information Alter Landscape Preferences?” *Society & Natural Resources* 21.1. Publisher: Routledge .eprint: <https://doi.org/10.1080/08941920701655700>, 34–49. DOI: 10.1080/08941920701655700.
- Hodgson, R. W. and R. L. Thayer (1980). “Implied human influence reduces landscape beauty”. *Landscape Planning* 7.2, 171–179. DOI: 10.1016/0304-3924(80)90014-3.
- Hoeser, T. and C. Kuenzer (2022). “SyntEO: Synthetic dataset generation for earth observation and deep learning – Demonstrated for offshore wind farm detection”. *ISPRS Journal of Photogrammetry and Remote Sensing* 189, 163–184. DOI: 10.1016/j.isprsjprs.2022.04.029.
- Hoeve, K. van (2023). “Unraveling Landscape Scenicness with Deep Learning”. Masters thesis. Wageningen: Wageningen University.
- Hosu, V., H. Lin, T. Sziranyi, and D. Saupe (2020). “KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment”. *IEEE Transactions on Image Processing* 29, 4041–4056. DOI: 10.1109/TIP.2020.2967829.
- Huang, T. (2022). “Detecting Neighborhood Gentrification at Scale via Street Views and POIs (Student Abstract)”. *Proceedings of the AAAI Conference on Artificial Intelligence* 36.11. Number: 11, 12969–12970. DOI: 10.1609/aaai.v36i11.21621.
- Huang, X. and Y. Liu (2022). “Livability assessment of 101,630 communities in China’s major cities: A remote sensing perspective”. *Science China Earth Sciences* 65.6, 1073–1087. DOI: 10.1007/s11430-021-9896-4.
- Huang, Y., J. Li, G. Wu, and T. Fei (2020). “Quantifying the bias in place emotion extracted from photos on social networking sites: A case study on a university campus”. *Cities* 102, 102719. DOI: 10.1016/j.cities.2020.102719.
- Hull, B. R. and G. R. B. Revell (1989). “Issues in sampling landscapes for visual quality assessments”. *Landscape and Urban Planning* 17.4, 323–330. DOI: 10.1016/0169-2046(89)90086-8.

- Huysmans, J., K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens (2011). “An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models”. *Decision Support Systems* 51.1, 141–154. DOI: 10.1016/j.dss.2010.12.003.
- Ilic, L., M. Sawada, and A. Zarzelli (2019). “Deep mapping gentrification in a large Canadian city using deep learning and Google Street View”. *PLOS ONE* 14.3. Publisher: Public Library of Science, e0212814. DOI: 10.1371/journal.pone.0212814.
- Imamoglu, N., M. Kimura, H. Miyamoto, A. Fujita, and R. Nakamura (2017). “Solar Power Plant Detection on Multi-Spectral Satellite Imagery using Weakly-Supervised CNN with Feedback Features and m-PCNN Fusion”. *arXiv:1704.06410 [cs]*. arXiv: 1704.06410.
- Jacobs, N., N. Roman, and R. Pless (2007). “Consistent Temporal Variations in Many Outdoor Scenes”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007 IEEE Conference on Computer Vision and Pattern Recognition. ISSN: 1063-6919, 1–6. DOI: 10.1109/CVPR.2007.383258.
- Jacques, D. (1980). “Landscape appraisal: the case for a subjective theory.” *Journal of Environmental Management* 10, 107–113.
- Jensen, R., J. Gatrell, J. Boulton, and B. Harper (2004). “Using Remote Sensing and Geographic Information Systems to Study Urban Quality of Life and Urban Forest Amenities”. *Ecology and Society* 9.5. Publisher: Resilience Alliance Inc.
- Jiang, L., S. Liu, and C. Chen (2019). “Recent research advances on interactive machine learning”. *Journal of Visualization* 22.2, 401–417. DOI: 10.1007/s12650-018-0531-1.
- Johannsen, G. (2009). “Human-machine interaction”. *Control Systems, Robotics, and Automation* 21, 132–62.
- Jordan, M. I. and T. M. Mitchell (2015). “Machine learning: Trends, perspectives, and prospects”. *Science* 349.6245. Place: US Publisher: American Assn for the Advancement of Science, 255–260. DOI: 10.1126/science.aaa8415.
- Kamp, I. van, K. Leidelmeijer, G. Marsman, and A. de Hollander (2003). “Urban environmental quality and human well-being: Towards a conceptual framework and demarcation of concepts; a literature study”. *Landscape and Urban Planning*. Urban environmental quality and human wellbeing 65.1, 5–18. DOI: 10.1016/S0169-2046(02)00232-3.
- Ke, J., K. Ye, J. Yu, Y. Wu, P. Milanfar, and F. Yang (2023). “VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining”. In: *Proceeding of the CVR 2023*. CVPR 2023. Vancouver: arXiv. DOI: 10.48550/arXiv.2303.14302. arXiv: 2303.14302[cs].
- Kendall, M. (1938). “A New Measure for Rank Correlation”. *Biometrika* 30.1, 81–93. DOI: 10.1093/biomet/30.1-2.81.
- Khurana, D., A. Koli, K. Khatter, and S. Singh (2023). “Natural language processing: state of the art, current trends and challenges”. *Multimedia Tools and Applications* 82.3, 3713–3744. DOI: 10.1007/s11042-022-13428-4.

