USING ADVERSARIAL NETWORKS AND INVARIANCE FOR VERTICAL CANOPY STRUCTURE REGRESSION

GEO 610 Master's Thesis

Author
Jim Buffat
13-714-134

Supervised by
Dr. Felix Morsdorf

Faculty representative
Prof. Dr. Alexander Damm

30.01.2021
Department of Geography, University of Zurich

University of Zurich^UZH

USING ADVERSARIAL NETWORKS AND INVARIANCE FOR VERTICAL CANOPY STRUCTURE REGRESSION

Jim Buffat
13-714-134

GEO 610 Master's Thesis

Supervision

Dr. Felix Morsdorf
Dr. Jan Dirk Wegner

Faculty Representative
Prof. Dr. Alexander Damm

January 2019

Abstract

Prediction of canopy structure variables from multispectral imagery is an interesting application in situations where high costs or access restrictions forbid the use of more precise LiDAR acquisition methods. The wide coverage and temporally well resolved nature of medium-resolution sensors such as Sentinel-2 MSI make inversion rather than direct measurements interesting for monitoring forests at country-wide or global scale. Good inversion performance from multispectral data is difficult to achieve because of its inherently ill-posed nature as well as atmospheric and illumination-viewing geometry induced perturbation. At smaller scales and at higher resolutions finding an inversion model is often challenged by data scarcity due spatio-temporal mismatch between ground truth and spectral imagery. This thesis presents a semi-supervised loss formulation for neural networks for inversion of multispectral TOA imagery and evaluates it for prediction of canopy height percentiles, cover fraction and vertical canopy index. The inversion problem is cast in an adversarial setting and regularized by an invariance based loss to impose implicit latent constraints. While the presented loss terms could not reduce absolute pixel-wise errors, the losses are shown to improve distributional validity both horizontally and vertically. Moreover, the losses are shown are beneficial to integrating non-corresponding data sets of spectral imagery and ground truth.

Contents

1 Introduction

1.1 Structural Canopy Variables

Physiological, biochemical and structural properties of forests measured using Remote Sensing products are closely linked to quantities relevant for the assessment of forests' role in climate change (Kumar [1], Damm et al. [2], Kükenbrink et al. [3], Antonarakis et al. [4]) and as a purveyor of ecological diversity (Brusa and Bunker [5], Singh et al. [6], Wang et al. [7], Hamraz et al. [8]). Accurately gauging these properties is vital for forest and habitat modelling as well as global climate change prediction Hence, both singular acquisitions as well as long-term monitoring of these quantities are important services for a range of topics. Of special interest in the study of forest canopies is its vertical structure (Ishii et al. [9]). Its vertical structure determines light availability at low canopy levels (Schneider et al. [10], Kükenbrink et al. [3]) which is a key quantity for the assessment of photosynthesis (Ellsworth and Reich [11]), micrometeorological applications (Chen et al. [12]), precise gas exchange estimates (Damm et al. [2]) and radiative transfer simulations (Wang and Li [13]). Maps at a global or even country-wide scale of canopy structure variables with a high spatial and temporal resolution are however either totally missing or are limited to high resolution in one of the domains (Lang et al. [14]).

Light Dectection and Ranging (LiDAR) and in particular Airborne Laser Scanning (ALS) has been a particularly successful and expanding technique to measure forest structure over large swaths since it is able to penetrate and capture forests in 3D by design and at relatively low costs (Morsdorf et al. [15, 16], Bae et al. [17]). Even though the use of ALS is expanding for acquisitions at multiple spatial scales, consistent time series measurements for large areas are currently only performed at low resolutions. Examples of this are the GEDI (Global Ecosystem Dynamics Investigation, Qi et al. [18]) and ICESAT-2 (Ice, Cloud and Land Elevation Satellite, Markus et al. [19]) missions. There also exist a growing number of ALS data sets acquired by state entities with resolutions that are sufficiently high for many large-scale vegetation applications. However, their wide-spread use is hampered by usage restrictions and non-centralized access. Furthermore, these acquisitions tend to have temporally irregular sampling patterns.

The ubiquitous and temporally well resolved availability of satellite aperture radar (SAR) and multispectral imagery is the basis for the interest in using these data sources to either improve or completely derive structural and phenological quantities (Lang et al. [14], Bae et al. [17], García et al. [20]). SAR is able to penetrate the forest to various degrees and thus indeed measures 3D structure as does ALS, though it does this at a lower resolution. The vertical structure information in spectral imagery is greatly reduced. In optical imagery, the only direct link to structural forest properties under the canopy are shadows on and light penetration through the canopy cover. However, multispectral imagery is also indirectly correlated to structure by species specific distributions and the internal correlation between structural and physiological quantities.

1.2 Physics and Model Based Approaches

Traditional remote sensing of optical reflective measurements makes use of physically and model-based approaches to analyse and monitor vegetation functioning and state. Model-based approaches rely on statistical relationships of hand-crafted spectral features to infer vegetation properties $\mathbf{x}$ from experimental observations $\mathbf{y} = g(\mathbf{x})$. Physics-based approaches make use of radiative transfer models (RTM) that find solutions to the radiative transfer equations given suitable parametrizations. RTMs for TOA (top of atmosphere) reflectance are forward models $\tilde{g}$ that depend, at least in the context of forest light regimes, on forest structure, topography, ground, wood and leaf spectra, atmospheric conditions and sun and viewing geometry. These variables will be summarized in $\omega \in \Omega$.

A range of different inversion strategies have been developed to allow for inference of vegetation structure with both statistics and physics-based models (Fawcett et al. [21], Wang and Li [22]). Formally, the goal is always to construct functions f to recover ground truth $\mathbf{y} = f(\mathbf{x} \,|\, \omega) \approx g^{-1}(\mathbf{y})$ from spectral acquisitions $\mathbf{x}$ (or $\tilde{g}^{-1}$ for RTM inversion) (Camps-Valls et al. [23]). However, both for model-based and physics-based retrieval this inversion is not well defined in a strict sense since the problem is ill-posed, i.e. even disregarding optimization problems, incomplete physical knowledge and noise there is no one-to-one relationship between spectral and ground truth domain. This invalidates the uniqueness constraint of a true inversion (Logvin et al. [24]). Moreover, the problem is most often trivially ill-posed due to the large number of free parameters that far outweigh the information content in the spectral imagery (Camps-Valls et al. [23]). This is accentuated by the fact that the level of sensitivity of radiance spectra to many structural forest quantities is hampered by noise, confounding variables or simply missing causal connections.

Model-based approaches rely on statistical modelling or simple regression of empirical data. A range of different regression and modelling techniques and algorithms are used in Remote Sensing. Large spectral data sets corresponding to biophysical ground truth exist at several spatial scales and spectral resolutions. Care must be taken that algorithms allow for generalizability since spectral data depends on multiple factors, several of which are hard to assess and varying strongly in time and across biomes (Wang and Li [22]). It is hard to construct robust statistical models across the full range of variation. This is true in particular due to the inherent structural complexity of forest canopies. With the advent of large multi- and hyperspectral data sets at increasing spatial resolution and the improving ground truth quality, the modelling techniques must prove versatile enough for pattern recognition in the spatial and spectral domain simultaneously.

Physical models offer the benefit of interpretable and plausible results in forward mode. The construction of the inversion f is normally done by sampling ω and constructing a mesh of forward simulations (look-up table LUT). An inversion of a spectral image y then proceeds by the minimization of some loss function ℓ in order to find the *closest* parameter configuration. ℓ is however a priori not more physically motivated than in the model-based case such that the measure of *closeness* may not be related to the underlying causal structure and bias the inversion. However, fine-grained control of the sampling of ω as well as the possibility of conditionally querying such LUTs improves interpretability and plausibility, also in reverse mode. Moreover, sensitivity studies over a broad range of parameters can be conducted relatively easily and at low cost once a suitable ground parameterization has been acquired. The costs and time associated with accurate 3D parametrizations make large scale application of accurate RTMs currently impossible.

Convolutional neural networks (CNNs) have proven to be very versatile function approximators for problems relying on multidimensional, spatially explicit data. CNNs are particularly well suited for imagery of optical Remote

Sensing as they offer a principled way to derive predictive fetures from both spatially and spectrally structured information upon which optical Remote Sensing applications draw. Optical remote sensing has a long tradition of defining predictive features by hand. Especially for the use of spectral data at high resolutions a more principled and target-specific way is needed since the domain knowledge cannot be encoded easily in terms of a small set of features. Automated pattern recognition by CNNs offer a way to develop models from the growing amount of remotely sensed data. In this perspective CNNs align with statistical models in that inversion functions f are derived from distributional properties of the training set. However, the use of physically motivated loss terms for the training of CNNs theoretically allows for the inclusion of prior knowledge of physical properties in the inversion. In the context of the inversion of structural canopy variables from spectral imagery, the physical constraints can refer to the interaction of radiation with heterogeneously vertically structured vegetation, but also refers to perturbative effects due to topography and atmospheric state.

1.3 Losses for Semi-Supervised Training

The data set size and distribution across the range of ground truth variability are important parameters that determine the generalizability of a CNN after training, i.e. its performance on data from unknown parts of the input distribution. In inverse problem settings in Remote Sensing, usually a spatio-temporal mismatch between spectral observations and ground truth acquisitions exists. Operating costs, difficult field conditions as well as temporal variability often pose major difficulties in assessing ground truth and spectral data simultaneously and at the same locations. On a global scale, spectral data sets with full coverage and consistent quality exist, while in situ data sets are locally restricted and often difficult to normalize to common standards. Such incompatibilities often reduce the amount of data that can be used in strictly supervised learning or, when used, induce label-noise and degrade prediction quality.

The inclusion of prior domain knowledge is a strategy to reduce the need for real samples (Ren et al. [25], Stewart and Ermon [26], Muralidhar et al. [27]). Such knowledge can be derived from pre-existing statistical or physical models. Two types of prior knowledge are of interest in this thesis. First, the prior knowledge can be **conditional** acting on input and co-domain simultaneously. Secondly, it can be solely related to the **co-domain**. This includes statistical models over forest structure variables in the co-domain. If the co-domain is multidimensional these models encode the joint distribution over the variable's observation. Note that in the multidimensional case physics-based models are possible as well.

In both cases, these models may be reformulated as losses and included as soft constraints in the global loss for which the network is optimized. The rationale of such a loss construction is to find networks which simultaneously fulfill all constraints (Karpatne et al. [28]). This strategy is also followed in applications where first-order logical statements are included (Li and Srikumar [29], Stewart and Ermon [26]) and a continuous loss formulation is not evident a priori.

The derivation of such constraints is however difficult for complex physical and observational systems. Relevant conditional constraints for prediction of forest structure variables are only superficially known, regionally different, sensor-dependent and pixel-wise, thus disregarding the information content of spatial distribution. The same restriction applies to co-domain constraints such as allometric models. On the other hand, use of physical simulation through RTMs is complicated by the fact that the causal relationship encoded in the RTM needs to be inverted as well.

1.4 Perturbation of Spectral Imagery

Perturbation of spectral input data consists in the modification of the measured signal by physical processes independent of the target variables. Modification of the spectral measurements due to changing atmospherical composition, illumination-surface-sensor geometry (Fawcett et al. [30], Dong et al. [31]) as well as effects due to adjacent terrain are conceptually most easily separable from the physical processes from which the inversion should be performed. Shadowing and adjacency effects caused internally by the canopy structure itself are known to affect common pixel-wise inversion estimates as well (Kukenbrink et al. [32], Schneider et al. [33]). However, for predictors acting on spatially explicit data, canopy-internal effects can arguably serve as additional features for structure prediction. This hypothesis applies both to spatially high and low resolved spectral imagery. In the latter case, such adjacency effects are mixed with the backscatter from other sources within the pixel neighbourhood and may in combination with surrounding pixels increase the information content of the measured signal composite per pixel. Non-canopy-related effects (atmosphere, illumination-geometry, terrain) are however considered *perturbative* and the Discussion addresses the effect of these sources of uncertainty.

While deep neural networks (DNNs) are powerful function approximators, studies have shown that DNNs in many visual intelligence tasks are vulnerable to adversarial attacks including classification, detection, segmentation and image-to-image translation (Wang et al. [34]). Small perturbations in the input can cause the network to fail, indicating that the failing networks are not stable within small regions around input samples. (Carmon et al. [35]). This is a problem that may persist even with sufficiently large data sets since the root of this phenomenon lies in the high-dimensionality of the input space (Goodfellow et al. [36]). Adversarial attacks have been proposed as a means to increase the predictor network's robustness. It was pointed out that this training setup leads to a trade-off between accuracy and adversarial robustness (Raghunathan et al. [37]). There is however recent work showing that the inclusion of unlabeled data with adversarial attacks can resolve this issue (Uesato et al. [38], Carmon et al. [35], Raghunathan et al. [37]). The inclusion of semi-supervised learning techniques is therefore argued to be an interesting pathway for reducing the network's sensitivity to perturbations.

It is still unclear to what extent lacking adversarial robustness of DNNs is relevant for regression of forest structure and how it compares to the generally large uncertainties involved in these tasks. While in a generative adversarial setting, perturbations are intentionally constructed to fool the predictor network, these were pointed out to be not necessarily valid physically and that they should be considered a worse case scenario (Mangal et al. [39]). Laugros et al. [40] also points out that increasing robustness is highly specific such that reducing the robustness to one kind of perturbation might increase the sensitivity to another kind.

1.5 Aim of Thesis

This thesis aims at training a DNN for inverting multispectral and hyperspectral images of temperate mixed forests in Switzerland to a set of canopy height percentiles p20, p50, p70 and p95 (20 %, 50 %, 70 % and 95 %), the fractional cover COV and the vertical canopy index VCI. Ground truth for these quantities is derived from ALS acquisition of the Cantons of Aargau and Fribourg. The spectral imagery is prepared from Sentinel-2 and APEX (Airborne Prism Experiment) acquisitions in Aargau.

The contribution of this thesis is the evaluation of two loss terms that aim to improve the distributional validity of the networks predictions. Both losses are compared to a simple ℓ_1 baseline. Although the training and evaluation are conducted on data from a restricted region and for a particular inversion task, the loss terms can be used more generally for inversion tasks in Optical Remote Sensing No explicit models or other prior knowledge is needed.

> **Adversarial Training:** A framework proposed by Ren et al. [25] for semi-supervised adversarial training is adopted. It resolves the difficulty of model derivation from the data by casting the inversion problem into an adversarial setting. Latent constraints in the data are expected to be learned implicitly during training by an adversary removing the need for user exploration of the data and automatically yielding constraints relevant to the learning task. Furthermore, this framework offers a way to include data sets with related input-target samples as well as data sets of unrelated samples in the input space and co-domain.
>
> **Invariance Constraint:** Given the low sensitivity of certain structure variables with respect to the spectral data, it can be expected that an extensive sampling of the perturbation distribution must be used during training in order to robustly disentangle the signal from perturbations. This is, however, not possible for many sensors for which no long-term time series exist or the ground truth is sparse in time. The present work therefore proposes a loss term that punishes the prediction difference under physically valid input perturbations. It is tested on naturally occurring perturbation between images in a time series of the same location and allows for the use of simulated data.

2 Data

The loss functions that will be investigated require a core set of corresponding ground truth and input samples. As will be described below, non-corresponding data sets of ground truth images without spectral counterpart or vice versa can be used as well. Furthermore, the loss functions will target perturbation in the spectral image over multiple acquisitions. Spectral images from multiple times are therefore required.

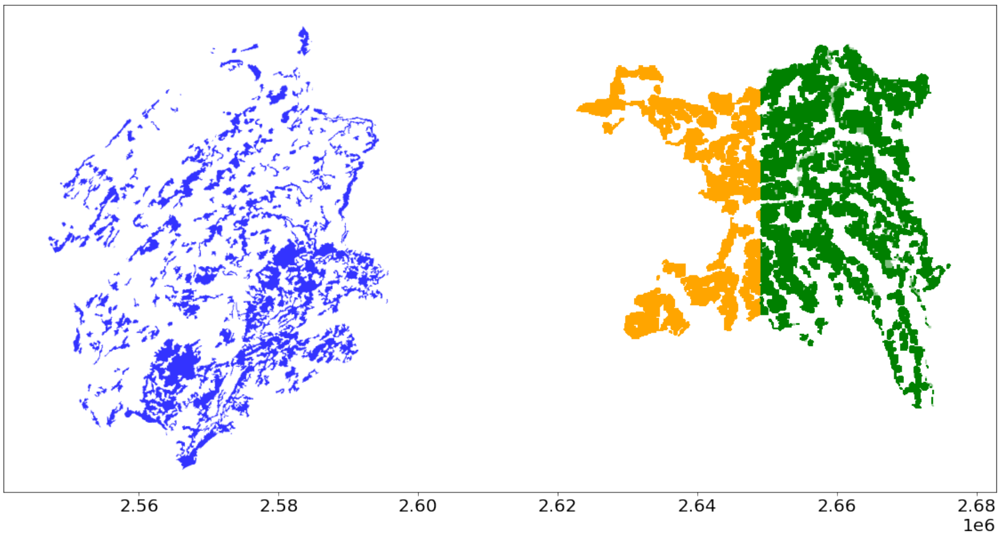

Figure 1: Sentinel-2 based data sets SINGLE and ON-OFF. In green, training images with spectral *and* ground truth coverage $\mathcal{D}_c$. In orange, validation images of $\mathcal{D}_c$ (with spectral and ground coverage). In blue, covered area of test $\mathcal{D}_c$.

Data sets consisting of 300 × 300 m square images of LiDAR derived structural canopy variables and multispectral acquisitions were assembled for training and evaluation. The spectral imagery was cropped from the Sentinel-2 MSI 1c product (Drusch et al. [41]) as well as from atmospherically corrected APEX acquisitions (Schaepman et al. [42]). All images were cropped over a grid of fixed positions such that a time series of images at fixed positions could be gathered in the case of the Sentinel images. The grid was defined such that neighbouring images overlapped by 150 m to make optimal use of the available covered forested regions. Care was taken to exclude non-forested regions: all images were spatially filtered and only those images were retained which intersected a forested region. The definition of forested regions was set as the land cover classes 311-313 and 324 of the CORINE (Büttner and Kosztra [43]) land cover classification. No other spatial selection was performed such that the images still contain other land cover classes bordering the forest mask. This includes notably human made infrastructure as well as water bodies.

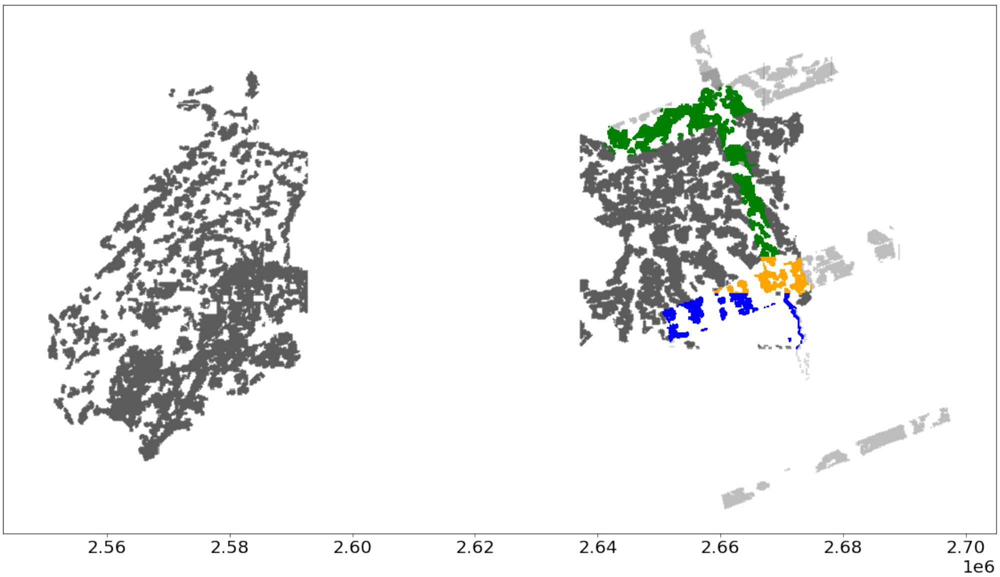

Figure 2: APEX data set. In green, training images with spectral *and* ground truth coverage $\mathcal{D}_c$. In light gray, training images with only spectral coverage $\mathcal{X}_{nc}$. In dark gray, ground truth coverage included as $\mathcal{Y}_{nc}$. In orange, validation $\mathcal{D}_c$. In blue, covered area of test $\mathcal{D}_c$.

2.1 Sentinel-2 L1c (SEN1c)

Sentinel-2 MSI Level 1c acquisitions [41] of the years 2018 - 2020 of the Canton of Aargau and from the years 2016 - 2018 of the Canton of Fribourg were used for the creation of a low resolution multispectral data set. The first 13 layers (B1 - B8, B8A, B9 - B12) were resampled to a uniform resolution of 10 m. Images covering 300×300 m were cropped at the the specified coordinates as mentioned above. All Sentinel-2 acquisitions within the time frame with a Sentinel cloud quality flag of over 80% were included. In order to reduce the impact of cloud and cloud shadows, images intersecting a previously derived cloud and cloud shadow mask as proposed by Hancher et al. [44] were excluded. The filtering by quality index and the cloud mask excluded most of the optically thick clouds, some images however still contained haze and cirrus.

The L1c product consists of radiometrically calibrated and otho-rectified TOA reflectance of Sentinel-2 acquisitions. The network is required to be robust to spectral variations arising from variable atmospheric composition and topographic conditions. The network is expected to perform an online atmospheric correction. This atmospheric correction is not explicit but task specific and assures that possible predictive features are not removed in the pre-processing. At the same time, the use of data that still contains atmospheric perturbations of the TOC signal is an interesting use-case to study the prediction variability of the network under atmospheric perturbation.

2.2 APEX

Gereoctified L2 TOC reflectance data acquired by the Airborne Imaging Spectrometer (AIS) from the Airborne Prism Experiment (APEX) (Schaepman et al. [42]) was prepared to gather a high resolution data set both in the spatial and spectral dimension. An atmospheric and topographic correction as in Richter and Schläpfer [45] was performed. With a pixel resolution of 2 m and covering a spectral range from 400 - 2500 nm over 284 bands, APEX is significantly more resolved than Sen1c. Due to the intersection of a region of the available LiDAR data set and its temporal proximity, a subset of the ECOTRANS (pers. comm. Dr. Andreas Hüni) missions were used. These were carried out in July 2018 and cover parts of the Canton of Aargau.. The missions partly overlapped. However, the overlapping regions were too small for a multi-view data set to be constructed.

2.3 Aargau and Fribourg ALS (AAR and FRI)

The ground truth for the targeted structural canopy variables was derived from two different ALs data sets covering the Canton of Aargau (AAR) (pers. comm. Dr. Felix Morsdorf) and Fribourg (FRI) [46] respectively. The AAR data set was acquired between February and March 2019, the FRI data set from October 2016 to February 2017. Both ALS acquisitions were thus taken under leaf-off conditions and include large swaths of temperate mixed forests.

	SINGLE			**ON-OFF**		
	Train.	**Val.**	**Test**	**Train.**	**Val.**	**Test**
Locations	10'985	5'719	11'682	9'042	4'479	11'682
Samples	347'431	183'525	93'050	285'956	124'566	93'050

Table 1: Table of the training, validation and test set sizes. Locations: window positions with at least one valid date. Samples: total number of cropped images.

	APEX		
	Train.	**Val.**	**Test**
$\mathcal{D}_c$	3'403	727	734
$\mathcal{Y}_{nc}$	31'511	–	–
$\mathcal{X}_{nc}$	7'251	–	–

Table 2: Table of the training, validation and test set sizes of the APEX data set.

From these point clouds the percentiles p20, p50, p70, p95, vertical canopy index VCI and fractional cover COV (Bruggisser et al. [47]) were derived. The preprocessing of the point clouds included height normalization and classification into vegetation and non-vegetation using the LASTools [48] software toolbox. All target variables were computed in pixel resolutions of 10m and 2m such that they could be aligned to the Sentinel-2 and APEX images. Subsequently, the maps of the target variables were cropped with the same image grid as the spectral data sets.

3 Methods

The thesis aims at training a Deep Neural Network (DNN) for the prediction of height percentiles (p20, p50, p70, p95), cover fraction COV and the vertical canopy index VCI of forested areas from multi- and hyperspectral imagery. No thorough investigation into the performance of the chosen architecture has been conducted. Instead, the contribution of the present project is the formulation and evaluation of two loss terms. A loss term $\mathcal{L}^{\text{COD}}$ acting in the co-domain and based in an adversarial setting is proposed. Further, a loss term $\mathcal{L}^{\text{CON}}$ operating on an invariance constraint under physically valid perturbations is introduced.

These loss terms in principle reduce the amount of ground truth labels and are closely linked to ideas of label-free training. It is argued that the application of these loss terms may prove useful for other optical inversion applications in Remote Sensing as they are based on general properties and address the standard short-comings of spectral imagery.

First, the proposed losses are required to constrain the model during training to a target space that approximates the observational distribution of the ground truth data sets. This requirement on distributional validity is addressed by casting the prediction into an adversarial setting.

A second requirement for the loss is that it encodes complex physical constraints from simulated data. Rather than deriving a loss from target-specific physically motivated constraints as in Karpatne et al. [28], a generic procedure is adopted.

Finally, the proposed loss terms address the problem of spatio-temporal mismatches in available data sources by offering the possibility to include non-corresponding data sets in the training procedure. These data sets are merely required to be sampled from the same input and output distributions.

3.1 General Procedure

Let $\mathcal{X} = \{\mathbf{x}_i | \mathbf{x}_i \sim p_x(\mathbf{x})\}_{i \in M}$ the data set of multispectral images of forest scenes $\mathbf{x}_i$ that were cropped from the spectral data sets. Here, p_x denotes the probability distribution over forest images of fixed size as seen by the acquiring sensor in a region with similar vegetative, topographic and atmospheric conditions. As described above, samples of the true target distribution p_y were derived from the LiDAR data sets. Let $\mathcal{Y} = \{\mathbf{y}_j | \mathbf{y}_j \sim p_y(\mathbf{y})\}_{j \in N}$ denote this data set. In case of spatial and temporal overlap, a data set of corresponding samples of spectral images and ground truth $\mathcal{D}_c$ can be constructed. Equivalently, non-corresponding data sets $\mathcal{X}_{nc}$ and $\mathcal{Y}_{nc}$ were constructed from images $\mathbf{y}$ and $\mathbf{x}$.

3.1.1 Supervised and Semi-Supervised Learning

A basic supervised training formulation only makes use of $\mathcal{D}_c$ by minimizing

$$\mathcal{L}^{\mathrm{S}} = \mathbb{E}_{\mathbf{x},\mathbf{y} \sim p(\mathbf{x},\mathbf{y})} \left[\ell_1\left(N_\theta(\mathbf{x}), \mathbf{y}\right)\right] \approx N^{-1} \sum_{\mathbf{x}_i, \mathbf{y}_i \in \mathcal{D}_c} \left[\ell_1\left(N_\theta(\mathbf{x}_i), \mathbf{y}_i\right)\right]$$

where ℓ_1 denotes the L1 loss (absolute error). Label-free learning as formulated by Stewart and Ermon [49] and Ermon et al. [50] reduces the need for data pairs in training by enforcing physical, logical or just observational constraints by means of regularization terms that are based on prior domain knowledge. Assuming there are probabilistic or physical models $m(\mathbf{y}|\,\mathbf{x})$ and $n(\mathbf{y})$ of target variables, the model fitting can be constrained by including terms that punish low likelihood of the output, i.e.

$$\begin{aligned}\mathcal{L}^{\mathrm{SC}} &= \mathcal{L}^{\mathrm{S}}(\mathcal{D}_c) - \lambda_m\, \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \log m(N_\theta(\mathbf{x})|\,\mathbf{x}) - \lambda_n\, \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \log n(N_\theta(\mathbf{x})) \\ &\approx \mathcal{L}^{\mathrm{S}}(\mathcal{D}_c) + \lambda_{\mathrm{CON}}\, \mathcal{L}^{\mathrm{CON}}(\mathcal{D}_c) + \lambda_{\mathrm{COD}}\, \mathcal{L}^{\mathrm{COD}}(\mathcal{X}_{nc}, \mathcal{Y}_{nc})\end{aligned}$$

where $\mathcal{L}^{\mathrm{COD}}$ and $\mathcal{L}^{\mathrm{CON}}$ are called co-domain loss and conditional loss and the λs are fixed weights. This formulation covers both the inclusion of observational constraints, as well as physical or causal relationships, since m and n can be derived from such relationships.

Note that in the present problem setting, $n(\mathbf{y})$ could be constructed **i)** from models of training set (observational constraints) and **ii)** from general allometric relationships encoding the target distribution. There is extensive work to develop allometric models for biomass prediction. The main explanatory variables used are diameter-at-breast-height (DBH), tree volume, crown area and tree height (Aabeyir et al. [51], Mugasha et al. [52], Barbosa et al. [53], Henry et al. [54]). The validity of such models is limited by environmental circumstances as well as the data resolution of the inversion problem. Upscaling, model inversion and model transformation to account for multispecies settings might be needed to adapt n to the inversion problem requiring additional validation. Conditional models for $m(\mathbf{y}|\mathbf{x})$ have been proposed (Fassnacht et al. [55], Köhler and Huth [56], Pascual et al. [57]). Similarly to the use of models for n, multiple problems and restrictions arise from using such relationships m in a general learning task. The present project instead follows another strategy to derive losses $\mathcal{L}^{\text{COD}}$ and proposes a $\mathcal{L}^{\text{CON}}$ term leveraging input sample invariance across transformations in the co-domain.

3.1.2 Co-Domain Loss $\mathcal{L}^{\text{COD}}$: Adversarial Regularization

While $\mathcal{L}^{\text{COD}}$ can be learned offline from the training distribution with the tools of traditional statistical inference, generative adversarial networks (GANs) have proven to be yield highly versatile distribution approximators for image data. As Mao et al. [58] reports, there has been extended work in the fields of image generation, but also semi-supervised tasks in texture recovery from downsampled images (Ledig et al. [59]) and classification (Salimans et al. [60]). Fundamentally, semi-supervised learning in the context of GANs aims at better approximating the label distribution by finding an implicit model of the label distribution (discriminator) alongside the the main model. This is necessary, if minimization of a supervised loss such as ℓ_1 is hard to obtain or inaccurate. P-norm losses are often used without physical basis in regression for lack of a physically sensible metric, such that the true distribution may be approximated but the solution is biased to satisfy properties implicitly defined by ℓ_1. Ren et al. [25] propose a semi-supervised setting for regression, where the explicit $n(\mathbf{y})$ is replaced with an implicit likelihood function that is learned during the training of the predictor network. They use the same adversarial setting that is used for the training of GANs to learn implicitly the relationships n.

The present thesis adopts this adversarial formulation to derive a codomain loss. As was pointed out above, the relationships n between structural forest variables are difficult to parametrize explicitly. However, while the limitation of validity of the implicit likelihood n is not resolved in principle, a fully general model for n might not be needed to constrain the prediction of the target variables sufficiently. In an adversarial setting constraints are learned implicitly. Thus, the degree of their complexity is explored by the discriminator and not arbitrarily set by a predefined model. The appeal of using an adversarial framework lies in its general and principled applicability especially in domains where the formulation of an explicit n is otherwise costly and difficult.

A GAN is defined by a minimax loss as in Arjovsky et al. [61]

$$\min_{\theta} \max_{\phi} \mathbb{E}_{p_y} \left[\log D_{\phi}(\mathbf{y})\right] + \mathbb{E}_{p_x} \left[\log(1 - D_{\phi}(N_{\theta}(\mathbf{x})))\right], \tag{1}$$

with D_{ϕ} a discriminator and N_{θ} the model that is trained for inversion. Note that in the present formulation the predictor network N_{θ} has the role of the generator in a GAN. As opposed to a standard GAN, the predictor does not act on a randomly sampled input vector.

Multiple variations of this formulation have been proposed as the loss in this formulation is not well behaved. D_{ϕ} tends to saturate fast, leading to uninformative gradients during optimization, e.g. both the least-square formulation (Mao et al. [58]) and the replacement of the above log-loss with a regularized Wasserstein loss (Gulrajani et al. [62]) aim at mitigating this risk.

The present thesis adopted the least-squares loss formulation first advanced by Mao et al. [58], which reformulates the minimax optimization over θ and ϕ as

$$\min_{\phi} \mathcal{L}^{\mathrm{D}}_{a,b} = \min_{\phi} \frac{1}{2} \mathbb{E}_{p_y} \left[(a - D_{\phi}(N_{\theta}(\mathbf{x})))^2\right] + \frac{1}{2} \mathbb{E}_{p_y} \left[(b - D_{\phi}(\mathbf{y}))^2\right] \tag{2}$$

$$\min_{\theta} \mathcal{L}^{\mathrm{N}}_{b} = \min_{\theta} \frac{1}{2} \mathbb{E}_{p_x} \left[(b - D_{\phi}(N_{\theta}(\mathbf{x})))^2\right], \tag{3}$$

where $a = 0$ and $b = 1$ are values for fake and true classes.

In the present project the application of this formulation has been extended from images to image regions at different scales by replacing p with probability distributions p^i defined over image regions at different subsampling scales. The discriminator is defined as a function

$$D_{\phi} : \mathbf{y} \to (\mathbf{c}_0, \dots \mathbf{c}_M; \hat{\mathrm{c}}) \quad \text{where} \quad \mathbf{c}_i \in [0,1]^{w_i \times w_i},\ w_i = \left\lfloor \frac{w_0}{2^{i+1}} \right\rfloor$$
$$\text{and} \quad \hat{\mathrm{c}} \in [0,1]$$

such that $\mathbf{c}_i$ are images at a decreasing scale. Each $\mathbf{c}_i$ covers a larger number of pixels in $\mathbf{y}$ such that each $\mathbf{c}_i$ is interpreting D's confidence in the veracity of an image region at subsampling level i. $\hat{c}$ represents a scalar output for the whole image as in a traditional discriminator. Accordingly, the adversarial regularization is performed on multiple scales by putting as loss terms for D_ϕ and N_θ

$$\begin{aligned}\mathcal{L}^{\mathrm{N}} &= |M+1|^{-1} \left(\sum_{i \in M} \sum_{j \leq W_i} W_i^{-1} \mathcal{L}_b^{\mathrm{N}}(c_{ij}(\mathbf{x})) \; + \; \mathcal{L}_b^{\mathrm{N}}(\hat{c}(\mathbf{x})) \right) \\ \mathcal{L}^{\mathrm{D}} &= \frac{|M+1|^{-1}}{2} \left(\sum_{i \in M} \sum_{j \leq W_i} W_i^{-1} \mathcal{L}_{a,b}^{\mathrm{D}}(c_{ij}(\mathbf{x}), c_{ij}(\mathbf{y})) \; + \; \mathcal{L}_{a,b}^{\mathrm{D}}(\hat{c}(\mathbf{x}), \hat{c}(\mathbf{y})) \right),\end{aligned}$$

where c_{ij} is a single pixel in the confidence map $\mathbf{c}_i$ and $W_i = w_i^2$ is $\mathbf{c}_i$'s size. Thus, the loss simply is the mean over all confidence maps $\mathbf{c}_i$ and pixels in the confidence maps.

The use of a non-scalar discriminator, i.e. the use of image regions of different scales as an argument for the discriminator rather than the whole image is different from the classical GAN architecture. It is motivated by the fact that in the present problem setting the distribution p over which D_ϕ acts, could be defined at multiple scales because characteristic scales of forest structure are typically smaller than the chosen window size. Contrary to classical GAN settings, where there is a single object instance sampled from p per image, there are multiple instances per image for the distributions p^i. The concept of discriminating single objects per image as in classical GANs can thus be extended by evaluating discriminator and predictor simultaneously on different scales.

While p covers the sub-distributions p^i jointly, it is argued that addressing p^i explicitly has benefits.

1. **Multi-Scale Features:** Pattern recognition by D_ϕ is constrained to windows of increasing size. Since the target variable distributions vary across scale, the discriminator is incited to find discriminative features at multiple scales. This can prove helpful, especially if an interpretation of discriminator features is required.

2. **Vanishing Gradients:** The simultaneous evaluation of many instances

smoothens the loss function, i.e. the predictor does not fail the discriminator test completely if it underperforms only in some image regions. This stabilizes training by suppressing the problem of vanishing gradients.

3. **Minibatch Discrimination:** Effectively, each image at each scale can be treated as a batch of multiple training samples. The simultaneous application of the discriminator to a batch of samples in traditional GANs is known as Minibatch Discrimination and was introduced in Salimans et al. [60] as a means to reduce collapse of the generator. While mode collapse is avoided at low image frequencies that can be captured by the ℓ_1 loss, it is potentially still a problem for high-frequencies. It is argued that evaluation on multiple scales may alleviate this problem.

4. **PatchGAN** Isola et al. [63] has introduced a GAN architecture that only acts on subpatches of the image and found an improvement in resolving high-frequency structures.

The simultaneous use of a pixel-wise loss with a GAN has been used for other conditional image-to-image translation tasks (for an overview see Isola et al. [63]). As Isola et al. [63] points out, the pixel-wise ℓ_1 loss in these cases approximates well the low frequency parts in the prediction, while the GAN architecture is responsible for capturing of high-frequency structures. In the present context, where the aim is regression, it is important that the high-frequency part of the prediction does not impact on the pixel-wise absolute accuracy that's well approximated by ℓ_1 alone. The loss in the next section can be understood as a regularization of the adversarial setting for this reason.

3.1.3 Conditional Loss $\mathcal{L}^{\mathrm{CON}}$: Invariance Consistency

The present project explores the possibility of using invariance properties of the inversion problem to train the regression model N_θ. The basic interest behind such a procedure lies in the implicit inclusion of physical constraints in the inversion. The same restrictions that were discussed for explicit codomain models n apply for conditional models m.

Constraining the model inversion physically has been shown to improve causal validity and generalizability in other applications (Daw et al. [64])

and could improve robustness in the present problem setting. As was pointed out above, small perturbations in the scenes can have large effects in CNN based regression. Hard physical constraints enforce guarantees that make a predictor more robust (Mohan et al. [65]). The effect of physical constraints is less clear if included as soft constraints, as is the case in this thesis. Due to the perturbation in the present problem setting being predominantly caused by atmospheric variation and illumination-viewing-geometry, the inclusion of some prior knowledge informing the network of the causal structure relating ground truth and spectral imagery is argued to be crucial. This argument evidently becomes even more acute when the training set is small and only partially representative of the input distribution used in future model application.

The presented invariance-based loss may prove especially useful for integrating training sets of simulated and true data. Simplified scene parametrization due to lacking ground knowledge can induce a significant domain gap between simulated and true imagery which biases the predictor during training. It is hypothesized that the impact of this problem can be reduced by making use of invariance. As will be clear from the design of $\mathcal{L}^{\text{CON}}$, the focus on invariance enhances the reduction of non-explanatory features rather than the difference in the absolute value of the prediction.

Let $s = (v_s, X_s)$ be a set of parameters needed for a simulation of a spectral image $\mathbf{y}_s$ of some forest scene. s is split in two sets in order to stress that the target variables $\mathbf{x}_s$ are derived from the subset X_s, whereas v_s denotes the set of parameters independent of the target variables $\mathbf{x}_s$. From this a data set

$$\mathcal{D}_{inv} = \{(\mathcal{M}(T_{\psi_j} \circ s_i), x)\}_{j \leq R} \quad \text{with } T_\psi \circ s_i = (T_\psi \circ v_{s_i}, X_{s_i})$$

can be constructed for the forest scenes s_i under transformations T_ψ parametrized by a vector ψ. Note that T_ψ only act on parameters that do not affect the target variables x. This effectively samples the space of possible realizations of y along the lines of the invariant x.

As the predictor is required to satisfy these invariances, the variational distance between the predictions can be formulated as

$$\mathcal{L}^{\text{CON}}(\mathcal{D}_{inv}) = \mathbb{E}_{\psi \times \psi'}\left[\ell_1\left(N_\theta \circ \mathcal{M} \circ T_\psi \circ s,\ N_\theta \circ \mathcal{M} \circ T_{\psi'} \circ s\right)\right] \tag{4}$$

It is clear that in this formulation the bootstrapping may cause problems for plain gradient descent. Preliminary tests have indeed shown this formulation to result in highly unstable gradients. This is to be expected since θ appears in both terms. In order to stabilize learning, the parameter update of θ was performed over only one of the two terms effectively treating the other as a ground truth image.

Three variations of this loss were implemented and evaluated

$$\mathrm{CON}_1 : \ell_1(\mathbf{y}_1, [\mathbf{y}_2]) \ \left[\frac{\ell_1(\mathbf{y}_1, \mathbf{y}) + \ell_1(\mathbf{y}_2, \mathbf{y})}{\max \ell_1(\mathbf{y}_1, \mathbf{y}) + \ell_1(\mathbf{y}_2, \mathbf{y})}\right]$$

$$\mathrm{CON}_2 : m \circ \ell_1(\mathbf{y}_1, [\mathbf{y}_2])$$

$$\mathrm{CON}_3 : m \circ \ell_1(\langle \mathbf{y}_1 \rangle, \langle [\mathbf{y}_2] \rangle)$$

where $\mathbf{y}_i = N_\theta(\mathbf{x}_i)$ for $\mathbf{x}_1$ and $\mathbf{x}_2$ are two predictions in random order, y is the ground truth, ℓ_1 and max are understood to be broadcast over all pixels, $[\,\cdot\,]$ denotes exclusion of the term from the backpropagation, $\langle\,\cdot\,\rangle$ denotes normalization with mean and standard deviation and m are three max pooling layers with a step size of 3. The following motivates the three formulations of $\mathcal{L}^{\mathrm{CON}}$:

> **CON$_1$** The prediction differences in an image are weighted. The weights are high for pixels with large differences, low-difference pixels are weighted less. The weights intend to reduce the risk of a high influence of low-difference pixels on the gradient if they are much more frequent than high-difference pixels.
>
> **CON$_2$** As above, pixels with high variation are selectively chosen for the gradient. Here the selection is not done globally but in different image regions.
>
> **CON$_3$** Instead of comparing absolute values, the loss now compares relative differences. The image normalization $\langle\,\cdot\,\rangle$ effectively redefines $\mathcal{L}^{\mathrm{CON}}$ to punish uniquely prediction differences that are caused by different relative height changes in the spatial dimension.