- Kim, B., M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres (2018). “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)”. In: *Proceedings of the 35th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, 2668–2677.
- Kingma, D. and J. Ba (2014). “Adam: A Method for Stochastic Optimization”. In: *Proceedings of the 3rd International Conference on Learning Representations*. ICLR. San Diego: Ithaca, 13.
- Kleerekoper, L. (2016). “Urban Climate Design: Improving thermal comfort in Dutch neighbourhoods”. *A+BE: Architecture and the Built Environment* 6. DOI: 10.7480/abe.2016.11.
- Kleerekoper, L., A. Koekoek, and J. Kluck (2018). “Een wijktypologie voor klimaatadaptatie”. *Stadswerk Magazine*, 28–30.
- Koh, P. W., T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang (2020). “Concept Bottleneck Models”. In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, 5338–5348.
- Krippendorf, J. (1984). *Die Ferienmenschen Für ein neues Verständnis von Freizeit und Reisen*. Orell Füssli.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc.
- Kuffer, M., D. R. Thomson, G. Boo, R. Mahabir, T. Grippa, S. Vanhuyse, R. Engstrom, R. Ndugwa, J. Makau, E. Darin, J. P. de Albuquerque, and C. Kabaria (2020). “The Role of Earth Observation in an Integrated Deprived Area Mapping “System” for Low-to-Middle Income Countries”. *Remote Sensing* 12.6. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, 982. DOI: 10.3390/rs12060982.
- Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller (2019). “Unmasking Clever Hans predictors and assessing what machines really learn”. *Nature Communications* 10.1. Number: 1 Publisher: Nature Publishing Group, 1096. DOI: 10.1038/s41467-019-08987-4.
- Law, S., C. I. Seresinhe, Y. Shen, and M. Gutierrez-Roig (2018). “Street-Frontage-Net: urban image classification using deep convolutional neural networks”. *International Journal of Geographical Information Science*, 1–27. DOI: 10.1080/13658816.2018.1555832.
- Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). “Gradient-based learning applied to document recognition”. *Proceedings of the IEEE* 86.11. Conference Name: Proceedings of the IEEE, 2278–2324. DOI: 10.1109/5.726791.

- Leidelmeijer, K. and J. Mandemakers (2020). *Leefbaarheid in Nederland 2020*. Amsterdam: Atlas Research, 79.
- Leidelmeijer, K., G. Marlet, R. Ponds, R. Schulenberg, and C. van Woerkens (2014). *Leefbaarometer 2.0: Instrumentenontwikkeling. 2*. Amsterdam: RIGO Research en Advies/Atlas voor Gemeenten, 151.
- Levering, A., D. Marcos, S. Lobry, and D. Tuia (2020). “Interpretable Scenicness from Sentinel-2 Imagery”. In: *Proceedings of the 2020 International Geoscience and Remote Sensing Symposium*. International Geoscience and Remote Sensing Symposium. Hawaii, 4.
- Levering, A., D. Marcos, and D. Tuia (2021a). “Liveability from Above: Understanding Quality of Life with Overhead Imagery and Deep Neural Networks”. In: *Proceedings of IGARSS 2021*. IGARSS. Brussels: IEEE.
- (2021b). “On the relation between landscape beauty and land cover: A case study in the U.K. at Sentinel-2 resolution with interpretable AI”. *ISPRS Journal of Photogrammetry and Remote Sensing* 177, 194–203. DOI: 10.1016/j.isprsjprs.2021.04.020.
- (2023). “Time Series Analysis of Urban Liveability”. In: *2023 Joint Urban Remote Sensing Event (JURSE)*. 2023 Joint Urban Remote Sensing Event (JURSE). ISSN: 2642-9535, 1–4. DOI: 10.1109/JURSE57346.2023.10144221.
- Li, G. and Q. Weng (2007). “Measuring the quality of life in city of Indianapolis by integration of remote sensing and census data”. *International Journal of Remote Sensing* 28.2. Publisher: Taylor & Francis eprint: <https://doi.org/10.1080/01431160600735624>, 249–267. DOI: 10.1080/01431160600735624.
- Li, W., X. Huang, Z. Zhu, Y. Tang, X. Li, J. Zhou, and J. Lu (2022). “OrdinalCLIP: Learning Rank Prompts for Language-Guided Ordinal Regression”. In:
- Lindemann-Matthies, P., R. Briegel, B. Schüpbach, and X. Junge (2010). “Aesthetic preference for a Swiss alpine landscape: The impact of different agricultural land use with different biodiversity”. *Landscape and Urban Planning* 98.2, 99–109. DOI: 10.1016/j.landurbplan.2010.07.015.
- Linnainmaa, S. (1976). “Taylor expansion of the accumulated rounding error”. *BIT Numerical Mathematics* 16.2, 146–160. DOI: 10.1007/BF01931367.
- Liu, S. and Q. Shi (2020). “Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China”. *ISPRS Journal of Photogrammetry and Remote Sensing* 164, 229–242. DOI: 10.1016/j.isprsjprs.2020.04.008.
- Liu, Z., H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie (2022). “A ConvNet for the 2020s”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). ISSN: 2575-7075, 11966–11976. DOI: 10.1109/CVPR52688.2022.01167.