While samples $\mathcal{M} \circ T_\psi(s)$ can be simulated, time series data can be used as well if the target variables y can be treated to be invariant across time. This effectively reinterprets $\mathcal{M} \circ T_\psi(s)$ as the operation connecting the acquisitions in the time series. Thus, contrary to a specific variation of the simulation parameterization, the use of time series data can be seen as sampling from the true parameter distribution in an unsupervised manner. While this removes the possibility to constrain the variation to specif domains, there is no risk of biasing the network by domain gap. Furthermore, the loss functions impact can be studied without the influence of the failures in the simulation.

However, it must be stressed that the training on simulated data sets $\mathcal{D}_{inv}$ that isolate specific perturbating physical phenomena is appealing. This would allow for including narrowly defined physical losses covering single phenomena. It would allow for an evaluation of the respective importance of physical processes and ground structure as features as compared to perturbators of the inversion problem. The present thesis only evaluates $\mathcal{L}^{\text{CON}}$ on time series and shows smaller experiments with simulated data.

3.2 Architecture

3.2.1 Predictor

The predictor network consists of an encoding input block, a middle block and an output block as depicted in Fig. 3. The middle block is an Xception type network as first presented by Chollet [66] and successfully used for a similar forest structure prediction task in Lang et al. [67].

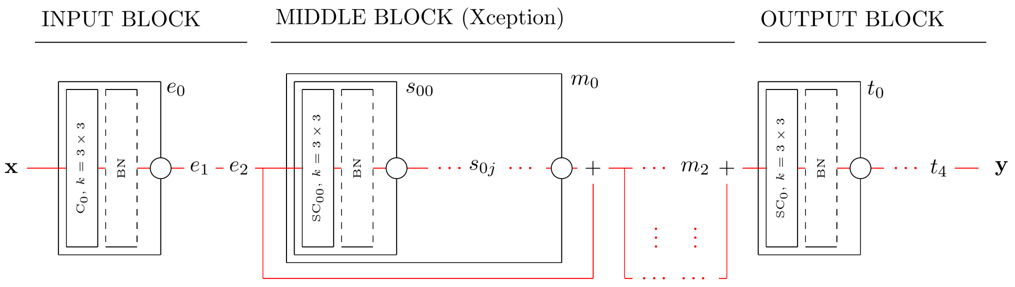

Figure 3: Schematic Overview of predictor N_θ. Circles denote ReLU nonlinearities. Letters at a box's top right define the a box. *BN* denotes Batch Norm. k denotes the convolutional window size. C denotes convolution, SC denotes separable convolution. For a complete specification of convolutional ouptut layers, see Tab. 3

The input block consists of stacked convolutional blocks e_i that project the low dimensional pixels of 13 bands into a high-dimensional space over which the the Xception network can act.

In the case of the ON-OFF runs, two input images are provided simulaneously, the aim being to leverage additional information relating to the seasonal image difference. Since image statistics greatly differ between the two seasons, each input was given a separate input block as can be seen in Fig. 4. These input blocks were then joined by stacking their output together and feeding it into a residual block m_{-1}.

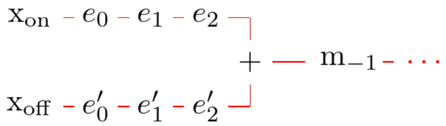
x_{on} – e_0 – e_1 – e_2
+ — m_{-1} – ···
x_{off} – e'_0 – e'_1 – e'_2

Figure 4: Double Input Layer with Residual block m_{-1} (SC). The + sign represents image stacking. Apostrophes highlight the weight independence.

An Xception network consists multiple residual blocks m_i each of which is a repetition of multiple Separable Convolutions (SC) as is depicted in Fig. 3. Residual blocks have been introduced by He et al. [68] and address the problem of vanishing gradients in DNNs by adding short-cut connections. These short-cut connections pose the fitting as residual mapping which is easier to optimize (He et al. [68], Zhang et al. [69]). SCs replace the default convolution over multiple layers by a depth-wise convolution followed by by a point-wise convolution. This effectively separates spatial and and channel-wise convolution under the assumption that channel and spatial correlations can be decoupled [66].

The output block's task is to reduce the high-dimensional representation of a scene to a point in the target space. It consists therefore of a cascaded reduction of the pixel dimensionality over SC blocks.

All convolutions in the Xception Block are followed by a batch norm layer (BN) (Ioffe and Szegedy [70]) and a ReLU non-linearity except for the last layer in the output block. The consistent use of BN layers is standard in Xception networks and has been shown to accelerate training in a such wide range of DNN architectures that it has become quasi-standard for use with ReLU non-linearities (Wu et al. [71]).

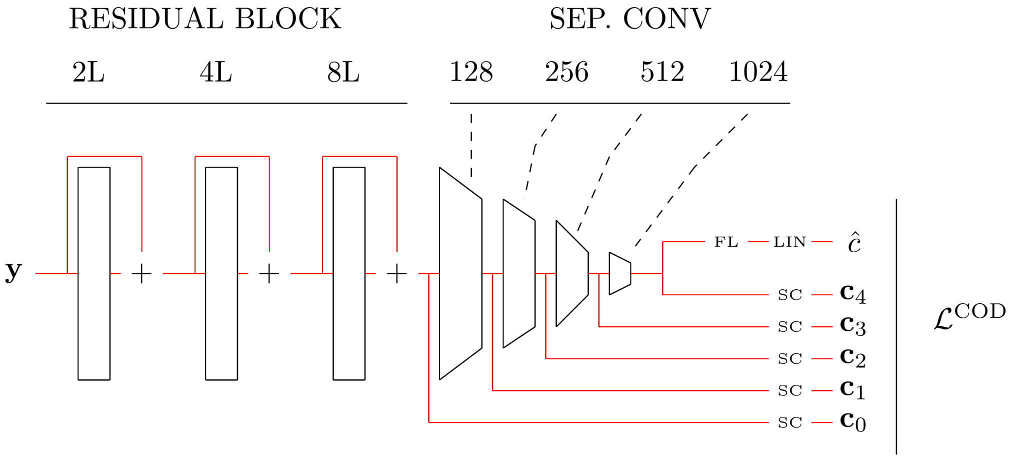

Figure 5: Schematic Overview of D_ϕ. *SC* denotes a separable convolution with sigmoid non-linearity. *LIN* denotes a linear layer with sigmoid non-linearity. *FL* denotes flattening over the image dimensions. *SC* denotes a separable convolution. $L = 6$ is the number of target variables. All numbers reference the number of convolutional output layers.

All convolutions were padded in such a manner that the output preserves the input image shape. For the same reason, all convolutions are applied with step size one.

3.2.2 Discriminator

The discriminator architecture consists of an input block of 3 residual layers that preserves the input image dimensions. A second block is is chained to the input block, which follows GAN architecture guidelines proposed by Radford et al. [72]. See Fig. 5 for a visualisation. The input block was prepended to yield an encoding for the first confidence map c_0. It consists of three residual SC layers (same architecture as t in Fig. 3 with short-cut connections). In order to account for the multi-scale output required by $\mathcal{L}^{\text{CON}}$ the Sep.Conv block has an output after every convolutional layer. Moreover, all convolutions were replaced by separable convolutions. Dropout layers with drop out probability of 0.25 % and BN layers were stacked on these SC layers. In order to regularize the discriminator, spectral regularization was applied to all convolutions. The non-linearities used in the discriminator are ReLUs apart from the last SC, where a sigmoid was used to satisfy the requirement $c \in [0, 1]$.

	-1	0	1	2	3	4
e_i	–	32	64	128	–	–
m_i	(128 3)	(128 2)	(128 2)	(256 3)	–	–
t_i	–	128	64	32	16	6

Table 3: Predictor model specification. For each block, number of output layers C_i, SC_{ij} and SC_i ordered by i as in Fig. 3. For m_i, the number of repetitions j is given as well (second number). t_4 has 6 output layers for each of p20, p50, p70, p95, VCI and COV.

4 Experiments

In the following section experiments on Sentinel and APEX data are evaluated and compared. The different experimental set-ups are listed in Tab. 4 and sample predictions are shown in Figs. 6 to 8, 25 and 26.

The goal of the evaluation is to answer the following questions

A How do augmented loss configurations impact prediction quality pixel-wise and distributionally in SINGLE, ON-OFF and APEX experiments?

B Are these differences consistent across runs with different CON loss in the SINGLE experiment?

C Can the adversarial scheme improve pixel-wise accuracy in a setting with $|\mathcal{D}_{inv}| \gg |\mathcal{D}_c|$, i.e. with a large non-coresponding data fraction due to spatio-temporal mismatch?

4.1 Quantitative Evaluation

4.1.1 Pixel-Wise Comparison

The Mean Absolute Error (MAE) as well as R^2 scores between prediction and LiDAR-derived ground truth are computed. A basic interest is whether the augmented losses can increase pixel-wise prediction accuracy of the regression. There is no guarantee that predictions matching distributional constraints such as those implicitly learned by the discriminator, yield improved pixel-wise prediction accuracy (such as MAE). The retrieval of relative variation within an image is not directly connected to pixel-wise performance. On the contrary, an increased distributional accuracy may reduce pixel-wise accuracy in the presented set-up. The loss configuration with ADV increase high-frequency sensitivity, but it is arguably the high-frequency features that expose the largest variation under perturbation. Furthermore it is known from training with adversarial defence that robustness and accuracy can be at odds (Zhang et al. [73]).

Good distributional matching independently of pixel-wise error reduction is a valuable feature on its own. It is argued that good matching of relative tar-

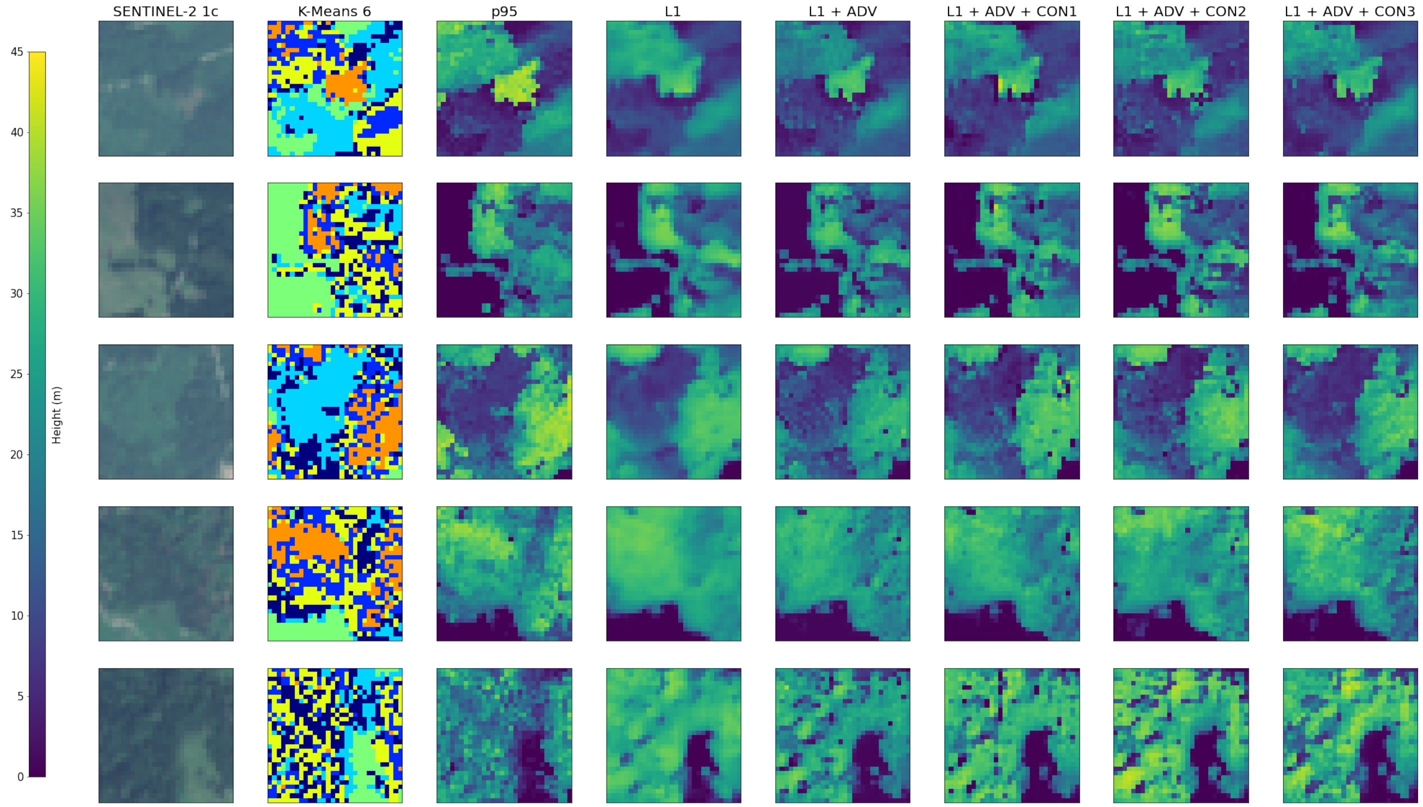

Figure 6: Sample p95 prediction results for SINGLE runs. Columns left - right: Sentinel2 input, classified ground truth, p95 ground truth, SINGLE runs. Classification colors as in Fig. 1.

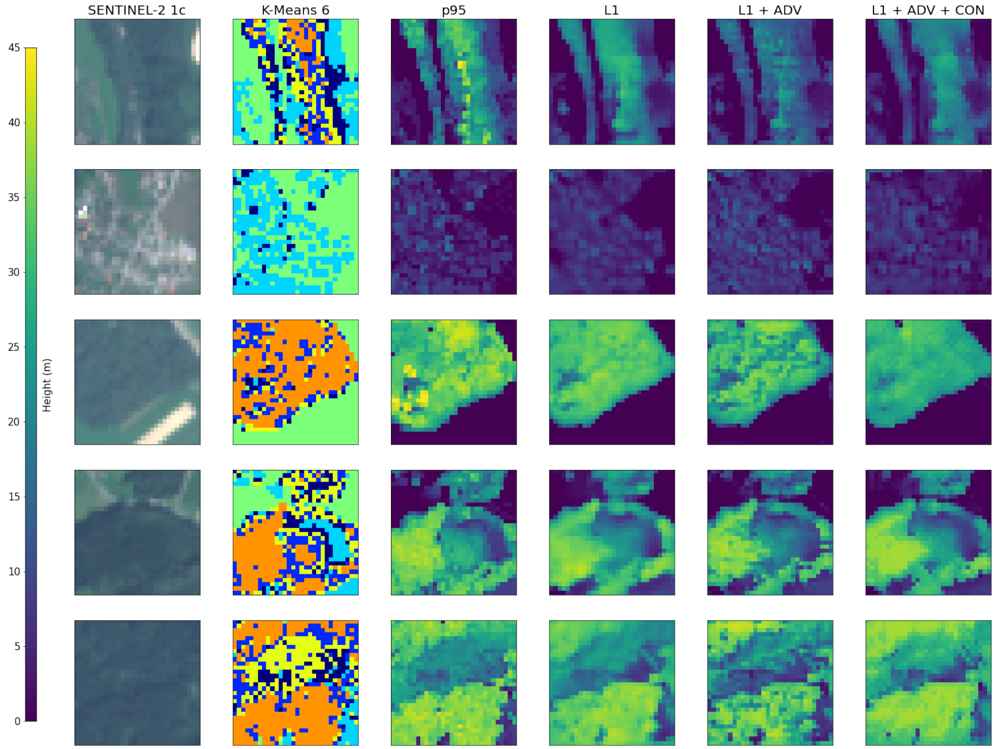

Figure 7: Sample p95 prediction results for ON-OFF runs. Columns left - right: Sentinel2 input, classified ground truth, p95 ground truth, SINGLE runs. Classification colors as in Fig. 1.

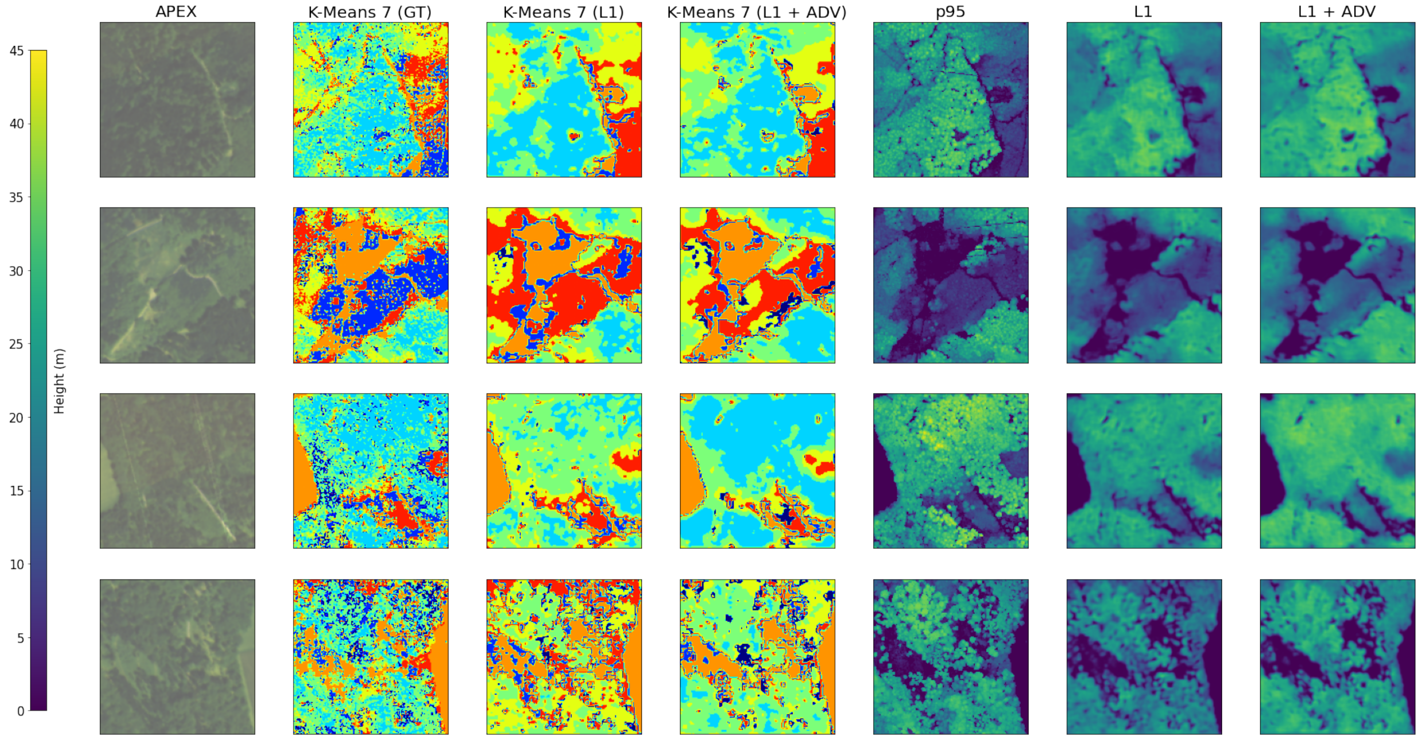

Figure 8: Sample p95 prediction results for APEX runs. Columns left - right: APEX input, classified ground truth, classified predictions, p95 predictions. Classification colors as in Fig. 2.

get variation as well as well as ecologically plausible vertical representation can influence analysis that is derived from regressed canopy variables significantly. From a practical view point for example, texture derived features have been shown to be good ecological predictors for biomass and degradation history. (Bourgoin et al. [74], Ploton et al. [75]).

Furthermore, since there is no trivial absolute height information in the spectral imagery, absolute height predictions are subject to be biased by the mean absolute height of the training data. It is therefore interesting to evaluate the predictions' sensitivity to local, relative target variation. This relative variation will be referred to as spatial structure. In order to evaluate the spatial structure of the prediction of each target variable, image statistics over the prediction were derived. The following analysis computes the local Pearson Correlation (PC) and the Kullback-Leibler divergence (KL) of normalized predictions to evaluate the recovery of spatial structures.

As the prediction of the target variables is done simultaneously (multi-target), a point of interest is whether the augmented losses improve the predictions in terms of the joint probability of the target variables. Multi-target prediction in principle allows the network to learn constraints between the target variables. The adversarial setting is expected to enhance the recovery of such constraints. To evaluate the validity of the joint percentile distribution, the predictions are subjected to a cross-entropy analysis.