- Lobry, S., B. Demir, and D. Tuia (2021). “RSVQA Meets Bigearthnet: A New, Large-Scale, Visual Question Answering Dataset for Remote Sensing”. In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. ISSN: 2153-7003, 1218–1221. DOI: 10.1109/IGARSS47720.2021.9553307.
- Loshchilov, I. and F. Hutter (2019). “Decoupled Weight Decay Regularization”. In: *ICLR*.
- Maaten, L. v. d. and G. Hinton (2008). “Visualizing Data using t-SNE”. *Journal of Machine Learning Research* 9.86, 2579–2605.
- Maggiori, E., Y. Tarabalka, G. Charpiat, and P. Alliez (2017). “High-Resolution Aerial Image Labeling With Convolutional Neural Networks”. *IEEE Transactions on Geoscience and Remote Sensing* 55.12. Conference Name: IEEE Transactions on Geoscience and Remote Sensing, 7092–7103. DOI: 10.1109/TGRS.2017.2740362.
- Marcos, D., S. Lobry, and D. Tuia (2019). “Semantically Interpretable Activation Maps: what-where-how explanations within CNNs”. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). ISSN: 2473-9944, 4207–4215. DOI: 10.1109/ICCVW.2019.00518.
- Marcos, D., R. Fong, S. Lobry, R. Flamary, N. Courty, and D. Tuia (2021). “Contextual Semantic Interpretability”. In: *Computer Vision – ACCV 2020*. Ed. by H. Ishikawa, C.-L. Liu, T. Pajdla, and J. Shi. Lecture Notes in Computer Science. Cham: Springer International Publishing, 351–368. DOI: 10.1007/978-3-030-69538-5_22.
- Marmanis, D., K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla (2018). “Classification with an edge: Improving semantic image segmentation with boundary detection”. *ISPRS Journal of Photogrammetry and Remote Sensing* 135, 158–172. DOI: 10.1016/j.isprsjprs.2017.11.009.
- Microsoft (2023). *Bing Image Creator*. Microsoft Image Creator.
- Miller, T. (2019). “Explanation in artificial intelligence: Insights from the social sciences”. *Artificial Intelligence* 267, 1–38. DOI: 10.1016/j.artint.2018.07.007.
- Ministry of Infrastructure and the Environment (2012). *35 icons of Dutch spatial planning*. The Hague: Ministry of Infrastructure and the Environment.
- Munoz, J. E. V., S. Srivastava, D. Tuia, and A. X. Falcao (2021). “OpenStreetMap: Challenges and Opportunities in Machine Learning and Remote Sensing”. *IEEE Geoscience and Remote Sensing Magazine* 9.1. Publisher: IEEE, 184–199. DOI: 10.1109/MGRS.2020.2994107.
- Nahuelhual, L., P. Laterra, D. Jiménez, A. Báez, C. Echeverría, R. Fuentes, L. Nahuelhual, P. Laterra, D. Jiménez, A. Báez, C. Echeverría, and R. Fuentes (2018). “Do people prefer natural landscapes? An empirical study in Chile”. *Bosque (Valdivia)* 39.2. Publisher: Universidad Austral de Chile, 205–216. DOI: 10.4067/S0717-92002018000200205.

- Naik, N., S. D. Kominers, R. Raskar, E. L. Glaeser, and C. A. Hidalgo (2017). “Computer vision uncovers predictors of physical urban change”. *Proceedings of the National Academy of Sciences* 114.29, 7571–7576. DOI: 10.1073/pnas.1619003114.
- Naik, N., J. Philipoom, R. Raskar, and C. Hidalgo (2014). “Streetscore – Predicting the Perceived Safety of One Million Streetscapes”. In: *CVPR 2014*. CVPR 2014. Columbus, OH, USA: IEEE, 793–799. DOI: 10.1109/CVPRW.2014.121.
- Nguyen, T.-A., B. Kellenberger, and D. Tuia (2022). “Mapping forest in the Swiss Alps treeline ecotone with explainable deep learning”. *Remote Sensing of Environment* 281. Publisher: Elsevier, 113217.
- Palmer, J. F. (2004). “Using spatial metrics to predict scenic perception in a changing landscape: Dennis, Massachusetts”. *Landscape and Urban Planning*. The Social Aspects of Landscape Change: Protecting Open Space Under the Pressure of Development 69.2, 201–218. DOI: 10.1016/j.landurbplan.2003.08.010.
- Paul, A. and J. Sen (2020). “A critical review of liveability approaches and their dimensions”. *Geoforum; Journal of Physical, Human, and Regional Geosciences* 117, 90–92. DOI: 10.1016/j.geoforum.2020.09.008.
- PDOK (2017). *NIEUW: hogere resolutie luchtfoto als open data bij PDOK*. Publication Title: NIEUW: hogere resolutie luchtfoto als open data bij PDOK - PDOK.
- Pearson, K. (1901). “LIII. On lines and planes of closest fit to systems of points in space”. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11. Publisher: Taylor & Francis eprint: <https://doi.org/10.1080/14786440109462720>, 559–572. DOI: 10.1080/14786440109462720.
- Penney, B. (2019). *English indices of deprivation 2015*. UK Office for National Statistics.
- Qiu, C., L. Mou, M. Schmitt, and X. X. Zhu (2019). “Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network”. *ISPRS Journal of Photogrammetry and Remote Sensing* 154, 151–162. DOI: 10.1016/j.isprsjprs.2019.05.004.
- Qiu, W., Z. Zhang, X. Liu, W. Li, X. Li, X. Xu, and X. Huang (2022). “Subjective or objective measures of street environment, which are more effective in explaining housing prices?” *Landscape and Urban Planning* 221, 104358. DOI: 10.1016/j.landurbplan.2022.104358.
- Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever (2021). “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, 8748–8763.
- Rahman, A., Y. Kumar, S. Fazal, and S. Bhaskaran (2011). “Urbanization and Quality of Urban Environment Using Remote Sensing and GIS Techniques in East Delhi-