4.1.2 Spatial Structure: Pearson Correlation

The PC between prediction and ground truth was evaluated over all images in the test set. Specifically, the PC was calculated as

$$s = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{5}$$

where σ_x, σ_y and σ_{xy} were calculated with

$$\sigma(x) = \sqrt{\mathbb{E}[x^2] - \mathbb{E}[x]^2} \tag{6}$$

$$\sigma(x, y) = \mathbb{E}[xy] - \mathbb{E}[x]\,\mathbb{E}[y]. \tag{7}$$

The expectation in the equations above were approximated with a gaussian kernel with $\sigma = 300$, 50 and 20 m. PC is invariant to linear transformations to the data (Sviridov et al. [76]). This is due to the subtraction of all means and subsequent normalization by the standard deviation of ground truth and prediction. For this reason, it was considered suitable to evaluate the trends in the predicted height distributions independently of the amplitude.

4.1.3 Spatial Structure: Kullback-Leibler Divergence

While the PC compares the local variation in the target variable pixel per pixel, the KL divergence measures the regional distributional difference. In the case of KL the pixel-wise association is ignored. Both KL and PC assess the local quality of the prediction's spatial structure. The KL divergence is however less restrictive as to the exact value of the prediction locality. The interest in using a KL divergence and weakening the importance of local accuracy in evaluation is based in the fact that the adversarial loss $\mathcal{L}^{\mathrm{COD}}$ is not localized but bound to image regions. While a priori this does not preclude a per pixel improvement, its main impact likely is to be found in the structure prediction over regions spanning multiple pixels. Equally, the predictive features in the present inversion problem are not locally explicit across multiple input realizations, e.g. textures with predictive power can shift in a multi-view setting due to shadows and viewing geometry.

The KL divergence per image between ground truth histograms q_l and prediction histograms p_l over L windows and B bins is defined as

$$D_{KL}(q\,||\,p) = \sum_{l \leq L} L^{-1} \sum_{i \leq B} q_{li} \log \frac{q_{li}}{p_{li}}. \tag{8}$$

The use of KL divergence for texture evaluation is well documented (Mathiassen et al. [77], Maliani et al. [78], Do and Vetterli [79]). When used in minimization, KL is often used along with a parametric density model to estimate the similarity. This model normally is first fitted to a distribution of predefined features. The KL divergence often can be computed analytically in this case. Here, no such parametric model was used. Instead, the empirical distribution was binned using a standard heuristic for setting the bin width. This can be considered to be a non-parametric density modelling. While the resulting density models depend on the binning procedure, it was preferred over a parametric model due to its ease of implementation and straightforward interpretability of failure modes, which would have been entangled with some fitting procedure for any parametric model.

Additionally to a plain KL divergence, the Jensen-Shannon (JS) divergence (Lin [80])

$$D_{JS}(q\,||\,p) = \frac{1}{2}D_{KL}(q\,||\,m) + \frac{1}{2}D_{KL}(p\,||\,m),\ m = \frac{q+p}{2} \tag{9}$$

was evaluated. JS can be considered a smoothed and bounded version of the KL divergence (Weng [81], Kadir et al. [82]). Since it was found to be more numerically stable during inspection of single samples it was included in the present analysis serving as cross reference.

The JS and KL divergence were applied on a moving window of size w_p with step size $\lfloor w_p/2 \rfloor$ on each image such that the evaluated windows within an image partly overlapped. Prior to applying the divergences, each window was normalized with its mean μ and standard deviation σ

$$w_l \leftarrow \frac{w_l - \mu(w_l)}{\sigma(w_l)}. \tag{10}$$

This was done in order to disregard absolute values and take into account only the local relative variation of the target variable. In order to minimize the relevance of non-vegetated pixels, only images with a mean height ≥ 5 m were included.

The bin sizes were dynamically chosen for each evaluated subwindow with the Freedman-Diaconis rule (Freedman and Diaconis [83]). The bins were fixed on the ground truth data such that comparability between different runs was guaranteed. Choosing dynamically the bin size prevented oversampling in regions with low structural variabilty such as grass lands and fields which could otherwise result in histograms with empty bins. On the other hand, a fixed bin assignment could cause overly simplified histograms leading to insensitivity of the evaluation to robust differences smaller than the bin width. Remaining empty bins were avoided by adding a minimum count of one and renormalization. No thorough investigation of the Freedman-Diaconis rule as a good histogram estimator for the KL (JS) calculation was performed, however.

4.1.4 Joint Target Distribution

The augmented losses in the present project aim at approximating the ground truth in distribution both spatially but also over the target space. In order to validate the predictions with respect to their distributional accuracy a probabilistic model of the target space is needed.

As was pointed out previously, the derivation of such models in general is complicated for forest structure. A descriptive analysis for vertical canopy structure in North-Western Switzerland is for example proposed by Leiterer et al. [84, 85]. While the specific nature of such models was deemed inappropriate for direct inclusion in general pyhsical loss terms, it is still possible to use them for evaluation.

A probabilistic model over the joint distribution p_p of p20, p50, p70 and p95 was derived to compare the performance of the different loss configurations. Due to its efficiency and ease of use with large amounts of data a K-Means clustering was used to find a cluster c_i. The number of clusters to fit n_c was determined by deriving the gap statistic as proposed by Tibshirani et al. [86] on a randomly sampled subset of the training set. The steps to calculate the gap statistic as outlined in [86] are shortly summarized in the following paragraph.

The gap statistic evaluates the difference g_k

$$g_k = \mathbb{E}[\log w_k(S)] - \log w_k(s) \text{ with } w_k(s) = \sum_{r \leq k} \sum_{x_i \in C_r} ||x_i - c_r||^2 \tag{11}$$

for a clustering with k clusters C_r. As argued for in Tibshirani et al. [86] the estimation over random reference distributions S can be reduced to a mean over randomly sampled uniform distributions. Thus, $n_S = 100$ random uniform distributions were sampled within the bounds of the empirical joint percentile distribution p_p. Subsequently, n_c was fixed as

$$n_c = \arg\min_k g_k \geq g_{k+1} - s_{k+1} \text{ with } s_k = \text{std}(w_k(S))\sqrt{1 + 1/n_S} \tag{12}$$

The final fitting was performed on the Aargau training set for SINGLE and ON-OFF runs as outlined in Fig. 1. For the high resolution APEX data set the same procedure was followed on the APEX training data. However, only a randomly sampled subset of this data set was chosen for the gap statistic (1'000 images). The criterion was satisfied for SINGLE and ON-OFF at $n_c = 6$ and for APEX $n_c = 7$. A visualization of the clusterings is shown in Figs. 23 and 24.

In order to cast the K-Means classification into a probabilistic model, the assignment probability distribution of pixel x_i to cluster C_k was defined to be

$$p(x_i \in C_j) = \frac{1}{||x_i - c_j||} \left(\sum_{r \leq k} \frac{1}{||x - c_r||} \right)^{-1} . \tag{13}$$

Both ground truth g and predicted samples p in the validation set were clustered yielding partitions $C_j(g)$ and $C_j(p)$. The multi-label cross entropy

$$\begin{aligned} c_e &= -\frac{1}{N} \sum_{i \leq N} \sum_{t \leq T} p(x_i \in C_t) \log p\left(N_\theta(y_i)\right) \in C_t) && (14) \\ &= -\frac{1}{N} \sum_{i \leq N} \sum_{t \leq T} p_g^t(x_i) \log p_p^t(x_i) && (15) \end{aligned}$$

between the ground truth distribution p_g and the predicted distribution p_p takes account of low density regions in the target space. c_e effectively is a log loss weighed by the uncertainty of the ground truth classification. For the present analysis this means that deviations in the percentile predictions are weighed by the probability model. Evaluating c_e rather than a classifcation based metric such as accuracy is precisely preferred because of this weighting. It accounts for the uncertainty in the ground truth labels. Effectively, it reduces the contribution of target structures, that were less present during training, to the evaluation (lying in low density regions of p_g).

4.1.5 Prediction Variability

As Richter and Schläpfer [87] state, atmosphere, solar illumination, sensor viewing geometry and terrain information have to be taken into account for an accurate retrieval of surface reflectance. Normalization of reflectance data, i.e. the removal of such effects by recalibration, is an integral part of many reflectance-based inversion algorithms. All of the above mentioned perturbations to a normalized image must be resolved by ancillary data, e.g. the sun and viewing geometry, atmospheric composition and surface BRDF. It can be expected that especially algorithms that rely on single pixel reflectance spectra as input suffer from any remaining non-predictive variability since these effects can not be mitigated by a spatial context. Indeed, misclassification and biased estimation of biophysical and chemical variables are reported for example due to anisotropy effects (Weyermann et al. [88]).

The method presented here does not use atmospheric correction to reduce perturbative effects in the inversion. Also no topographic correction is applied apart from the geometric Sentinel L1c ortho-rectification. The reason for this is that the CNN and its optimization are expected **i)** to express predictive features that are approximately invariant under these perturbations and **ii)** to be able to extract a set of predictive features that apply to different perturbative regimes. The use of non-corrected data was preferred because it allowed for $\mathcal{L}^{\text{CON}}$ to effectively act as physically based loss, while any correction would have represented a non-interpretable intermediate step.

In order to assess the impact of above mentioned perturbations on the prediction, the Mean Variation Error (MVE) and the ratio c_{MVE} shown in Figs. 15 and 16. MVE is defined as the standard deviation of predictions at the same pixel location over all images in the test set. MVE can thus be considered a proxy of $\mathcal{L}^{\text{CON}}$ which measures the deviation of same-location predictions. The ratio

$$\mathrm{c}_{\text{MVE}} = \frac{\mathrm{std}(\hat{\mathbf{t}})}{\langle \ell_1(\hat{\mathbf{t}}, \mathbf{t}) \rangle} \tag{16}$$

is a proxy for the importance of perturbation in units of MAE. Here the mean $\langle \cdot \rangle$ and standard deviation are applied over all predicted pixels covering the same location in the time series, $\hat{\mathbf{t}}$ is the prediction and $\mathbf{t}$ the ground truth.

By observing MVE the impact of perturbations are singled out while the forest structure is kept constant. Hence, c_{MVE} can be understood to quantify the predictor's robustness to these perturbations and as a proxy for the fraction of MAE due to perturbative effects. While c_{MVE} does not quantify this fraction exactly, it allows a qualitative assessment of its magnitude for different loss configurations.

4.2 Datasets

The gathered data sets were assembled in different data configurations in order to assess the impact of the loss terms $\mathcal{L}^{\text{CON}}$ and $\mathcal{L}^{\text{COD}}$ for prediction. Two experiments were conducted on Sentinel-2 data and one smaller experiment on hyperspectral APEX data.

The present work does not include a thorough grid search over the training parameters and the loss weights λ_{CON} and λ_{COD}. The parameter values are reported in Tab. 4.

All data sets were split in a training, validation and test part. During training of the predictor models, the training data set was used for backpropagation and the validation data set to monitor the training progress. The evaluation presented in the Results section is based on the model predictions on the test set. As can be seen in Fig. 1 care was taken to separate geographically the test set. This could not be done for the APEX data set as can be observed in Fig. 2.

4.2.1 Sentinel-2 1c: Single Image (SINGLE)

To evaluate the model on single cloud free images, an experiment was conducted on SEN1c leaf-on images from Aargau and Fribourg. The mean time series length per image was 31 (compare to Tab. 1). This length varies as a function of the atmospheric conditions over the covered time span.

> **Loss Configurations:** Runs with base line ℓ_1, (L1), semi-supervised $\ell_1 + \lambda_{\text{COD}}\,\mathcal{L}^{\text{COD}}$ (L1 + ADV) and $\ell_1 + \lambda_{\text{COD}}\,\mathcal{L}^{\text{COD}} + \lambda_{\text{CON}}\,\mathcal{L}^{\text{CON}}$ (L1 + ADV + CON) loss were evaluated. All variations of $\mathcal{L}^{\text{CON}}$ were included.
>
> **Data Sets:** SEN1c images were combined with the ground truth data sets AAR and FRI to form a data set $\mathcal{D}_c$ of corresponding images. Non-corresponding data sets $\mathcal{X}_{nc}$ and $\mathcal{Y}_{nc}$ were constructed for $\mathcal{L}^{\text{COD}}$ from the same SEN1c images by randomly sampling in either of the LiDAR and spectral data sets. A data set $\mathcal{D}^{T}_{inv}$ for $\mathcal{L}^{\text{CON}}$ was constructed by randomly sampling and grouping pairs of images from the time series of spectral images corresponding to the same ground truth image. Only spectral images from the months Mai to August were used. An overview over the dataset sizes can be found in Tab. 1. As can be seen in Fig. 1, images from Aargau were split in training and validation set, while images from Fribourg were used exclusively as test data set.
>
> **Normalization:** Both input and target were subjected to a gaussian normalization before use with the model. This normalization was derived from the training set.

In preliminary tests, a substantial improvement could be noted when applying CON losses only over the percentiles p20 - p95 and not over VCI and COV. This possibly is due to high temporal variability of VCI and COV. In all presented runs VCI and COV are therefore excluded from CON.

4.2.2 Sentinel-2 1c: Leaf-On/Leaf-Off Image (ON-OFF)

Phenological changes can inform the predictor network of forest structure. Instead of basing the prediction on single spectral images during leaf-on conditions, individual samples during leaf-on and leaf-off conditions were stacked. This experiment tests whether the presented loss configurations on stacked images yield comparable prediction properties to SINGLE runs.

> **Data Sets**: Images from acquisitions in the months Mai to August were considered leaf-on and images from November to February leaf-off. The same procedure as for SINGLE was followed to assemble the leaf-on part of $\mathcal{D}_c$, $\mathcal{X}_{nc}$, $\mathcal{Y}_{nc}$ and $\mathcal{D}^T_{inv}$ data sets. The leaf-off counterpart was randomly sampled from leaf-off images at the same location. Due to more frequent exclusion of leaf-off images (disadvantageous cloud conditions in autumn and winter) not all locations could be used, which explains the smaller number of training samples (see Tab. 1).
>
> **Loss Configurations:** Runs with base line ℓ_1, (L1), semi-supervised $\ell_1 + \lambda_{\text{COD}}\,\mathcal{L}^{\text{COD}}$ (L1 + ADV) and $\ell_1 + \lambda_{\text{COD}}\,\mathcal{L}^{\text{COD}} + \lambda_{\text{CON}}\,\mathcal{L}^{\text{CON}}$ (L1 + ADV + CON$_2$) loss were evaluated. Only CON$_2$ was evaluated as $\mathcal{L}^{\text{CON}}$.
>
> **Comparison to SINGLE:** As discussed above, the model architecture was changed in order to account for the differential nature of the adapted input. Both input and target were subjected to a separate gaussian normalization before use with the model. The validation data sets of SINGLE and ON-OFF were aligned along the leaf-on dates to assure a consistent comparison, i.e. the data sets of both consisted of the same leaf-on images with corresponding ON-OFF leaf-off images being drawn randomly. Conclusions deriving from this comparison must be taken with care as **i)** the training data sets differ and **ii)** the model architecture was changed in order to accommodate for the different input shape. Both affect the optimization. Hence, performance comparison

> between SINGLE and ON-OFF cannot differentiate with certainty between data and architecture caused changes.
>
> The additional leaf-off images arguably introduce new sources of noise and perturbation. By restricting the data sets to the same leaf-on images, the same perturbing effects yielding the spread in SINGLE input domain are present in the ON-OFF case. Arguably, ON-OFF inputs additionally also exhibit non-predictive variation in the leaf-off part of the input, including possibly snowy images.

4.2.3 APEX

The APEX experiment evaluates the prediction performance of percentiles p20, p50, p70 and p95 from APEX TOC images. The corresponding data sets $\mathcal{D}_c$ and and $\mathcal{X}_{nc}$ are substantially smaller than $\mathcal{D}_c$ in the SINGLE and ON-OFF experiments. On the other hand, a large amount of forest structure information $\mathcal{Y}_{nc}$ without spectral counterpart is available. The aim is to evaluate whether the semi-supervised adversarial setting is a useful tool to integrate non-corresponding data sets for improved pixel-wise and distributional performance.

The application of $\mathcal{L}^{\text{COD}}$ on spectral data with increased resolution in the spatial domain is interesting because it can be expected that $\mathcal{L}^{\text{COD}}$ improves the high frequency part of the prediction. In SINGLE and ON-OFF the image frequency spectrum is restricted by the low spatial resolution. Even the highest frequencies in Sentinel images correspond to large forest structures when compared to the single tree level. The application of $\mathcal{L}^{\text{COD}}$ could prove particularly beneficial if applied to images with high spatial resolution (i.e. tree level frequencies) since the causal background of the high frequencies with respect is arguably less confounded by mixing.

It is unclear if higher resolution in the spectral domain is beneficial in the context of structural canopy retrieval. Theoretically, the higher spectral resolutions could have benefits.

> **Species Distribution**: Biochemical features in the reflectance spectrum can be matched with more accuracy and thus increase the predictors sensitivity to target distributions that correlate well with species distributions.

Canopy Penetration 3D structure retrieval especially at low and sparse canopy levels is expected to be in part based on the light penetration through the canopy and backscatter from ground surfaces. The unmixing of canopy and and ground signal is improved at higher spectral resolution in linear mixing theory. Similarly, an increased discriminativeness of the predictor could be the achieved.

No extensive research has been performed so far to evaluate these factors. Halme et al. [89] finds that higher spectral resolution benefits mainly variables that are strongly correlated to species differences. Since it is unclear to what extent the prediction of hidden vertical canopy structure is driven by species differences in the spectral signal, it is not clear if [89] applies specifically to the present inversion problem.

Loss Configurations: Since the data lacked enough overlapping flight lines to construct $\mathcal{D}_{inv}^{T}$, no $\mathcal{L}^{\text{CON}}$ was included in this experiment. Runs with a base line ℓ_1, (L1) and semi-supervised $\ell_1 + \lambda_{\text{COD}}\,\mathcal{L}^{\text{COD}}$ (L1 + ADV) losses were performed.

Data Sets: $\mathcal{D}_c$ was constructed by pairing APEX images with corresponding ground truth from the AAR data set. Since $\mathcal{D}_c$ is significantly smaller than in the SINGLE and ON-OFF case due to lacking APEX imagery, the possibility to use $\mathcal{L}^{\text{COD}}$ on data with spatio-temporal mismatch is particularly interesting. $\mathcal{X}_{nc}$ and $\mathcal{Y}_{nc}$ were constructed so as to include images without spectral or co-domain counter part as well. Specifically, the whole ground truth coverage of FRI and AAR were included, as well as some ECOTRANS section outside of the FRI and AAR coverage as can be seen in Fig. 2.

Normalization: The data set $\mathcal{D}_c$ with the APEX acquisitions was preceded by a PCA transform of the cropped APEX images in the training set $\mathcal{X}_{nc}$. The PCA was fitted on the individual pixels of all images in the training set and restricted to the first 30 components.

4.3 Training

Optimization of both predictor and discriminator weights was performed with the the ADAM optimization as defined in Kingma and Ba [90] with fixed hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The training was performed until no further improvement of the MAE in p95 could be observed on the validation part of $\mathcal{D}_c$. During training, the discriminator loss needed to stay stable. If this was not the case, the experiment was performed with a different set of learning rates for predictor and discriminator. See Tab. 4 for the parameters defining the learning rates and loss configuration for each run.

SINGLE	$l_r(N_\theta)$	$l_r(D_\phi)$	λ_{COD}	λ_{CON}
L1	5×10^{-4}	–	–	–
L1 + ADV	5×10^{-4}	5×10^{-6}	1	–
L1 + ADV + CON$_1$	5×10^{-4}	5×10^{-6}	1	0.3
L1 + ADV + CON$_2$	5×10^{-4}	5×10^{-6}	1	0.016
L1 + ADV + CON$_3$	5×10^{-4}	5×10^{-6}	1	0.016
ON-OFF				
L1	5×10^{-4}	–	–	
L1 + ADV	5×10^{-4}	5×10^{-6}	1	–
L1 + ADV + CON$_2$	5×10^{-4}	5×10^{-6}	1	0.16
APEX				
L1	5×10^{-4}	5×10^{-6}	–	–
L1 + ADV	5×10^{-3}	5×10^{-5}	1	–

Table 4: Parameters of experiments. l_r ednotes the step size of the ADAM optimizers of predictor and discriminator networks. λ's denote the weights of $\mathcal{L}^{\text{CON}}$ and $\mathcal{L}^{\text{COD}}$.

4.4 Pixelwise Validation

4.4.1 Height Percentiles (SINGLE and ON-OFF)

Mean absolute errors (MAE) between prediction and ground truth were calculated for all pixels in the test set. Fig. 14 summarizes the results for the MAE percentile results and Tab. 6 presents the corresponding R^2 scores. The predictions trained with the augmented losses didn't outperform the base line L1 experiments, neither in the SINGLE nor in the ON-OFF case. This is true

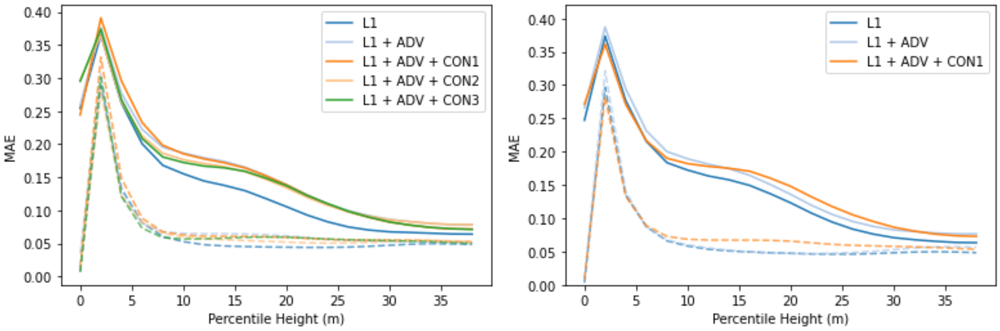

Figure 9: Mean Absolute Error (solid) and Median Absolute Error (dashed) of VCI stratified over true p95 (left: SINGLE, right: ON-OFF). The reduction was performed in 2m bins.

for all considered percentiles. MAE of the L1 + ADV run in the SINGLE experiment outperforms slightly all other CON runs. This is different to the ON-OFF experiment where the CON$_2$ loss outperforms the ADV run in all percentiles but p20.

ON-OFF			**SINGLE**	
	COV	VCI	COV	VCI
L1	0.15	0.12	0.15	0.11
L1 + ADV	0.17	0.13	0.16	0.14
L1 + ADV + CON$_1$	–	–	0.16	0.13
L1 + ADV + CON$_2$	0.17	0.14	0.16	0.14
L1 + ADV + CON$_3$	–	–	0.16	0.14

Table 5: Global MAE of SINGLE and ON-OFF runs for VCI and COV

Figs. 11 and 13 show the MAE and Median Absolute Error (MedAE) stratified over the true height (p95) for all predicted percentiles. SINGLE and ON-OFF experiments show a similar behaviour over the height range. For the percentiles p50 - p95 the baseline L1 runs show minima around 0 - 5 m and 25 - 30 m in MAE.

The minimum at 0 m is trivially explained by the good distinction of non-forested pixels. The minimum at 25 - 30 m coincides with the maximum of

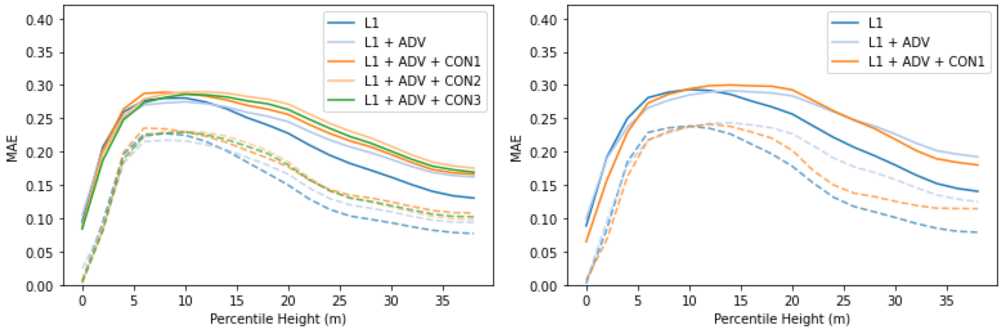

Figure 10: Mean Absolute Error (solid) and Median Absolute Error (dashed) of COV stratified over true p95 left: SINGLE, right: ON-OFF). The reduction was performed in 2m bins.

the empirical distribution of the validation set as can be checked for p95 in Fig. 21.

The ADV run flattens the second minimum in both SINGLE and ON-OFF and is characterized by a rise in MAE (MedAE) at lower heights than in the baseline. The L1 + ADV runs outperform the baselines at heights $\lesssim 20$ m. The same is true for runs with CON_2 and CON_3 losses in SINGLE but not for CON_2 loss on ON-OFF.

4.4.2 COV and VCI

As for the prediction performance on percentiles, the baseline L1 runs outperform the runs with augmented losses (see Tab. 5) for COV and VCI. Both COV and VCI predictions perform significantly worse than the percentile predictions when comparing MAE to the empirical standard deviation. Moreover, contrary to the percentile predictions, the inclusion of $\mathcal{L}^{\text{CON}}$ doesn't improve MAE of the adversarial run in any experiment globally. Only certain height sections are marginally improved as is visible in Fig. 10.

The large difference between MedAE and MAE for VCI and COV predictions observable in Fig. 9 shows that these predictions are heavily impacted by outliers. The empirical COV and VCI distributions have two isolated peaks around 0 and 1. Large outliers can easily occur when prediction is located in the wrong peak.

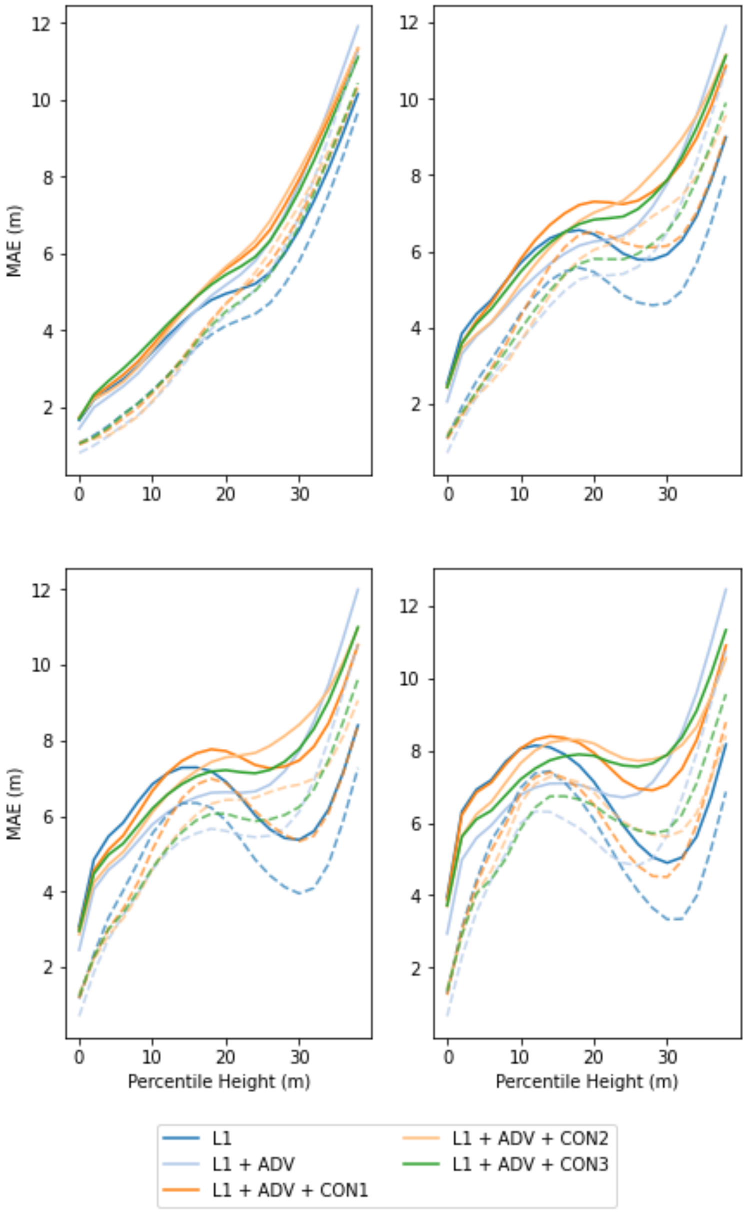

Figure 11: **SINGLE**. Mean Absolute Error (solid) and Median Absolute Error (dashed) stratified over true p95. The reduction was performed in 2m bins. Shown are results for predicted percentiles p20, p50, p70 and p95 (top-bottom, left-right).

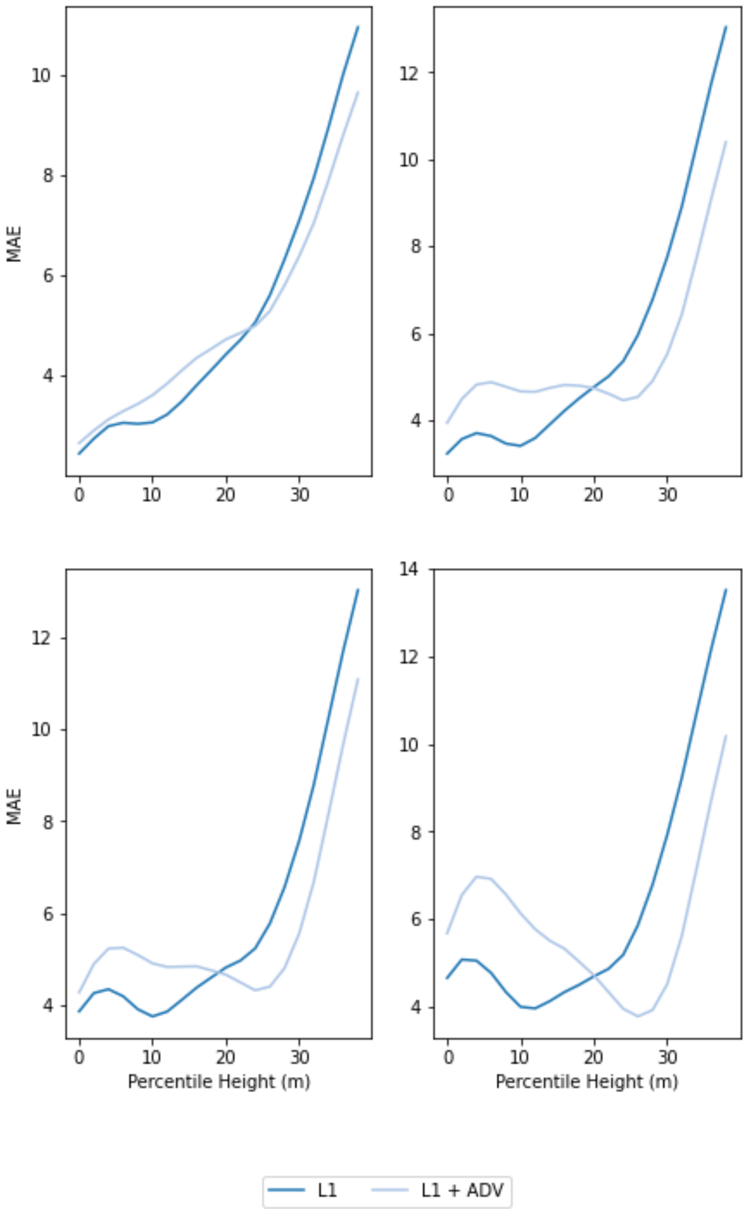

Figure 12: **APEX**. Mean Absolute Error (solid) over true p95. The reduction was performed in 2m bins. Shown are results for predicted percentiles p20, p50, p70 and p95 (top-bottom, left-right).

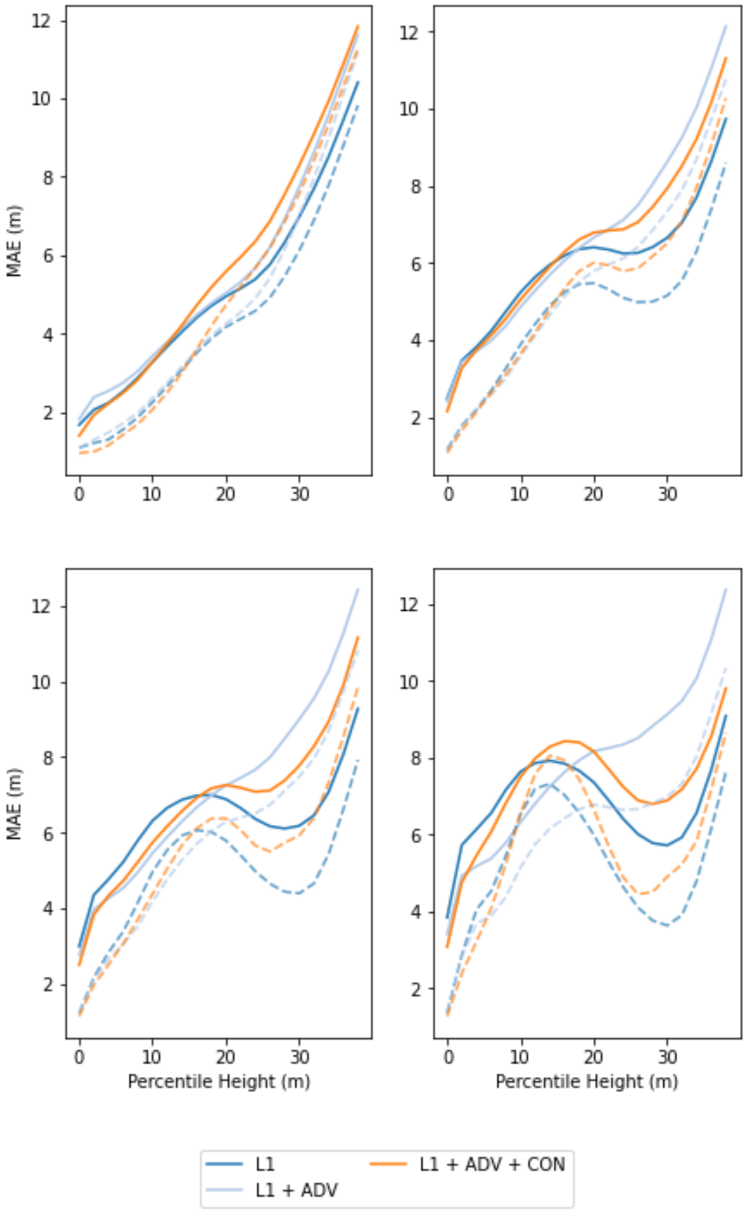

Figure 13: **ON-OFF**. Mean Absolute Error (solid) and Median Absolute Error (dashed) stratified over true p95. The reduction was performed in 2m bins. Shown are results for predicted percentiles p20, p50, p70 and p95 (top-bottom, left-right).

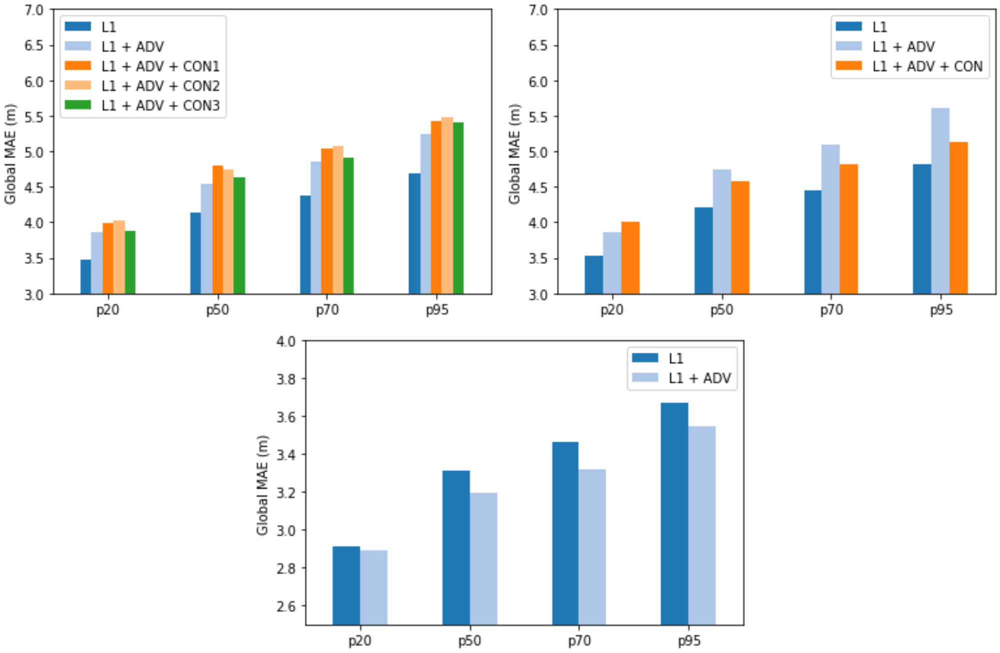

Figure 14: Global MAE of SINGLE (left) and ON-OFF (right) and APEX (bottom) runs

Globally, the proposed losses $\mathcal{L}^{\text{CON}}$ and $\mathcal{L}^{\text{COD}}$ were not useful to enhance pixel-wise MAE and R^2 of COV and VCI.

4.4.3 APEX

Fig. 14 and Fig. 12 show pixel-wise prediction performance of the APEX baseline and augmented run. The baseline run L1 can be seen to be outperformed globally by L1 + ADV for all percentiles. This is in contrast to the experiments SINGLE and ON-OFF.

Looking at MAE as a function of height in Fig. 12, it can be noted that L1 + ADV improves predictions in the height range $h \gtrsim 20$ m. At heights $h \lesssim 20$ m, the ADV run's performance in terms of MAE is worse than the baseline. As the observation above, this behaviour is qualitatively different to SINGLE and ON-OFF. Since the two runs were trained on data sets with large size differences it is not clear whether the specific loss configurations or the different data sizes are responsible for the prediction differences.

4.5 Spatial Structure Validation

4.5.1 Results: KL and JS

The results for the KL and JS evaluation on the training sets of SINGLE and ON-OFF are summarized in Tabs. 7 and 11. The evaluation shows that in the SINGLE configuration, CON runs outperform both the baseline run L1 and the adversarial run L1 + ADV. In ON-OFF the CON$_2$ run does not improve spatial distribution measures KL and JS. This is true at all scales, for all target variables and observable both in JS and KL.

This result is in line with the visual impression of the sample predictions in Figs. 6, 7, 25 and 26. It is noteworthy to point out that this result is opposite to the pixel-wise performance order. The adoption of an adversarial setting along with the physical regularization in $\mathcal{L}^{\text{CON}}$ thus improves regional distri-

R^2

SINGLE	p20	p50	p70	p95	VCI	COV
L1	0.28	0.55	0.61	0.67	0.64	0.63
L1 + ADV	-0.04	0.33	0.41	0.50	0.53	0.56
L1 + ADV + CON$_1$	0.10	0.38	0.46	0.52	0.56	0.54
L1 + ADV + CON$_2$	0.08	0.35	0.43	0.53	0.51	0.52
L1 + ADV + CON$_3$	0.12	0.38	0.45	0.51	0.51	0.54
ON-OFF						
L1	0.21	0.48	0.55	0.62	0.60	0.57
L1 + ADV	-0.03	0.26	0.33	0.38	0.55	0.45
L1 + ADV + CON$_2$	-0.01	0.37	0.45	0.58	0.53	0.48
APEX						
L1	0.32	0.47	0.56	0.61	–	–
L1 + ADV	0.49	0.63	0.67	0.72	–	–

Table 6: R^2 scores for all experiments.

bution approximation at all scales even though the pixel-wise performance decreases.

There are differences in the prediction performance between the different CON losses in SINGLE. Results for KL and JS at the image scale (30 × 30) suggest that CON3 performed best on percentile predictions. At smaller scales, this statement remains true for p70 and p95. For p20 and p50 the orderings between CON losses differ in KL and JS, so that no conclusions can be made for those cases.

In the APEX runs, in the inclusion ADV can be observed to improve KL and JS on percentiles p20 and p95 on all scales. Performance p50 and p95 reveals the ADV to perform slightly worse. While it is intuitively explainable why the uppermost canopy layer's target distribution is best mapped, the connection to performance on lower layers is unclear.

4.5.2 Results: Pearson Correlation

The results for the PC evaluation are shown in Tabs. 12 and 13. Contrary to the results for KL and JS, the baseline L1 runs are not outperformed by any augmented loss configuration. This statement applies to all target variables and both evaluated experiments SINGLE and ON-OFF. The baseline runs consistently show higher correlation between ground truth and prediction windows. While in the ON-OFF experiment the inclusion of $\mathcal{L}^{\mathrm{CON}}$ can recover the decay of PC on p95, this is not the case in SINGLE and for any other target variable in ON-OFF.

4.6 Joint Distribution Matching

Tab. 14 shows the mean and median cross-entropy per image of SINGLE, ON-OFF and APEX configurations.

The cross-entropy is reduced in all augmented SINGLE runs. This is true for the mean as well as the median. In order to test for a consistent reduction of the median across all predictions, a binomial sign test was performed between each augmented run and the base line. As can be seen in Fig. 17 the use of the more powerful Wilcoxon signed-rank test not possible due to the asymmetry of the CE distribution (only one SINGLE L1 is shown, the symmetry is consistent across all runs). Notably, the sign test doesn't

JS SINGLE

30 × 30	p20	p50	p70	p95	VCI	COV
L1	0.10	0.10	0.10	0.10	0.15	0.54
L1 + ADV	0.14	0.13	0.13	0.14	0.20	0.55
L1 + ADV + CON$_1$	0.08	0.08	0.08	0.09	0.11	0.54
L1 + ADV + CON$_2$	0.09	0.09	0.10	0.10	0.12	0.54
L1 + ADV + CON$_3$	0.08	0.07	0.08	0.08	0.13	0.52
15 × 15						
L1	0.17	0.17	0.17	0.18	0.22	0.59
L1 + ADV	0.18	0.19	0.20	0.21	0.25	0.61
L1 + ADV + CON$_1$	0.14	0.14	0.15	0.16	0.17	0.58
L1 + ADV + CON$_2$	0.14	0.15	0.16	0.16	0.17	0.58
L1 + ADV + CON$_3$	0.13	0.13	0.14	0.14	0.18	0.56
7 × 7						
L1	0.24	0.24	0.25	0.26	0.27	0.53
L1 + ADV	0.23	0.24	0.25	0.27	0.29	0.55
L1 + ADV + CON$_1$	0.21	0.22	0.23	0.24	0.22	0.48
L1 + ADV + CON$_2$	0.20	0.21	0.22	0.23	0.23	0.48
L1 + ADV + CON$_3$	0.20	0.21	0.22	0.22	0.23	0.46

Table 7: Mean Jensen-Shannon Divergence for all SINGLE runs calculated over window sizes 30 × 30, 15 × 15 and 7 × 7 for all targets.

require the data to be symmetrical around its median. In order to assure sample independence in the test, single predictions were randomly sampled from locations with multiple predictions. The reduction of the median CE of the augmented loss runs over the baseline in SINGLE was found to be statistically significant with vanishing p-value.