- India". *Journal of Geographic Information System* 03.1. Number: 01 Publisher: Scientific Research Publishing, 62. DOI: 10.4236/jgis.2011.31005.
- Ramesh, A., M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever (2021). "Zero-Shot Text-to-Image Generation". In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, 8821–8831.
- Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat (2019). "Deep learning and process understanding for data-driven Earth system science". *Nature* 566.7743. Number: 7743 Publisher: Nature Publishing Group, 195–204. DOI: 10.1038/s41586-019-0912-1.
- Rijsbergen, C. van (1979). "Information Retrieval". *Journal of the American Society for Information Science* 30.6. _eprint: <https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.4630300621>, 374–375. DOI: <https://doi.org/10.1002/asi.4630300621>.
- Rodriguez Lopez, J. M., K. Heider, and J. Scheffran (2017). "Frontiers of urbanization: Identifying and explaining urbanization hot spots in the south of Mexico City using human and remote sensing". *Applied Geography* 79, 1–10. DOI: 10.1016/j.apgeog.2016.12.001.
- Roe, J. J., C. W. Thompson, P. A. Aspinall, M. J. Brewer, E. I. Duff, D. Miller, R. Mitchell, and A. Clow (2013). "Green Space and Stress: Evidence from Cortisol Measures in Deprived Urban Communities". *International Journal of Environmental Research and Public Health* 10.9. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, 4086–4103. DOI: 10.3390/ijerph10094086.
- Roscher, R., B. Bohn, M. F. Duarte, and J. Garcke (2020). "Explainable Machine Learning for Scientific Insights and Discoveries". *IEEE Access* 8. Conference Name: IEEE Access, 42200–42216. DOI: 10.1109/ACCESS.2020.2976199.
- Rosenblatt, F. (1958). "The perceptron: A probabilistic model for information storage and organization in the brain". *Psychological Review* 65.6. Place: US Publisher: American Psychological Association, 386–408. DOI: 10.1037/h0042519.
- Rosier, J. F., H. Taubenböck, P. H. Verburg, and J. van Vliet (2022). "Fusing Earth observation and socioeconomic data to increase the transferability of large-scale urban land use classification". *Remote Sensing of Environment* 278, 113076. DOI: 10.1016/j.rse.2022.113076.
- Samek, W. and K.-R. Müller (2019). "Towards Explainable Artificial Intelligence". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller. Lecture Notes in Computer Science. Cham: Springer International Publishing, 5–22. DOI: 10.1007/978-3-030-28954-6_1.

- Sapena, M., M. Wurm, H. Taubenböck, D. Tuia, and L. A. Ruiz (2021). “Estimating quality of life dimensions from urban spatial pattern metrics”. *Computers, Environment and Urban Systems* 85, 101549. DOI: 10.1016/j.compenvurbsys.2020.101549.
- Sariyildiz, M. B., K. Alahari, D. Larlus, and Y. Kalantidis (2023). “Fake it till you make it: Learning transferable representations from synthetic ImageNet clones”. In: *CVPR 2023-IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Toronto.
- Scepanovic, S., S. Joglekar, S. Law, and D. Quercia (2021). “Jane Jacobs in the Sky: Predicting Urban Vitality with Open Satellite Data”. *ACM Human-Computer Interaction* 5, 1–25. DOI: 10.1145/3449257.
- Schirpke, U., E. Tasser, and U. Tappeiner (2013). “Predicting scenic beauty of mountain regions”. *Landscape and Urban Planning* 111, 1–12. DOI: 10.1016/j.landurbplan.2012.11.010.
- Schroeder, H. and T. C. Daniel (1981). “Progress in Predicting the Perceived Scenic Beauty of Forest Landscapes”. *Forest Science* 27.1. Publisher: Oxford Academic, 71–80. DOI: 10.1093/forestscience/27.1.71.
- Schuhmann, C., R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev (2022). *LAION-5B: An open large-scale dataset for training next generation image-text models*. DOI: 10.48550/arXiv.2210.08402. arXiv: 2210.08402[cs].
- Seresinhe, C. I., T. Preis, G. MacKerron, and H. S. Moat (2019). “Happiness is Greater in More Scenic Locations”. *Scientific Reports* 9.1, 1–11. DOI: 10.1038/s41598-019-40854-6.
- Seresinhe, C. I., T. Preis, and H. S. Moat (2015). “Quantifying the Impact of Scenic Environments on Health”. *Scientific Reports* 5.1, 1–9. DOI: 10.1038/srep16899.
- (2017). “Using deep learning to quantify the beauty of outdoor places”. *Royal Society Open Science* 4.7. Publisher: Royal Society, 170170. DOI: 10.1098/rsos.170170.
- Shuttleworth, S. (1979). “The evaluation of landscape quality”. *Landscape Research* 5.1. Publisher: Routledge _eprint: <https://doi.org/10.1080/01426397908705925>, 14–15. DOI: 10.1080/01426397908705925.
- Singleton, A., D. Arribas-Bel, J. Murray, and M. Fleischmann (2022). “Estimating generalized measures of local neighbourhood context from multispectral satellite images using a convolutional neural network”. *Computers, Environment and Urban Systems* 95, 101802. DOI: 10.1016/j.compenvurbsys.2022.101802.
- Solecka, I. (2019). “The use of landscape value assessment in spatial planning and sustainable land management — a review”. *Landscape Research* 44.8. Publisher: Routledge _eprint: <https://doi.org/10.1080/01426397.2018.1520206>, 966–981. DOI: 10.1080/01426397.2018.1520206.