The ON-OFF L1 baseline is outperformed by L1 + ADV + CON2, but not by L1 + ADV, exactly as for the KL divergence. The same binomial sign test as above was applied to the the CE differences between L1 and L1 + ADV. As above the reduction of CE in L1 + ADV + CON proves to be statistically significant under a binomial sign test with vanishing p-value.

The augmented APEX run outperforms the baseline as in SINGLE and ON-OFF. Under the same binomial test as above, the reduction is statistically significant with vanishing p-value.

4.7 Prediction Variability

In order to assess the impact of perturbations on prediction accuracy, MVE and c_{MVE} are computed. Figs. 15 and 16 show MVE and c_{MVE} for SINGLE and ON-OFF runs stratified over canopy height.

The MVE of baseline L1 runs in ON-OFF and SINGLE configuration are very similar. Different behaviour can be observed for the runs with augmented losses. MVE is expected to be smaller in L1 + ADV + CON runs than in L1 + ADV runs as $\mathcal{L}^{\mathrm{CON}}$ explicitly punishes prediction differences under perturbed input. From Figs. 15 and 16 we see this to hold mostly for p50 - p95 in the ON-OFF configuration. In the SINGLE case, $\mathcal{L}^{\mathrm{CON}}$ doesn't reduce MVE under the levels of L1 + ADV. This result is comparable to the performance ordering in MAE where we see a positive impact of $\mathcal{L}^{\mathrm{CON}}$

JS ON-OFF

30 × 30	p20	p50	p70	p95	VCI	COV
L1	0.08	0.07	0.07	0.07	0.11	0.53
L1 + ADV	0.07	0.06	0.07	0.08	0.09	0.51
L1 + ADV + CON$_2$	0.09	0.08	0.08	0.08	0.13	0.55
15 × 15						
L1	0.14	0.14	0.14	0.14	0.17	0.57
L1 + ADV	0.13	0.11	0.13	0.14	0.14	0.55
L1 + ADV + CON$_2$	0.15	0.14	0.14	0.14	0.18	0.59
7 × 7						
L1	0.21	0.22	0.22	0.22	0.24	0.48
L1 + ADV	0.19	0.18	0.19	0.21	0.21	0.46
L1 + ADV + CON$_2$	0.22	0.22	0.22	0.22	0.23	0.50

Table 8: Mean Jensen-Shannon Divergence per image for all ON-OFF runs calculated over window sizes 30 × 30, 15 × 15 and 7 × 7 for all targets.

KL SINGLE						
30 × 30	p20	p50	p70	p95	VCI	COV
L1	0.50	0.54	0.54	0.59	1.11	2.56
L1 + ADV	0.97	0.92	0.87	0.93	1.43	2.83
L1 + ADV + CON$_1$	0.50	0.41	0.41	0.44	0.78	2.10
L1 + ADV + CON$_2$	0.46	0.44	0.46	0.43	0.74	1.97
L1 + ADV + CON$_3$	0.34	0.32	0.34	0.35	0.89	1.77
15 × 15						
L1	2.04	2.14	2.19	2.27	2.80	5.57
L1 + ADV	2.34	2.34	2.35	2.31	3.10	5.70
L1 + ADV + CON$_1$	1.65	1.46	1.47	1.49	1.98	2.88
L1 + ADV + CON$_2$	1.02	1.04	1.13	1.11	1.49	2.81
L1 + ADV + CON$_3$	1.14	1.01	1.02	0.77	2.04	2.84
7 × 7						
L1	3.85	4.05	4.16	4.30	4.12	7.64
L1 + ADV	4.15	4.32	4.42	4.34	5.09	8.24
L1 + ADV + CON$_1$	3.32	3.12	3.36	3.34	3.16	2.95
L1 + ADV + CON$_2$	2.25	2.28	2.69	2.60	2.66	3.80
L1 + ADV + CON$_3$	2.52	2.37	2.44	1.61	3.28	3.54

Table 9: Mean Kullback-Leibler Divergence per image for all SINGLE runs calculated over window sizes 30 × 30, 15 × 15 and 7 × 7 for all targets.

KL ON-OFF

30 × 30	p20	p50	p70	p95	VCI	COV
L1	0.31	0.31	0.31	0.32	0.72	1.87
L1 + ADV	0.30	0.28	0.31	0.33	0.60	1.58
L1 + ADV + CON$_2$	0.42	0.32	0.35	0.34	0.53	2.21
15 × 15						
L1	0.94	1.54	1.40	1.46	1.96	3.50
L1 + ADV	0.71	0.64	0.68	0.77	1.28	2.06
L1 + ADV + CON$_2$	1.38	1.33	1.29	0.81	1.69	3.85
7 × 7						
L1	1.66	2.53	2.48	2.51	2.70	4.39
L1 + ADV	1.16	1.13	1.13	1.25	1.83	2.45
L1 + ADV + CON$_2$	2.26	2.26	2.32	1.55	2.59	4.28

Table 10: Mean Kullback-Leibler Divergence per image for all ON-OFF runs calculated over window sizes 30 × 30, 15 × 15 and 7 × 7 for all targets.

KL APEX					**JS APEX**			
150 × 150	p20	p50	p70	p95	p20	p50	p70	p95
L1	2.31	2.46	2.75	2.26	0.26	0.25	0.27	0.31
L1 + ADV	1.41	2.61	3.06	1.87	0.23	0.25	0.28	0.29
75 × 75								
L1	3.12	3.00	3.60	2.98	0.30	0.28	0.32	0.34
L1 + ADV	1.81	3.19	4.01	2.48	0.27	0.28	0.32	0.33
30 × 30								
L1	3.70	3.30	4.07	3.64	0.34	0.31	0.35	0.38
L1 + ADV	2.03	3.44	4.54	2.89	0.32	0.31	0.35	0.37

Table 11: Mean Kullback-Leibler and Jensen-Shannon Divergence per image for APEX runs calculated over window sizes 150 × 150, 75 × 75 and 30 × 30 for all targets.

PC SINGLE

$\sigma = 10$	p20	p50	p70	p95	vci	cov
L1	0.59	0.64	0.66	0.68	0.60	0.64
L1 + ADV	0.51	0.58	0.61	0.64	0.53	0.60
L1 + ADV + CON$_1$	0.50	0.57	0.59	0.62	0.57	0.61
L1 + ADV + CON$_2$	0.49	0.56	0.58	0.61	0.53	0.58
L1 + ADV + CON$_3$	0.50	0.56	0.59	0.61	0.52	0.59
$\sigma = 5$						
L1	0.53	0.58	0.59	0.61	0.52	0.56
L1 + ADV	0.45	0.52	0.54	0.56	0.44	0.52
L1 + ADV + CON$_1$	0.43	0.50	0.52	0.54	0.48	0.52
L1 + ADV + CON$_2$	0.43	0.49	0.52	0.54	0.44	0.50
L1 + ADV + CON$_3$	0.44	0.50	0.52	0.54	0.44	0.51
$\sigma = 2$						
L1	0.39	0.42	0.43	0.45	0.40	0.39
L1 + ADV	0.32	0.36	0.38	0.40	0.31	0.36
L1 + ADV + CON$_1$	0.30	0.34	0.36	0.38	0.35	0.35
L1 + ADV + CON$_2$	0.30	0.35	0.36	0.38	0.32	0.34
L1 + ADV + CON$_3$	0.30	0.34	0.36	0.38	0.32	0.34

Table 12: Overview over mean PC computed as outlined in Section 4.1.2.

in ON-OFF runs but not in SINGLE. The baseline L1 runs of ON-OFF and SINGLE are comparable in terms of c_{MVE}. Both show a minimum at 0 m, a plateau from 0 - 20 m, a distinct peak around 30 m and a fast decay of c_{MVE} for the extreme height range above 30 m. Considering the targets p70 and p95, a general feature across all augmented runs in SINGLE is the reduction of the peak height and smaller variance around the mean value.

PC ON-OFF

$\sigma = 10$	p20	p50	p70	p95	vci	cov
L1	0.56	0.61	0.64	0.66	0.60	0.62
L1 + ADV	0.50	0.55	0.58	0.62	0.56	0.58
L1 + ADV + CON$_2$	0.48	0.58	0.61	0.65	0.56	0.59
$\sigma = 5$						
L1	0.49	0.53	0.55	0.57	0.51	0.52
L1 + ADV	0.42	0.47	0.49	0.53	0.47	0.49
L1 + ADV + CON$_2$	0.41	0.50	0.53	0.57	0.46	0.49
$\sigma = 2$						
L1	0.32	0.35	0.37	0.38	0.37	0.33
L1 + ADV	0.26	0.29	0.31	0.34	0.33	0.31
L1 + ADV + CON$_2$	0.26	0.33	0.35	0.38	0.33	0.29

Table 13: Overview over mean PC computed as outlined in Section 4.1.2.

CE

	SINGLE		**ON-OFF**		**APEX**	
	mean	**median**	**mean**	**median**	**mean**	**median**
L1	1.53	1.61	1.51	1.60	1.30	1.29
L1 + ADV	1.48	1.55	1.53	1.61	1.27	1.27
L1 + ADV + CON$_1$	1.49	1.57	–	–	–	–
L1 + ADV + CON$_2$	1.51	1.58	1.48	1.56	–	–
L1 + ADV + CON$_3$	1.52	1.60	–	–	–	–

Table 14: Mean and Median CE per image for SINGLE and ON-OFF runs. Listed are results for $n_c = 6$ (SINGLE / ON-OFF) and $n_c = 7$ (APEX).

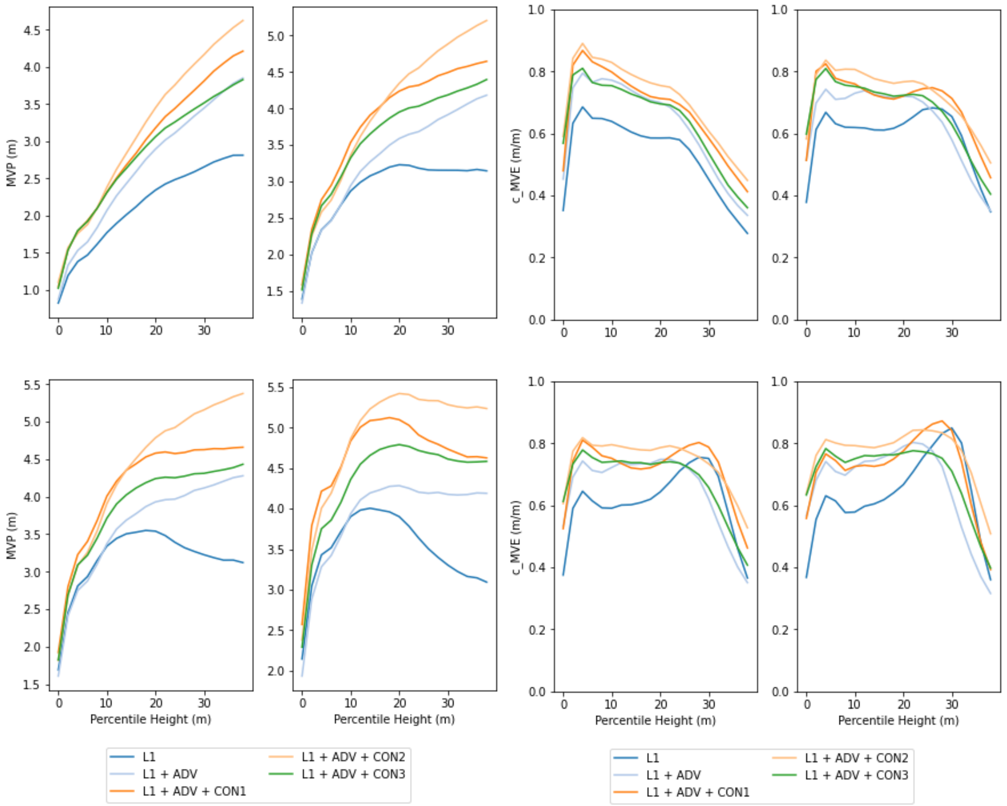

Figure 15: **SINGLE**. *Left*: Mean Variation Error stratified over true p95. The reduction was performed in 2m bins. Shown are results for predicted percentiles p20, p50, p70 and p95 (top-bottom, left-right). *Right:* c_{MVE} stratified over true p95.

5 Discussion

5.1 Results Sentinel Experiments

While SINGLE and ON-OFF augmented runs improved distributional approximation, they performed worse in terms of absolute regression errors as measured by MAE and R^2. The stratification of absolute errors over the true canopy height as shown in Figs. 11 and 13 reveils that the augmented runs perform worse in particular in the height range ∼ 30m where the baseline runs have a characteristic minimum. The ADV runs' minima in this region

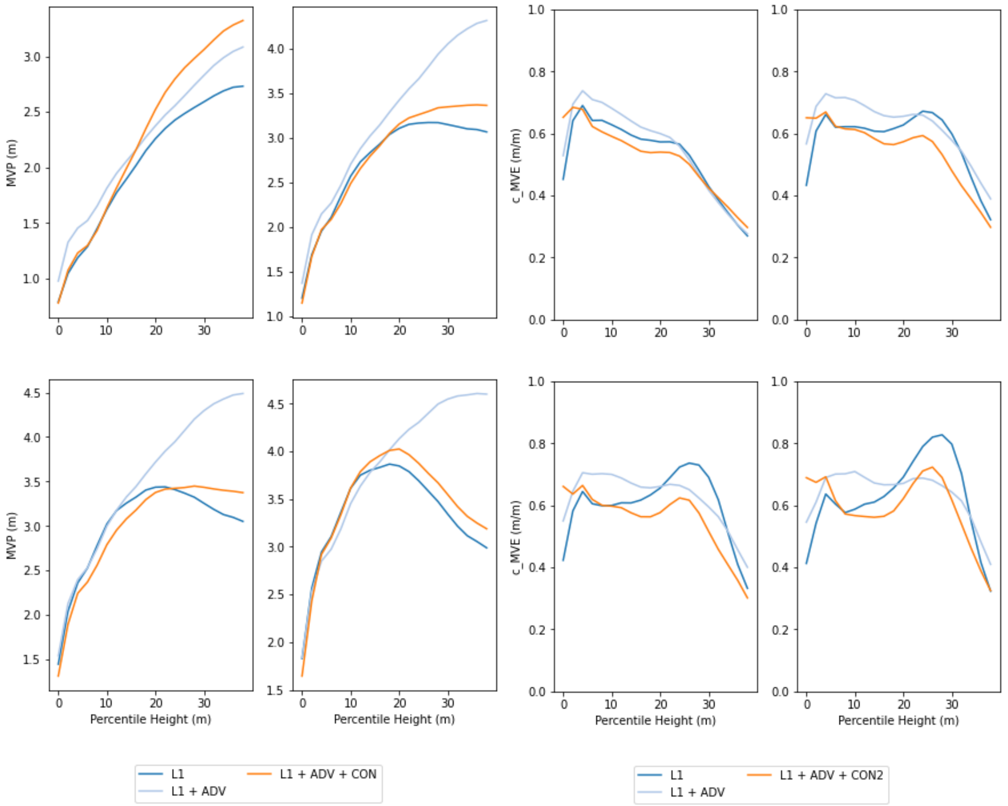

Figure 16: **ON-OFF**. *Left*: Mean Variation Error stratified over true p95. The reduction was performed in 2m bins. Shown are results for predicted percentiles p20, p50, p70 and p95 (top-bottom, left-right). *Right:* c_{MVE} stratified over true p95.

are levelled and CON runs expose a reduced minimum. Constraining the following argument to p95, it can be noted in Fig. 21 that $\sim$ 30m corresponds the maximum of the empirical training distribution. This minimum in the baseline run's MAE could therefore be due to a bias in the training set rather than the canopy at heights $h \sim 20$ m exposing less predictive features. If the minimum is due to training sample bias, the reduction of the minimum indicates a reduction of the impact that the biased data set has on the prediction accuracy with the adversarial setting.

A possible explanation for this is the fact that $\mathcal{L}^{\mathrm{CON}}$ and $\mathcal{L}^{\mathrm{COD}}$ depend on absolute height non-trivially and arguably only weakly (through height-dependent shadowing, specific vegetation, light penetration to ground, etc.). It is expected that predictive features mainly relate to relative target variation, while the absolute height, that is lost in the perspective projection to 2D, is only weakly present in species related features and shadowing. The absolute height is thus especially impacted by bias in the training data set. The height bias present in the baseline L1 run may therefore be reduced with a larger weight of the loss components $\mathcal{L}^{\mathrm{COD}}$ ($\mathcal{L}^{\mathrm{CON}}$) that do not depend (reduce the dependency) on absolute errors. Other than affecting the minima, bias reduction could explain as well the lower MAE of ADV runs at heights $h \lesssim 20$ m and higher MAE at h $\gtrsim 20$ m.

The correlations between both VCI and COV and the rest of the target space distribution were expected to be the most relevant predictors for COV and VCI prediction. These implicit soft constraints did not successfully improve the prediction within forested areas. On the one hand, the information content in the optical image might be insufficient, on the other it must be taken into account that VCI and COV are expected to vary more than the percentiles across acquisition times and geometries, i.e. label-noise is expected to be larger.

5.1.1 Spatial Distribution

The two losses $\mathcal{L}^{\mathrm{CON}}$ and $\mathcal{L}^{\mathrm{COD}}$ could not be shown to improve either of MAE and PC over the baseline runs. Similarly, it appears from Figs. 6 and 7 that models trained on the augmented losses improve the overall texture rather than the absolute values per pixel. This qualitative statement is corroborated by the results of JS and MAE. While MAE increases in all augmented loss runs, there is always a combination of $\mathcal{L}^{\mathrm{CON}}$ and $\mathcal{L}^{\mathrm{COD}}$ that reduces the JS divergence. Hence, the experimental results of SINGLE and ON-OFF suggest that the adoption of an adversarial setting can improve the spatial distribution approximation.

Similarly to MAE, PC could not be improved in any ON-OFF or SINGLE run. Since PC as well as JS measure relative target variations, a possible explanation for the inverted ordering in performance is the location dependency in σ_{xy}. The location dependency in JS is only weak, i.e. the neighbourhood defined by the window is the only dependency. Following this argument, the adversarial setting ($\mathcal{L}^{\text{COD}}$) reduces the impact of the locally explicit error terms (ℓ_1 and $\mathcal{L}^{\text{CON}}$) and enhances regional distribution approximation at the expense of locally correct estimates. The local and non-local parts of the total loss could not be reduced at the same time.

5.1.2 Weighing $\mathcal{L}^{\text{CON}}$ and $\mathcal{L}^{\text{COD}}$

Whether an augmented run indeed reduces JS divergence depends on the interplay between $\mathcal{L}^{\text{CON}}$ and $\mathcal{L}^{\text{COD}}$ and differs in SINGLE and ON-OFF. For example, the comparison of L1 + ADV + CON with L1 + ADV in the ON-OFF case suggests that the use of a consistency loss $\mathcal{L}^{\text{CON}}$ along with the adversarial loss $\mathcal{L}^{\text{COD}}$ is beneficial in terms of MAE and PC. This behaviour cannot be observed in the SINGLE runs where no CON configuration outperforms the L1 + ADV run. On the contrary, in SINGLE the inclusion of CON benefits the reduction of KL divergence and worsens locality-aware metrics MAE and PC.

It can be argued that the invariance loss CON effectively acts as regularizor to constrain the adversarial minimax game to locally valid solutions by increasing the locality dependency of the total loss. The weights λ_{CON} and λ_{COD} determine this trade-off. But at the same time, this relative importance can be argued to be driven by different input distributions. In the case of ON-OFF, $\mathcal{L}^{\text{CON}}$ compares the prediction difference over inputs with different variability than in the SINGLE case. Perturbations in ON-OFF inputs are different to SINGLE since in addition to the leaf-on perturbations they contain the perturbation from leaf-off images. The sensitivity of the loss to local accuracy is effectively (in expectation) changed by optimizing over a different input distribution. Hence, the qualitative difference in the performance ordering is also driven by effectively different weights given to local and non-local loss parts. The trade-off in relative importance of locality-dependent and independent losses does not translate easily to a trade-off between JS and MAE.

5.1.3 ON-OFF and SINGLE

As was pointed out above, due to the different input and architecture it is not possible to disentangle data and architecture in comparisons of SINGLE and ON-OFF evaluation. Since the evaluation of SINGLE and ON-OFF is based on the same data set with ON-OFF only altering the input imagery (stacked leaf-on / leaf-off) a comparison can still be useful even though differences can not be causally attributed to input or architecture.