- Song, H., L. Dong, W. Zhang, T. Liu, and F. Wei (2022). “CLIP Models are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Dublin, Ireland: Association for Computational Linguistics, 6088–6100. DOI: 10.18653/v1/2022.acl-long.421.
- Srinivasan, K., K. Raman, J. Chen, M. Bendersky, and M. Najork (2021). “WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21. New York, NY, USA: Association for Computing Machinery, 2443–2449. DOI: 10.1145/3404835.3463257.
- Srivastava, S., J. E. Vargas Muñoz, S. Lobry, and D. Tuia (2020). “Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data”. *IJGIS* 34.6, 1117–1136. DOI: 10.1080/13658816.2018.1542698.
- Srivastava, S., J. E. Vargas-Muñoz, and D. Tuia (2019). “Understanding urban landuse from the above and ground perspectives: a deep learning, multimodal solution”. *Remote Sensing of Environment* 228, 129–143. DOI: 10.1016/j.rse.2019.04.014. arXiv: 1905.01752.
- Suel, E., S. Bhatt, M. Brauer, S. Flaxman, and M. Ezzati (2021). “Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas”. *Remote Sensing of Environment* 257, 112339. DOI: 10.1016/j.rse.2021.112339.
- Suel, E., J. W. Polak, J. E. Bennett, and M. Ezzati (2019). “Measuring social, environmental and health inequalities using deep learning and street imagery”. *Scientific Reports* 9.1. Number: 1 Publisher: Nature Publishing Group, 6229. DOI: 10.1038/s41598-019-42036-w.
- Sumbul, G., M. Charfuelan, B. Demir, and V. Markl (2019). “Bigearthnet A Large-Scale Benchmark Archive for Remote Sensing Image Understanding”. *IGARSS*, 5901–5904.
- Sundararajan, M., A. Taly, and Q. Yan (2017). “Axiomatic attribution for deep networks”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 3319–3328.
- Svendsen, D. H., P. Morales-Álvarez, A. B. Ruescas, R. Molina, and G. Camps-Valls (2020). “Deep Gaussian processes for biogeophysical parameter retrieval and model inversion”. *ISPRS Journal of Photogrammetry and Remote Sensing* 166, 68–81. DOI: 10.1016/j.isprsjprs.2020.04.014.
- Taubenböck, H., T. Esch, A. Felbier, M. Wiesner, A. Roth, and S. Dech (2012). “Monitoring urbanization in mega cities from space”. *Remote Sensing of Environment*. Remote Sensing of Urban Environments 117, 162–176. DOI: 10.1016/j.rse.2011.09.015.
- Teeffelen, K. v. (2021). “Een algoritme is niet neutraal, ook een overheidsalgoritme niet”. *Trouw*. Section: binnenland.

- Thompson, B., S. G. Roberts, and G. Lupyan (2020). “Cultural influences on word meanings revealed through large-scale semantic alignment”. *Nature Human Behaviour* 4.10. Number: 10 Publisher: Nature Publishing Group, 1029–1038. DOI: 10.1038/s41562-020-0924-8.
- Thompson, S. and J. Kent (2014). “Healthy Built Environments Supporting Everyday Occupations: Current Thinking in Urban Planning”. *Journal of Occupational Science* 21.1, 25–41. DOI: 10.1080/14427591.2013.867562.
- Tian, Y., N.-E. Tsendbazar, E. van Leeuwen, R. Fensholt, and M. Herold (2022). “A global analysis of multifaceted urbanization patterns using Earth Observation data from 1975 to 2015”. *Landscape and Urban Planning* 219, 104316. DOI: 10.1016/j.landurbplan.2021.104316.
- Tuia, D., G. Camps-Valls, G. Matasci, and M. Kanevski (2010). “Learning relevant image features with multiple-kernel classification”. *IEEE Transactions on Geoscience and Remote Sensing* 48.10. Publisher: Institute of Electrical and Electronics Engineers, 3780–3791. DOI: 10.1109/TGRS.2010.2049496.
- Tuia, D., B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt, R. Kays, H. Klinck, M. Wikelski, I. D. Couzin, G. van Horn, M. C. Crofoot, C. V. Stewart, and T. Berger-Wolf (2022). “Perspectives in machine learning for wildlife conservation”. *Nature Communications* 13.1. Number: 1 Publisher: Nature Publishing Group, 792. DOI: 10.1038/s41467-022-27980-y.
- Tuia, D., C. Persello, and L. Bruzzone (2016). “Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances”. *IEEE Geoscience and Remote Sensing Magazine* 4.2, 41–57. DOI: 10.1109/MGRS.2016.2548504.
- “Toward a Collective Agenda on AI for Earth Science Data Analysis” (2021). *IEEE Geoscience and Remote Sensing Magazine*. Ed. by D. Tuia, R. Roscher, J. D. Wegner, N. Jacobs, X. Zhu, and G. Camps-Valls. DOI: 10.1109/MGRS.2020.3043504.
- Uitermark, J., C. Hochstenbach, and W. van Gent (2017). “The statistical politics of exceptional territories”. *Political Geography* 57, 60–70. DOI: 10.1016/j.polgeo.2016.11.011.
- United Nations, Department of Economic and Social Affairs, Population Division (2019). *World Urbanization Prospects: The 2018 Revision*. ST/ESA/SER.A/420. New York.
- Vasu, B., F. U. Rahman, and A. Savakis (2018). “Aerial-CAM: Salient Structures and Textures in Network Class Activation Maps of Aerial Imagery”. In: *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. 2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP). Aristi Village, Zagorochoria, Greece: IEEE, 1–5. DOI: 10.1109/IVMSPW.2018.8448567.
- Veenhoven, R., J. Ehrhardt, M. S. D. Ho, and A. de Vries (1993). *Happiness in nations: Subjective appreciation of life in 56 nations 1946–1992*. Happiness in nations: Subjective

- appreciation of life in 56 nations 1946–1992. Pages: 365. Rotterdam, Netherlands: Erasmus University Rotterdam. 365 pp.
- Velarde, M. D., G. Fry, and M. S. Tveit (2007). “Health effects of viewing landscapes – Landscape types in environmental psychology”. *Urban Forestry & Urban Greening* 6.4, 199–212.
- Verma, D., A. Jana, and K. Ramamritham (2018). “Quantifying Urban Surroundings Using Deep Learning Techniques: A New Proposal”. *Urban Science* 2.3. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, 78. DOI: 10.3390/urbansci2030078.
- Volpi, M. and D. Tuia (2017). “Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks”. *IEEE Transactions on Geoscience and Remote Sensing* 55.2. Conference Name: IEEE Transactions on Geoscience and Remote Sensing, 881–893. DOI: 10.1109/TGRS.2016.2616585.
- (2018). “Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images”. *ISPRS Journal of Photogrammetry and Remote Sensing* 144, 48–60. DOI: 10.1016/j.isprsjprs.2018.06.007.
- Vries, S. de, M. de Groot, and J. Boers (2012). “Eyesores in sight: Quantifying the impact of man-made elements on the scenic beauty of Dutch landscapes”. *Landscape and Urban Planning* 105.1, 118–127. DOI: 10.1016/j.landurbplan.2011.12.005.
- Wei, J., W. Yue, M. Li, and J. Gao (2022). “Mapping human perception of urban landscape from street-view images: A deep-learning approach”. *International Journal of Applied Earth Observation and Geoinformation* 112, 102886. DOI: 10.1016/j.jag.2022.102886.
- Wherrett, J. R. (1998). “Natural landscape scenic preference: techniques for evaluation and simulation.” PhD thesis. Aberdeen: Robert Gordon University.
- Wijnands, J. S., K. A. Nice, J. Thompson, H. Zhao, and M. Stevenson (2019). “Streetscape augmentation using generative adversarial networks: Insights related to health and wellbeing”. *Sustainable Cities and Society* 49, 101602. DOI: 10.1016/j.scs.2019.101602.
- Wilson, J. and G. Kelling (1982). “BROKEN WINDOWS: THE POLICE AND NEIGHBOURHOOD SAFETY”. *The Atlantic Monthly*.
- Workman, S., M. U. Rafique, H. Blanton, and N. Jacobs (2022). “Revisiting near/remote sensing with geospatial attention”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1778–1787.
- Workman, S., R. Souvenir, and N. Jacobs (2017). “Understanding and Mapping Natural Beauty”. In: *ICCV*. ICCV. Venice: IEEE, 5590–5599. DOI: 10.1109/ICCV.2017.596.
- Yao, Y., J. Zhang, Y. Hong, H. Liang, and J. He (2018). “Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data”. *Transactions in GIS* 22.2. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12330>, 561–581. DOI: 10.1111/tgis.12330.

- Zhang, F., L. Wu, D. Zhu, and Y. Liu (2019). “Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns”. *ISPRS Journal of Photogrammetry and Remote Sensing* 153, 48–58. DOI: 10.1016/j.isprsjprs.2019.04.017.
- Zhang, F., B. Zhou, L. Liu, Y. Liu, H. H. Fung, H. Lin, and C. Ratti (2018). “Measuring human perceptions of a large-scale urban region using machine learning”. *Landscape and Urban Planning* 180, 148–160. DOI: 10.1016/j.landurbplan.2018.08.020.
- Zhang, T., S. Zhang, Y. Wang, H. Yu, H. Ju, and H. Xue (2022). “Selection of the Most Scenic Viewpoints on an Island Based on Space–Time Perception: The Case of Nan’ao Island, China”. *International Journal of Environmental Research and Public Health* 19.3. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, 1309. DOI: 10.3390/ijerph19031309.
- Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba (2016). “Learning Deep Features for Discriminative Localization”. In: *CVPR*. CVPR. Las Vegas, NV, USA: IEEE, 2921–2929. DOI: 10.1109/CVPR.2016.319.
- Zhou, B., A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba (2018). “Places: A 10 Million Image Database for Scene Recognition”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6, 1452–1464. DOI: 10.1109/TPAMI.2017.2723009.
- Zhou, K., J. Yang, C. C. Loy, and Z. Liu (2022a). “Conditional Prompt Learning for Vision-Language Models”. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 16795–16804. DOI: 10.1109/CVPR52688.2022.01631.
- (2022b). “Learning to Prompt for Vision-Language Models”. *International Journal of Computer Vision* 130.9, 2337–2348. DOI: 10.1007/s11263-022-01653-1.
- Zhu, X. X., D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer (2017). “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources”. *IEEE Geoscience and Remote Sensing Magazine* 5.4. Conference Name: IEEE Geoscience and Remote Sensing Magazine, 8–36. DOI: 10.1109/MGRS.2017.2762307.
- Zhu, X. X., Y. Wang, M. Kochupillai, M. Werner, M. Häberle, E. J. Hoffmann, H. Taubenböck, D. Tuia, A. Levering, N. Jacobs, A. Kruspe, and K. Abdulahhad (2022). “Geoinformation Harvesting From Social Media Data: A community remote sensing approach”. *IEEE Geoscience and Remote Sensing Magazine* 10.4. Conference Name: IEEE Geoscience and Remote Sensing Magazine, 150–180. DOI: 10.1109/MGRS.2022.3219584.

Summary

Landscapes are an ever-present aspect of life, and whether for daily living or for relaxation, the qualities of landscapes influence those that use them. They are an important contributor to the health and well-being of humans. For instance, deprived living spaces can cause detrimental health effects, and landscapes with a high degree of scenic beauty are known to positively impact the health and mood of viewers. Evidently, understanding how landscapes can contribute to human well-being will have significant public health benefits. For this purpose, opinions about landscapes and their qualities need to be gathered from people. One way to do this is by showing landscape photos to volunteers, and to collect ratings from them. By then relating the contents of the landscape photos to their ratings, it can be better understood what drives the qualities of landscapes.

The process of deriving inferences from landscape photos with reference scores historically required a lot of manual processing. However, recent advances in machine learning methods have enabled models to automatically extract relevant information from photos, and to use this information to perform landscape quality (LQ) ratings on new photos. As such, the deep learning models that are capable of such assessments hold tremendous potential for large-scale inventorying of LQ. However, the reasoning of these models is hard to understand, and they require an enormous amount of training examples before they are usable. Models also typically only use one source of information such as photos, while often extra information is available which can be used to train better models. These problems have thus far limited the practical applicability of deep learning models for LQ assessments. The methods brought forth in this thesis are meant to address these shortcomings, and to improve the practical utilisation of deep learning models for LQ assessments.