The magnitude of the impact that the adoption of the adversarial setting has on KL divergence is different in ON-OFF and SINGLE experiments. The ON-OFF runs show less difference between baseline and augmented runs than the SINGLE runs mainly because the ON-OFF baseline start at lower KL. Generally, the KL divergence minima are lower in ON-OFF, and approximately equal in MAE, PC and CE.

A hypothesis explaining the improved baseline runs is that the inclusion of stacked leaf-on and leaf-off data provides the network with additional structural information, e.g. whether a pixel contains mainly coniferous (evergreen) trees. While the pixel-wise measures are only slightly affected by this additional information, divergence evaluation results suggest it is beneficial to the recovery of regional target distribution. The impact of adversarial regularization on ON-OFF runs may therefore be minor because the stacked input already allows for predictive features in the baseline runs.

However, the observation of improved baseline runs may as well be explained by the different predictor architecture. An experimental design to answer this question should make sure ON-OFF and SINGLE have the same model capacity.

5.1.4 Joint Distribution

The percentile distribution can be assumed to be structured due to ecologically based correlations between canopy top and lower canopy heights. These structures were attempted to be used as implicit constraints. Furthermore, they were fitted with a K-Means model for the CE evaluation. The results show that the inclusion of the adversarial setting improves CE generally under this model. As for the spatial distribution measures it can be observed that the inclusion of $\mathcal{L}^{\text{CON}}$ is needed in some cases. It can be concluded that the runs with augmented losses indeed approximate the joint percentile distribution better and thus yield ecologically more plausible results on a pixel-wise basis than the baseline runs. However, as for spatial distribution measures, it can also be stated that improving the joint distribution is qualitatively different to finding constraints for the prediction of absolute values. The reduction in CE was not accompanied by a reduction in MAE or R^2.

The CE measure depends on a model that was fitted in an unsupervised fashion. Inaccurate fitting and lacking model capacity due to overly simplistic assumptions regarding the joint percentile distribution, would impact the CE evaluation directly. In future work, the evaluation with a more rigorously tested model should be preferred.

5.2 Results APEX Experiment

The experiment on APEX data was intended to evaluate the adversarial setting on non-corresponding data sets. The comparison shows that the use of the non-corresponding data sets $\mathcal{X}_{nc}$ and $\mathcal{Y}_{nc}$ was beneficial in terms of the global MAE. At the same time, the distributional measures CE and KL are improved generally (except KL on p50 and p70). While in the case of SINGLE and ON-OFF absolute error reduction and distributional approximation could not be reduced at the same time, this is the case for the APEX experiment. From the experimental set-up it is however not clear whether the reason for simultaneous improvement relies uniquely in the larger training data set or if it is also caused by distributional properties of the higher resolved data.

The use of $\mathcal{L}^{\mathrm{COD}}$ on training data sets complemented by non-corresponding samples could therefore be shown to possibly improve absolute error reduction and distributional approximation on small data sets in similar Remote Sensing inversion tasks. Furthermore, it was shown in the ON-OFF experiment that a regularization of $\mathcal{L}^{\mathrm{COD}}$ by $\mathcal{L}^{\mathrm{CON}}$ can improve distributional approximation. The introduction of $\mathcal{L}^{\mathrm{CON}}$ in the APEX experiment could therefore further improve the ADV run by constraining the non-locality property of $\mathcal{L}^{\mathrm{COD}}$.

5.3 Spatial Structure and Joint Distribution

Both CE and PC depend on location in contrast to JS that evaluates regional properties. The results of SINGLE and ON-OFF on the test set have shown that the loss configurations could not yield predictors that outperformed the baseline simultaneously in terms of pixel-wise (MAE) and distributional ac-

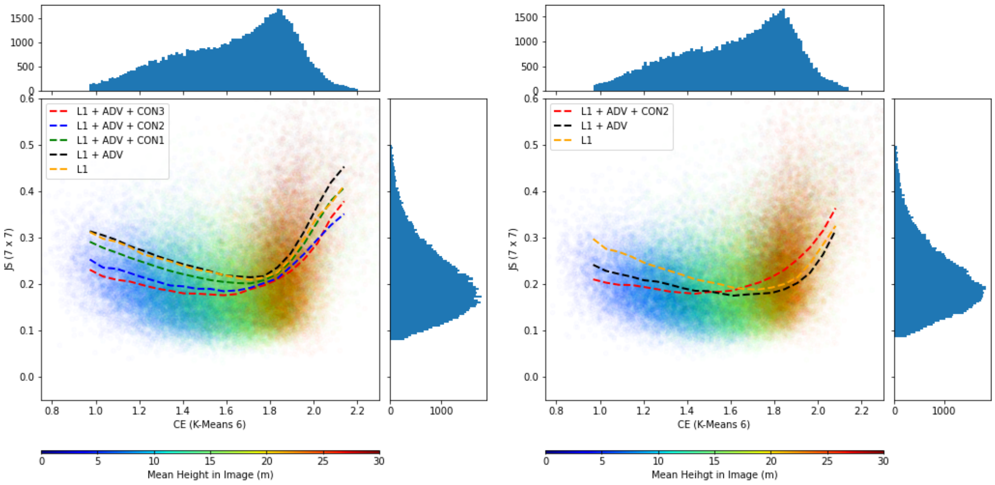

Figure 17: Scatter plot and marginal empirical distributions of JS and CE of SINGLE L1 + ADV + CON3 (left) and ON-OFF L1 + ADV + CON (right). Dashed lines are mean JS(CE).

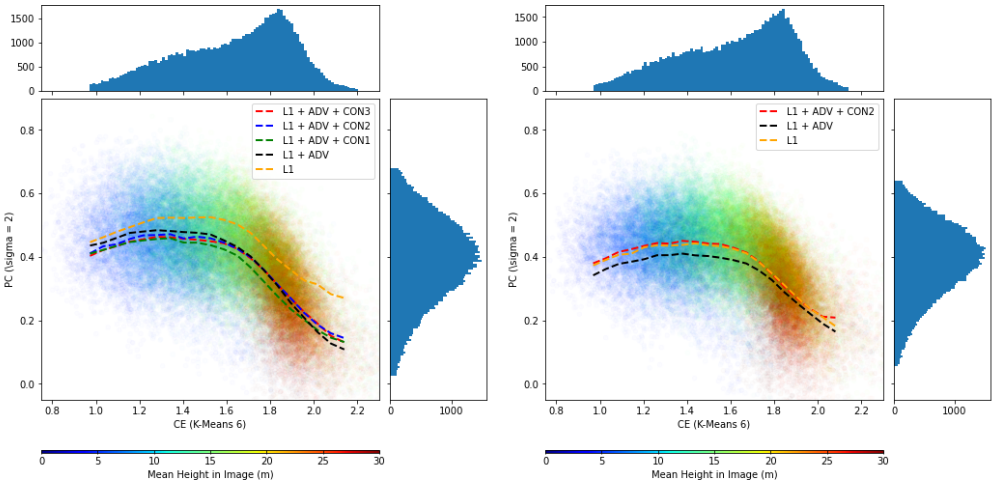

Figure 18: Scatter plot and marginal empirical distributions of PC and CE of SINGLE L1 + ADV + CON3 (left) and ON-OFF L1 + ADV + CON2 (right). Dashed lines are mean JS(PC).

curacy (JS/KL). Rather, the inclusion of $\mathcal{L}^{\text{CON}}$ and $\mathcal{L}^{\text{COD}}$ could be shown to improve the distributional match between prediction and target at the cost of deteriorating MAE. This is both true for the spatial target distribution (JS/KL) as well the vertical and thus pixel-wise distribution (CE). It is therefore interesting to ask whether a trade-off between the two concepts of structural and pixel-wise approximation exists at the level of single images.

In order to assess the effect of different loss configurations on the weighting of pixel-wise and structural accuracy the relationship between JS and CE and between PC and CE are plotted in Figs. 17 and 18. The scatter plots are shown in a color scale corresponding to true height to show the high degree of height stratification of CE and PC.

Fig. 17 shows both for SINGLE and ON-OFF that for all runs there is a negative correlation between JS and CE up to some critical CE c_{crit}. At CE $\leq c_{\text{crit}}$ a weak trade-off between pixel-wise and distributional accuracy could therefore exist. Since CE is highly correlated to p95 height, this correlation can be translated to a correlation over the specific height range $h \lesssim 20$. Hence, the negative correlation indicates a trade-off in low height ranges. Furthermore, it can be observed that $\mathcal{L}^{\text{CON}}$ reduces this correlation both in SINGLE and ON-OFF which highlights the importance of $\mathcal{L}^{\text{CON}}$ for recovery of canopy structure at low heights.

Comparing PC to CE shows a similar behaviour. However, the correlation at $h \lesssim 20$ m is much smaller. A difference in correlation between the baseline and augmented runs is still visible in SINGLE but not in ON-OFF. This indicates that the pixel-wise accuracy is rather at odds with the non-locality of JS/KL than the fitting of spatially contingent structure.

Concluding it can be stated that, at the single image level, $\mathcal{L}^{\text{CON}}$ decreases the negative correlation between pixel-wise and distributional accuracy at $h \leq 20$ m. Contrarily, the interdependency between spatial structure measured in PC and pixel-wise accuracy can only weakly be observed in this height range. Furthermore, no qualitative change can be observed in the PC-MAE correlation with the inclusion of $\mathcal{L}^{\text{COD}}$ or $\mathcal{L}^{\text{CON}}$. Hence, the augmented loss configurations indeed reduce a trade-off between spatial distribution approximation and pixel-wise accuracy. But their impact on reducing the gap between recovery of contingent spatial structures and pixel-wise accuracy is not observable. While this latter point could already be observed in absolute terms in Tabs. 12 and 13, this analysis shows that the qualitative behaviour of the predictor over different height ranges also doesn't change significantly with the proposed losses.

5.4 Prediction Variability under General Perturbations

The causal relationship between canopy height distribution and the sensitivity of the predictive features to perturbation (MVE) are not clear, especially since these features are not explicitly known. There is no trivial reason for different perturbative impacts between pixels with different heights in an approximately homogeneous canopy. Indeed, MVE stabilizes in most runs at $\sim$ 15 m in Figs. 19 and 20. Introspection from predictions as in Figs. 19 and 20 reveals the close relationship between strong spatial gradients and MAE respectely MVE. Differences in MVE are therefore attributed to feature variability under perturbation being particularly strong in heterogeneous canopy regions, e.g. strongly different shadows around local canopy maxima or woodland margins of specific heights. Visual inspection reveals that the strong correlation of MVE with spatial gradients does not clearly extend to $\mathrm{c_{MVE}}$. The height dependency of $\mathrm{c_{MVE}}$ being dependent on MAE(h) can not be interpreted with confidence.

Interrestingly, a decoupling of the height dependency can be observed in L1 + ADV runs at heights $h \lesssim 25$ m in Figs. 19 and 20. Only the extrema of 0 m and the highest canopy tops are reduced. This decoupling can also be observed to various degree for CON runs in SINGLE and ON-OFF. Similarly to the discussion above for the pixel-wise losses, this may indicate a generalization of the DNN features accross the height range, i.e. a reduction of the height bias in the unbalanced target data set.

This decoupling proves to deteriorate the prediction accuracy in SINGLE. $\mathcal{L}^{\mathrm{CON}}$ and $\mathcal{L}^{\mathrm{COD}}$ deteriorate $\mathrm{c_{MVE}}$ and MAE over all h in SINGLE. $\mathrm{c_{MVE}}$ in ON-OFF augmented runs is smaller than in SINGLE, especially at lower percentiles (p20, p50, p70). Furthermore, the CON run in ON-OFF effectively outperforms ON-OFF's L1 run in terms of $\mathrm{c_{MVE}}$. It is therefore argued that the ON-OFF set-up is shown to increase robustness against perturbations. As discussed above, it is however not possible whether the additional phe-

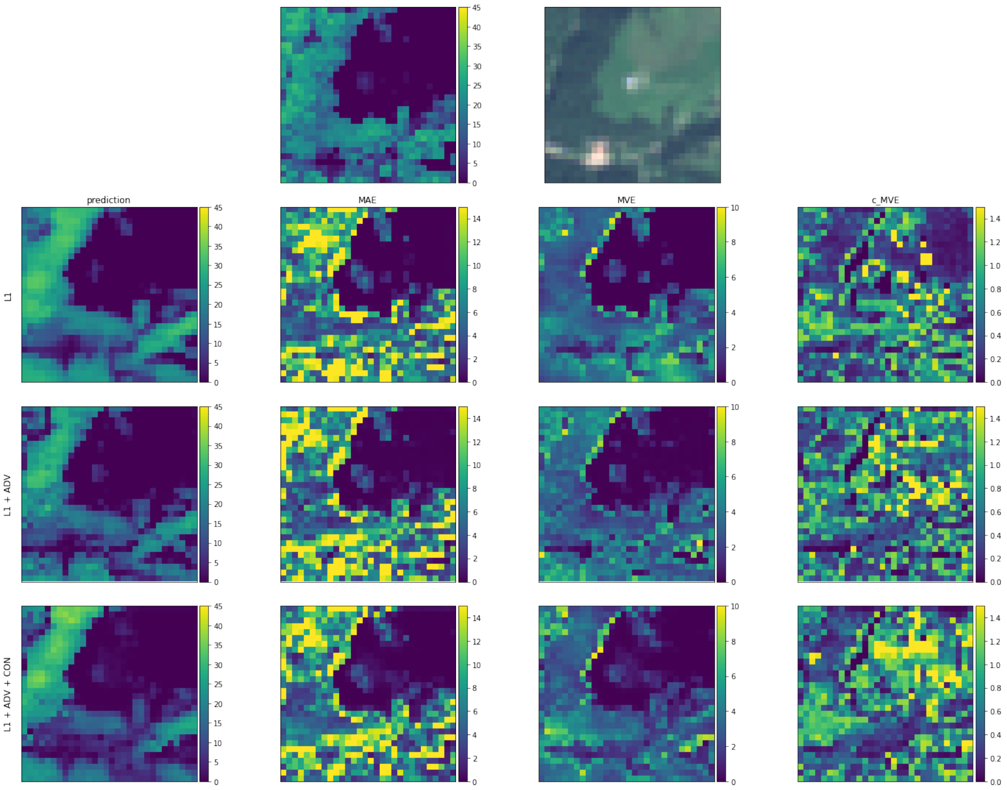

Figure 19: **ON-OFF**. First row: true p95, Sentinel-2 image. Lower rows, columns left - right: predicted p95, MAE, MVE, c_{MVE}. MAE and MVE are understood to be applied over all images in the time series.

nological information or the architecture modification is the reason for this difference to SINGLE.

5.5 Prediction under Topographic and Geometric Perturbation

In the previous sections perturbative effects across time were studied. However, also topography and illumination-viewing-sensor geometry play an important role in the prediction uncertainty quantified by MAE. Of particular interest is the prediction behaviour under varying slope s and aspect a.

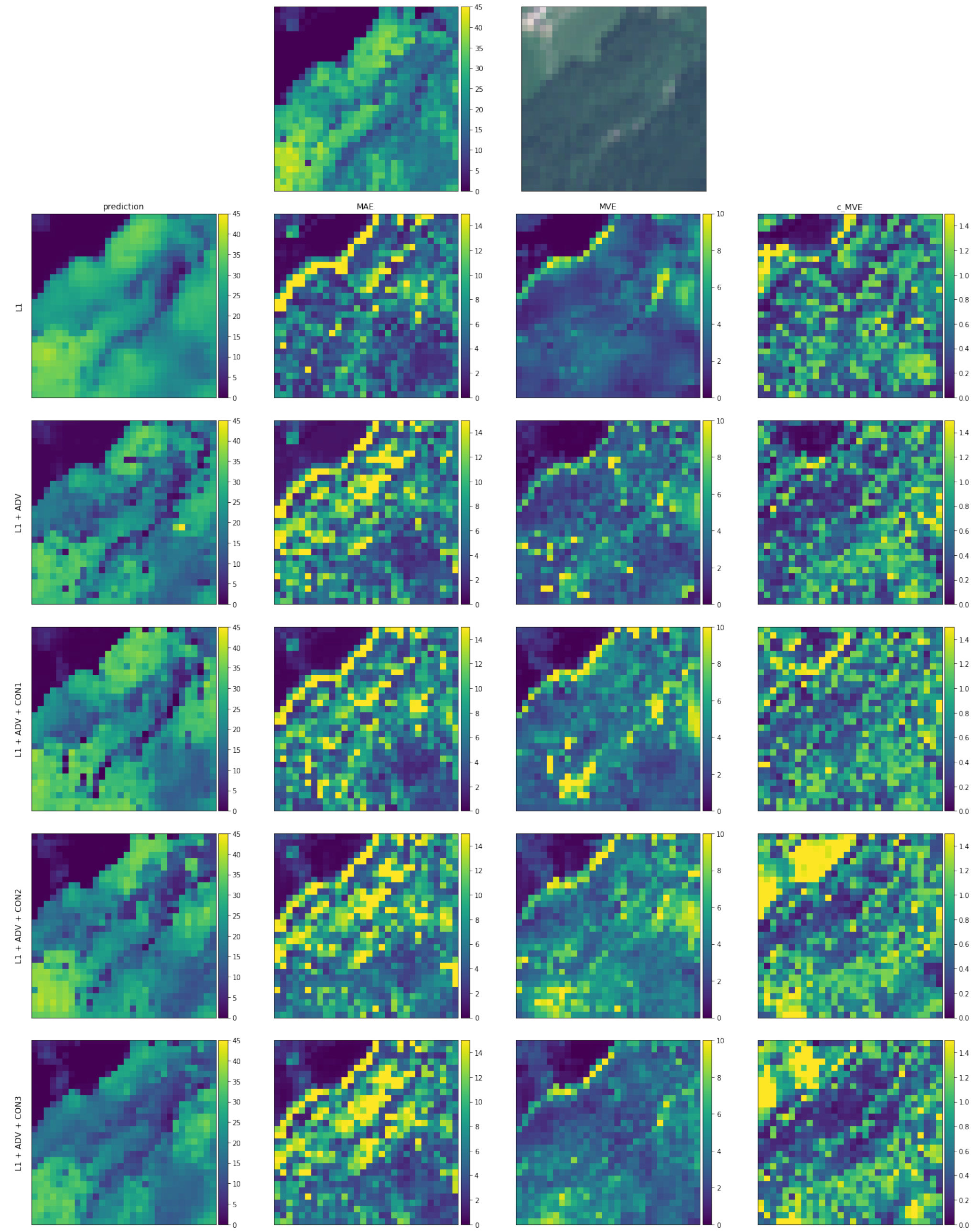

Figure 20: **SINGLE**. First row: true p95, Sentinel-2 image. Lower rows, columns left - right: predicted p95, MAE, MVE, c_{MVE}. MAE and MVE are understood to be applied over all images in the time series.

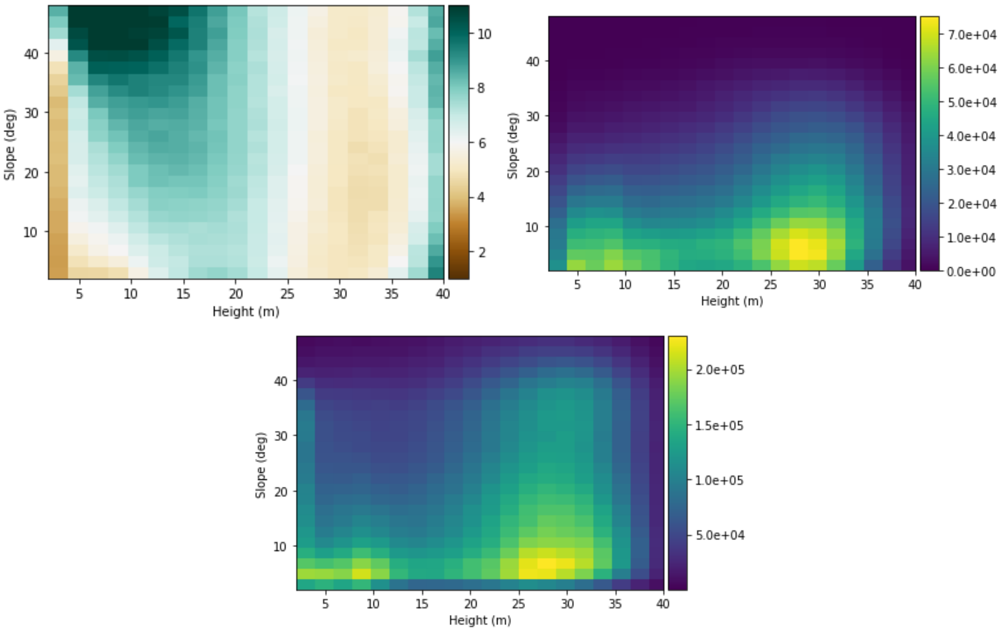

Figure 21: 2d histograms over Slope and true Height. Left: MAE of SINGLE L1. Counts of pixels in training (right) and test set (bottom).