Chapter 2 considers the prediction of liveability from aerial images for the entirety of the Netherlands. In particular, it addresses the lack of methods suitable for explaining models trained on remote sensing (RS) data. Firstly, it introduces a model which explicitly learns to relate liveability domain information (e.g. housing quality, physical environment) to liveability ratings. This model performed better than models which learned to only predict liveability from aerial images, which demonstrates that extra task information helps to train better models. By relating the model's learned reasoning to neighbourhood typologies, it could then be understood how liveability varies for different types of neighbourhood

layouts. This chapter has demonstrated that models trained on RS images can be used to extrapolate liveability surveys, and that interpretability can be part of the design process for models.

Chapter 3 continues with improving the interpretability of DL models trained on RS images, in particular for landscape beauty (or scenicness) assessments. It extends the interpretable model of chapter 2 to account for multiple possible linear relations between the intermediate task and the final scores. This idea was tested by using land cover prediction as the intermediate task, and to let the model learn 3 possible options per land cover class. The model could be used to find examples which contribute positively or negatively to scenicness. This chapter highlights that deep learning models can be used to find more complex patterns while remaining interpretable.

In **chapter 4**, the performance benefits of approaches which use multiple sources of information are tested for the task of housing quality detection. Several combinations of natural and RS images were used. Using only Google Streetview (GSV) images resulted in the best performance, approximately 30% better than using only aerial images. However, when fusing Flickr image features with aerial images, this performance gap can be brought down to 15%. While GSV images are the best-performing data-source, it is impractical for large-scale use as it is proprietary. Therefore, these results are encouraging, as the combination of Flickr and aerial overhead images demonstrates that open-source data sources are able to perform competitively.

Chapter 5 explores the potential that multimodal models pre-trained on information at the scale of the world-wide-web can offer for LQ assessments. Firstly, experiments were conducted with data-efficient learning regimes. Models trained on just a few hundred samples performed competitively compared to models trained on hundreds of thousands of examples. The findings demonstrate that it is possible to perform LQ assessments using magnitudes less reference data than previously considered possible. This chapter also proposed Landscape Prompt Ensembling, a multimodal rating approach using text and images that relies on the mind's eye view of participants. It uses text descriptions with scenicness ratings gathered from volunteers. It then uses a multimodal model which relates the given text descriptions to the contents of images in order to provide image ratings. The resulting dataset of image ratings was found to be concordant with a well-studied dataset of image scenicness ratings, both numerically (R^2 of 0.68) as well as thematically (high similarity in land cover class preferences). The findings of LPE highlight that multimodal approaches can be used to perform landscape ratings in entirely new ways in order to acquire insights beyond just the contents of images.

This thesis demonstrates that LQ assessments can be performed using less data, with better interpretability, and using approaches which better leverage multiple sources of information. These findings advance DL models towards LQ assessments at any scale that can be relied upon.

Acknowledgements

I'd like to take this opportunity to thank all the wonderful people that have helped me to complete my PhD journey, as well as those that enabled me to pursue an academic career in general.

Firstly, I'd like to thank my supervision team for their support and guidance throughout the PhD process. To my promotor Devis, thank you for all your help along the way all these years, even before the PhD started. You've been both a supervisor and a friend, and you've always treated me as a peer. Your optimism, enthusiasm, and social outlook really went a long way to making my PhD journey an unforgettable one. Despite the distance between Wageningen and Sion, and of course the pandemic, you always made sure that I was part of the team. I enjoyed the short stay in Switzerland, and I will for sure be back for more. To Diego, thank you for being my supervisor all these years. It has been very inspiring to witness your love for academia and cooperation. You were always there to listen to my silly ideas, and if I didn't have any then you were always ready to offer yours. You gave me much-needed reflection, especially during the pandemic, and you've taught me a lot about academia. In turn, I hope that I inspired you to always have some peanuts at hand in your office!

My thanks go out to all of the people I've worked together. To my friends at ECEO, thank you for making my short stay in Sion feel like home. If only it could have been longer! Thank you to the team at Washington University, I enjoyed my stay in St. Louis and for giving me new perspectives on my work. In particular, thank you Nathan for giving me the opportunity to come, I enjoyed our cooperation! And of course, thank you to everyone at GIRS. It's been a pleasure to pursue a PhD at the group that I've come to know so well during my master's programme. Thank you for the coffee table banter, the Veluweloop, the social events, and of course the GIRS band!

My heartfelt thanks go out to my friends and family for supporting me in this process. To my parents Piet and Wendy, I am forever grateful for supporting me in my pursuit of academia. Thank you Dan, for supporting me every step of the way during all these years, and for sitting out the pandemic with me. Thank you to my brother, my grandparents, cousins, and other family members for your care and support.

About the author

Alex Levering was born in 1993 in Oss, Noord-Brabant, the Netherlands. Throughout his early school years, he always had a keen interest in geography and history. In 2012 he moved to Vlissingen to study water management at the Hogeschool Zeeland. Here, he deepened his interests in geography and specialised in geo-information sciences through a minor at Wageningen University. From then on, he started using GIS methods throughout the remainder of his undergraduate studies. He graduated in 2016 with an award-winning thesis on coastal erosion in Colombia. He followed up on his interests in the geo-sciences by moving to Wageningen to pursue a master's degree in geo-information sciences. During his studies he further specialised in geo-spatial programming, and along the way he picked up a particular interest in machine-and deep learning methods. He graduated in 2019 with a thesis on recognising road incidents from the perspective of self-driving vehicles. After a brief period of work as an engineer, he returned to Wageningen University to pursue a PhD in landscape quality prediction using deep learning methods. In particular, his work aims to make deep learning methods more accessible for assessments using a variety of data sources, such as photos and remote sensing images.

Peer-reviewed Journal Publications

- Levering, A., D. Marcos, J. van Vliet, and D. Tuia (2023b). "Predicting the liveability of Dutch cities with aerial images and semantic intermediate concepts". *Remote Sensing of Environment* 287, 113454. DOI: 10.1016/j.rse.2023.113454.
- Zhu, X. X., Y. Wang, M. Kochupillai, M. Werner, M. Häberle, E. J. Hoffmann, H. Taubenböck, D. Tuia, A. Levering, N. Jacobs, A. Kruspe, and K. Abdulahhad (2022). "Geoinformation Harvesting From Social Media Data: A community remote sensing approach". *IEEE Geoscience and Remote Sensing Magazine* 10.4. Conference Name: IEEE Geoscience and Remote Sensing Magazine, 150–180. DOI: 10.1109/MGRS.2022.3219584.
- Levering, A., D. Marcos, and D. Tuia (2021b). "On the relation between landscape beauty and land cover: A case study in the U.K. at Sentinel-2 resolution with interpretable AI".

ISPRS Journal of Photogrammetry and Remote Sensing 177, 194–203. DOI: 10.1016/j.isprsjprs.2021.04.020.

Levering, A., M. Tomko, D. Tuia, and K. Khoshelham (2021c). “Detecting Unsigned Physical Road Incidents From Driver-View Images”. *IEEE Transactions on Intelligent Vehicles* 6.1. Conference Name: IEEE Transactions on Intelligent Vehicles, 24–33. DOI: 10.1109/TIV.2020.2991963.

Stronkhorst, J., A. Levering, G. Hendriksen, N. Rangel-Buitrago, and L. Rosendahl Appelquist (2018). “Regional coastal erosion assessment based on global open access data: a case study for Colombia”. *Journal of Coastal Conservation* 22, 787. DOI: 10.1007/s11852-018-0609-x.

Other Scientific Publications

Levering, A., D. Marcos, I. Havinga, and D. Tuia (n.d.). “Cross-Modal Learning of Housing Quality in Amsterdam”. In: *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GeoAI 2021*. 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GeoAI 2021, 1–4. DOI: 10.1145/3486635.3491067.

Levering, A., D. Marcos, and D. Tuia (2023a). “Time Series Analysis of Urban Liveability”. In: *2023 Joint Urban Remote Sensing Event (JURSE)*. 2023 Joint Urban Remote Sensing Event (JURSE). ISSN: 2642-9535, 1–4. DOI: 10.1109/JURSE57346.2023.10144221.

– (2021a). “Liveability from Above: Understanding Quality of Life with Overhead Imagery and Deep Neural Networks”. In: *Proceedings of the 2021 International Geoscience and Remote Sensing Symposium*. IGARSS. Brussels: IEEE.

Levering, A., D. Marcos, S. Lobry, and D. Tuia (2020). “Interpretable Scenicness from Sentinel-2 Imagery”. In: *Proceedings of the 2020 International Geoscience and Remote Sensing Symposium*. International Geoscience and Remote Sensing Symposium. Hawaii, 4.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)



Review of literature (4.5 ECTS)

- Predicting Urban Demographics Change in the Netherlands using a Multimodal Approach

Post-graduate courses (3 ECTS)

- S. Srivastava defence workshops on urban remote sensing; GIRS (2020)
- Joint Urban Remote Sensing Event 2022 Online Lounge; University of Medellin; (2022)
- Animal Tracking; WENR (2022)
- Introduction to Computer Vision; Washington University in St. Louis(2023)
- R and big data; WENR (2023)

Laboratory training and working visits (4.5 ECTS)

- Large language and vision models for predicting landscape qualities: Washington University in St. Louis (2023)

Invited review of (unpublished) journal manuscript (2 ECTS)

- TGARS: Built-up area segmentation (2021)
- RSE: Domain adaptation for global urban mapping (2021)

Competence strengthening / skills courses (1.5 ECTS)

- Scientific Integrity; WGS (2021)
- Unboxing your PhD; PE&RC (2021)
- Intensive Writing Week; WGS (2023)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.2 ECTS)

- PE&RC last year retreat (2022)
- PhD Course Workshop Carousel (2023)
- PE&RC Day (2023)

Discussion groups / local seminars / other scientific meetings (4.5 ECTS)

- Agro Food Robotics reading group; WUR (2019–2020)
- ECEO discussion seminars; EPFL Lausanne (2022)
- Symposium Satellite Classification; VU Amsterdam (2022)
- WUSTL Computer Vision spring seminar; WUSTL (2023)
- WUSTL CSE spring recruitment talks; WUSTL (2022-2023)

International symposia, workshops and conferences (8.2 ECTS)

- IGARSS; oral presentation; Online (2020)
- IGARSS; oral presentation; Online (2021)
- ACM SIGSpatial; oral presentation; Online (2022)
- JURSE; oral presentation; Crete, Greece (2023)

Societally relevant exposure (1.5 ECTS)

- Werkgroep Digitaal stiching GemeenteNL (2019-2020)

Lecturing / supervision of practical's / tutorials (0.9 ECTS)

- FTE-35306 Machine Learning (2020)
- FTE-35306 Machine Learning (2021)

Supervised MSc theses (3 ECTS)

- Calculated Experiences: using deep learning for perceived liveability in Amsterdam
- Transient attributes - Could the weather affect our subjective perception of an urban place?
- Unraveling landscape scenicness with deep learning - On the relation between weather, image aesthetics, and scenicness

This research received funding from the Laboratory of Geo-information Science and Remote Sensing of Wageningen University

Financial support from Wageningen University for the printing of this thesis is gratefully acknowledged.

Thesis printed by ProefschriftMaken

Cover illustration: "*Grasmere From The Rydal Road*", Francis Towne (1789)