Fig. 21 shows MAE of p95 as a function of s and true height (p95) h (for the run SINGLE L1). The wave pattern in height direction has already been discussed above. Two distinct regions in terms of MAE dependence on s can be observed. MAE at $h \lesssim 20$ m is positively correlated to s but there is no correlation at greater canopy heights. While this could be caused by the underlying p95 distribution in the test set, geometry-related phenomena may be the reason as well. The following explanations are hypothesized to play a role

> **Feature quality decay**. In the ortho-rectification process single pixel information is mixed which can impact both spatial and spectral features. The dependence of this effect on canopy height h not trivial.
>
> **Spectral Mixing** Single pixel spectra are likely to contain a growing amount of back scattered light from lower canopy layers or even ground surface at larger slopes s. This effect's height dependence is eminently related to height and vertical canopy structure.

Moreover, different distributions in vegetation over s in test and training set might deteriorate MAE as well. The present experimental set-up does not allow to study geometrical effects individually since the perturbation can not be controlled. However, the empirically found impacts do not exclude the hypothesis that MAE deteriorates due to geometric effects at heights $\lesssim 20$ m.

Similarly to the relationship of slope and MAE, two distinct regimes of correlation can be observed in Fig. 22, where the MAE distribution over aspect a and height h is shown. At $h \lesssim 20$ m MAE is decreased for more south looking pixels and inversely at $h \gtrsim 20$ m the minimum MAE is north-looking.

The radiance intensity and thus the SNR of Sentinel 1c reflectance can be assumed to be maximal in south-looking pixels. Shadow structure and backscatter from lower canopy layers is altered depending on the view geometry. A possible explanation for the different behaviour of MAE(a) could therefore be a trade off in signal capacity between SNR and geometry dependent feature creation. It must be noted as before that the present analysis cannot extract causal structure from the MAE distribution. Given the distribution of the training set exposing a underrepresentation of pixels at $h \lesssim 20$ m (see empirical histogram in Fig. 22), above discussed features of the MAE distribution may be an artefact of the imbalanced training set as well.

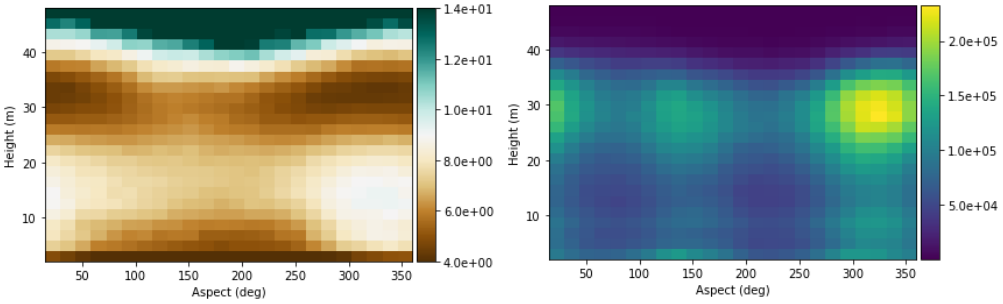

Figure 22: 2d histograms over Height and Aspect. *Left*: Median Absolute Error of SINGLE L1 over test set. Aspect definition: 0° north-looking, 180° south-looking. *Right*: Counts of pixels in training set.

This analysis has highlighted the need for control over perturbation for training and evaluation to find and reduce perturbation impact on the prediction.

Theoretically, perfect perturbation control can be achieved with the inclusion of simulated data. The use of RTM simulations is however itself prone to induce bias as a result of domain gap between true and simulated spectral data. The successful integration of RTM simulations with true spectral data is thus an important step towards controlled perturbation learning.

6 Conclusion

This thesis has evaluated the use of an adversarial setting and an invariance based loss for regression of canopy height percentiles, cover fraction and vertical canopy index. The losses were used to train an Xception type network on different types of multispectral imagery. It could be shown that in all experiments a combination of these losses lead to a better distributional approximation of percentiles, VCI and COV. This could be observed both in the empirical joint distribution over percentiles and the spatial distributions over locally normalized individual targets. The proposed losses constrained the network optimization to an empirically valid space.
This result is not bound to a specific model of the target distribution since the losses don't depend on an explicit formulation of statistical or physical constraints. It is therefore argued that improved distributional validity can be achieved with the same adversarial setting for other inversion tasks.

The distributional improvement (KL) did not translate into an improvement of pixel-wise absolute (MAE) or pixel-wise structural (PC) accuracy, however. Indeed, no supporting evidence was found for the assumption that the use of distributional constraints improves pixel-wise accuracy of a baseline run exclusively based on ℓ_1. It remains therefore unclear whether the inclusion of distributional constraints can help reduce pixel-wise errors in DNN derived maps of canopy structure.

However, the evaluated scheme could be useful in cases where ecological validity of canopy variable predictions is a concern. In a wider sense, the proposed scheme proved useful for any regression task where the distributional validity of the prediction is deemed important. The performance differences between different parameter configurations suggested that the relative weighting of loss components that depend on location (ℓ_1, $\mathcal{L}^{\text{CON}}$) and parts that do not have location dependency ($\mathcal{L}^{\text{COD}}$) can be used to fine tune a trade-off between vertical and horizontal distributional performance.

Moreover, the thesis also presented an example of a successful semi-supervised integration of non-related data sets in a supervised learning learning scheme. It could be shown that the presented adversarial setting for regression benefited the training both in terms of MAE and distributional validity. Since many inversion tasks in Remote Sensing are constrained by data availability caused by spatio-temporal mismatch, the presented adoption of an adversarial setting in inversion is particularly interesting for Remote Sensing applications.

Variations of the invariance based $\mathcal{L}^{\text{CON}}$ were evaluated. The variations yielded qualitatively comparable results. Differences were mainly observed in different performances for spatial and vertical distribution measures. Local gaussian normalization (CON_3) improved spatial distribution validity but decreased it over the vertical. Hence, the thesis' results concerning the formulation of $\mathcal{L}^{\text{CON}}$ showed that application specific normalization in the invariance loss can prove useful.

$\mathcal{L}^{\text{CON}}$ loss was tested on time series data, but can in principle be applied to simulated data. Simulated data would allow for control over atmospheric and geometric perturbations. While the use of time-series data prevented a distinction between perturbations during training and evaluation, perturbation control would allow to specifically identify and target failing modes of the predictor network. A concern of RTM integration is the high computational cost for simulations of suffcient quality. Low quality simulations introduce domain gap issues in the input domain. In the present thesis, a discriminator architecture that acts on subpatches of the sample window was used. The use of simulations of spectral imagery covering smaller regions than the input sample images could therefore be envisaged.

The thesis also evaluated the adversarial setting with input samples consisting of stacked leaf-on and leaf-off spectral imagery. Due to a architecture modification preventing direct comparison of SINGLE and ON-OFF experiments as well as lacking control over perturbation in the leaf-on and leaf-off time-series no final evaluation regarding the usefulness of such an approach could be made. Nevertheless it could be observed that distributional validity and robustness against perturbation may be better addressed with additional phenological information of leaf-off imagery. Distributional validity of simple L1 runs improved in the setting with stacked inputs. Furthermore, the analysis of the predictors' variation under perturbation revealed that $\mathcal{L}^{\text{CON}}$ was reduced only with stacked inputs. A thorough investigation what beneficial information phenological change for vertical structure prediction might have is therefore interesting independently of the performance of the proposed adversarial setting.

Finally, this thesis showed that that the dependence of MAE on slope and aspect is characteristically structured across canopy height. In absence of successful RTM integration, future work for canopy structure inversion should compare this result to models trained on topographically corrected data to address this issue.

Appendices

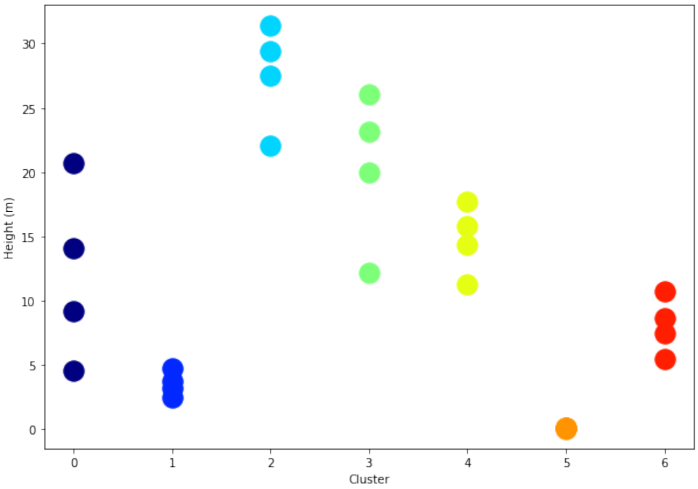

Figure 23: K-Means Percentile Clusters over APEX data training set $\mathcal{X}_{nc}$.

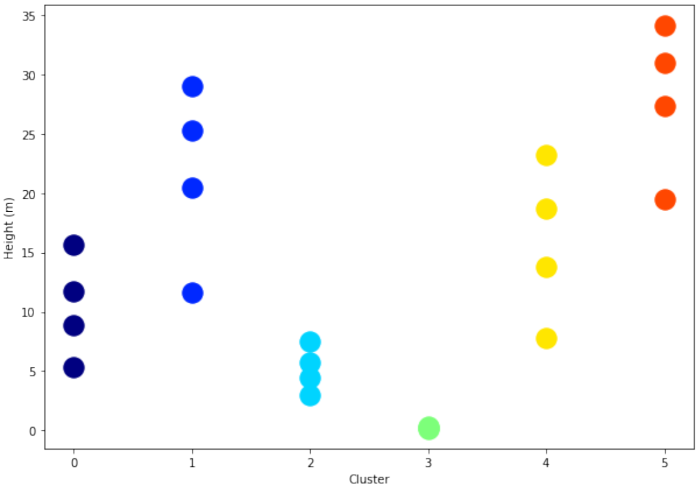

Figure 24: K-Means Percentile Clusters over SENTINEL data training set $\mathcal{X}_{nc}$.

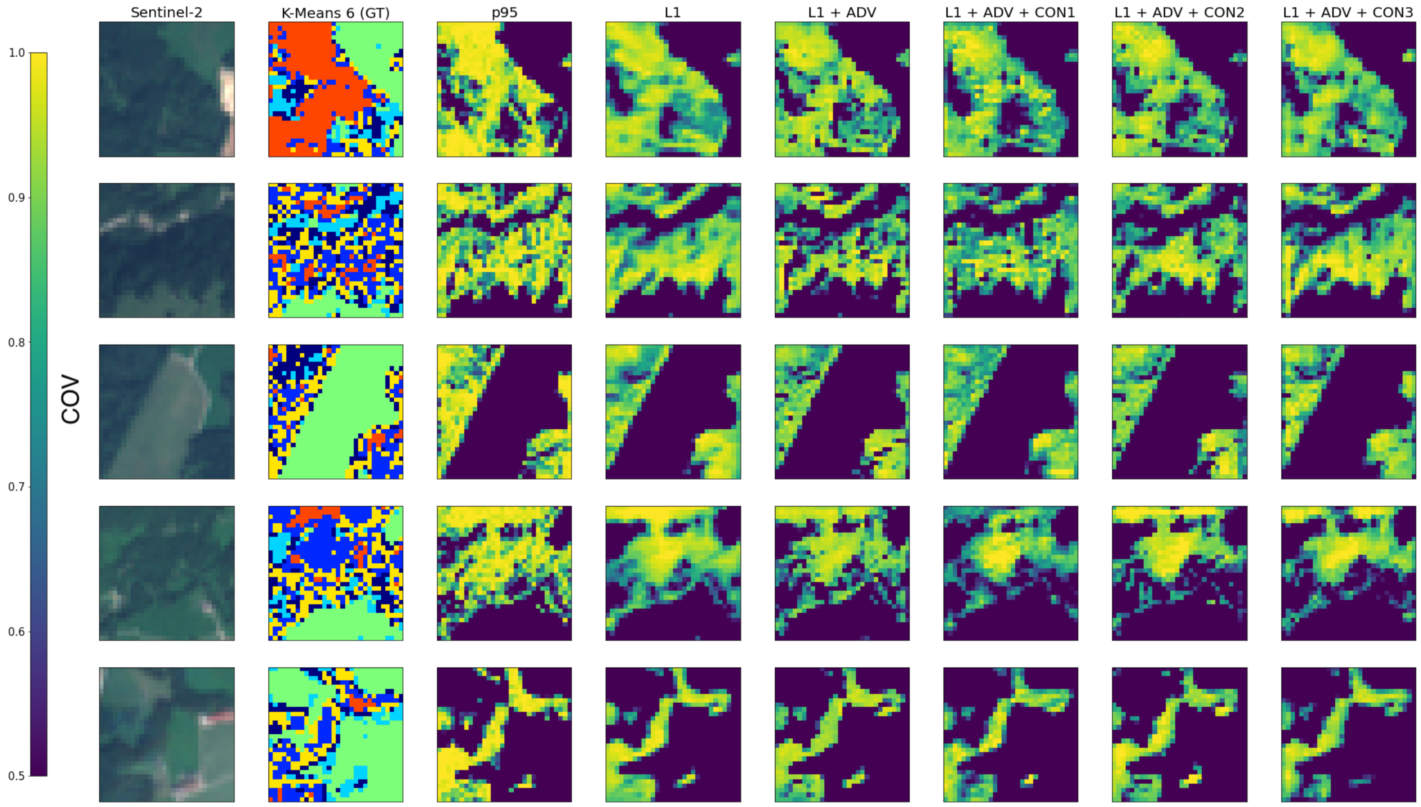

Figure 25: Sample COV prediction results for SINGLE runs. Columns left - right: Sentinel2 input, classified ground truth, p95 ground truth, SINGLE runs

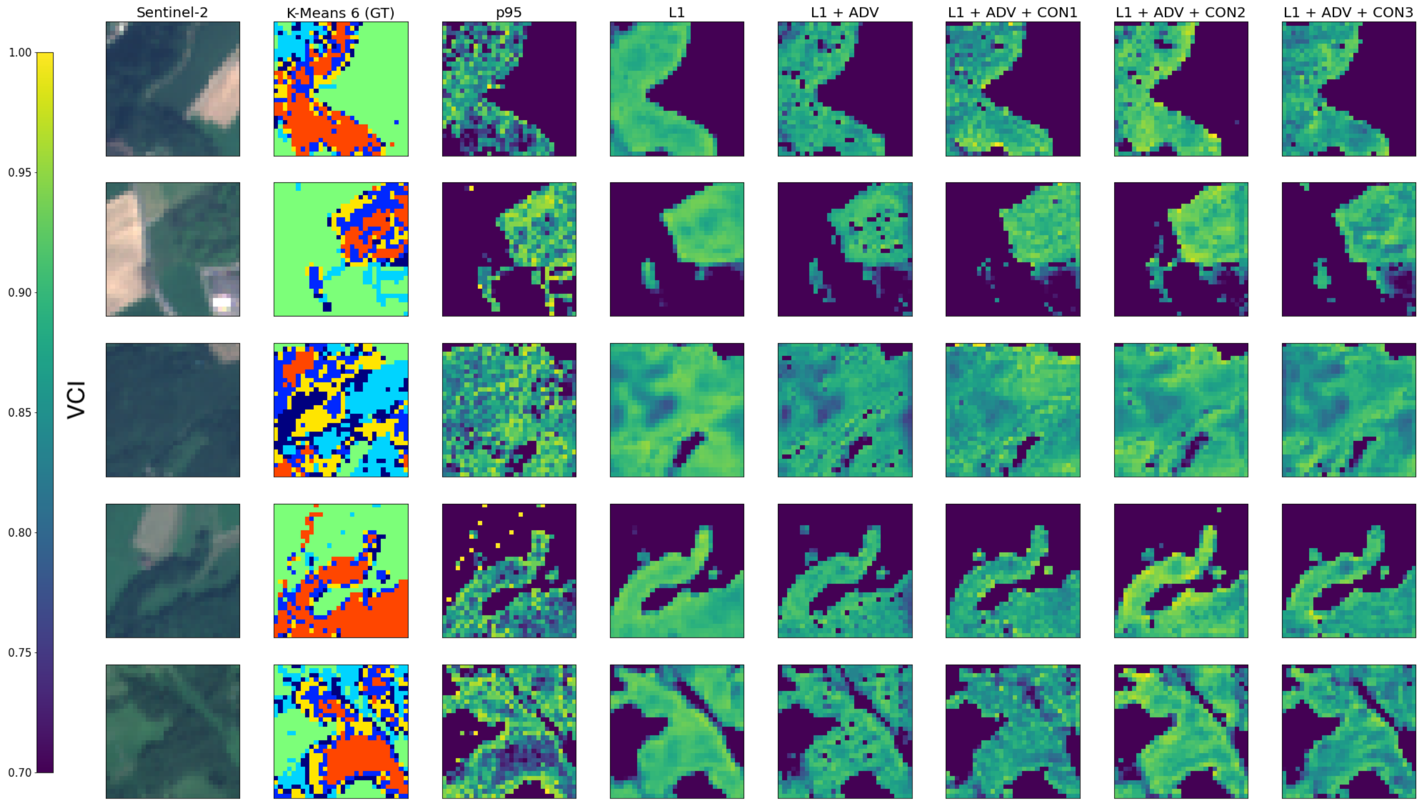

Figure 26: Sample VCI prediction results for SINGLE runs. Columns left - right: Sentinel2 input, classified ground truth, p95 ground truth, SINGLE runs

7 Declaration

I hereby declare that the submitted Thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the Thesis.

Jim Buffat 29.1.21

References

[1] D. Kumar, "Monitoring Forest Cover Changes Using Remote Sensing and GIS: A Global Prospective," *Research Journal of Environmental Sciences*, vol. 5, no. 2, 2011.

[2] A. Damm, E. Paul-Limoges, D. Kükenbrink, C. Bachofen, and F. Morsdorf, "Remote sensing of forest gas exchange: Considerations derived from a tomographic perspective," *Global Change Biology*, vol. 26, no. 4, 2020.

[3] D. Kükenbrink, F. D. Schneider, R. Leiterer, M. E. Schaepman, and F. Morsdorf, "Quantification of hidden canopy volume of airborne laser scanning data using a voxel traversal algorithm," *Remote Sensing of Environment*, vol. 194, 2017.

[4] A. S. Antonarakis, J. W. Munger, and P. R. Moorcroft, "Imaging spectroscopy- and lidar-derived estimates of canopy composition and structure to improve predictions of forest carbon fluxes and ecosystem dynamics," *Geophysical Research Letters*, vol. 41, no. 7, 2014.

[5] A. Brusa and D. E. Bunker, "Increasing the precision of canopy closure estimates from hemispherical photography: Blue channel analysis and under-exposure," *Agricultural and Forest Meteorology*, vol. 195-196, 2014.

[6] K. K. Singh, A. J. Davis, and R. K. Meentemeyer, "Detecting understory plant invasion in urban forests using LiDAR," *International Journal of Applied Earth Observation and Geoinformation*, vol. 38, 2015.

[7] Y. Wang, H. Weinacker, and B. Koch, "A Lidar point cloud based procedure for vertical canopy structure analysis and 3D single tree modelling in forest," *Sensors*, vol. 8, no. 6, 2008.

[8] H. Hamraz, M. A. Contreras, and J. Zhang, "Vertical stratification of forest canopy for segmentation of under-story trees within small-footprint airborne LiDAR point clouds," 2016.

[9] H. T. Ishii, S. I. Tanabe, and T. Hiura, "Exploring the relationships

among canopy structure, stand productivity, and biodiversity of temperate forest ecosystems," 2004.

[10] F. D. Schneider, D. Kükenbrink, M. E. Schaepman, D. S. Schimel, and F. Morsdorf, "Quantifying 3D structure and occlusion in dense tropical and temperate forests using close-range LiDAR," *Agricultural and Forest Meteorology*, vol. 268, 2019.

[11] D. S. Ellsworth and P. B. Reich, "Canopy structure and vertical patterns of photosynthesis and related leaf traits in a deciduous forest," *Oecologia*, vol. 96, no. 2, 1993.

[12] J. Chen, S. Jin, and P. Du, "Roles of horizontal and vertical tree canopy structure in mitigating daytime and nighttime urban heat island effects," *International Journal of Applied Earth Observation and Geoinformation*, vol. 89, 2020.

[13] Q. Wang and P. Li, "Canopy vertical heterogeneity plays a critical role in reflectance simulation," *Agricultural and Forest Meteorology*, vol. 169, pp. 111 – 121, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168192312002961

[14] N. Lang, K. Schindler, and J. D. Wegner, "Country-wide high-resolution vegetation height mapping with Sentinel-2," *ArXiv*, vol. abs/1904.13270, 2019.

[15] F. Morsdorf, E. Meier, B. Kötz, K. I. Itten, M. Dobbertin, and B. Allgöwer, "LIDAR-based geometric reconstruction of boreal type forest stands at single tree level for forest and wildland fire management," *Remote Sensing of Environment*, vol. 92, no. 3, pp. 353 – 362, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425704001464

[16] F. Morsdorf, B. Kötz, E. Meier, K. I. Itten, and B. Allgöwer, "Estimation of LAI and fractional cover from small footprint airborne laser scanning data based on gap fraction," *Remote Sensing of Environment*, vol. 104, no. 1, 2006.

[17] S. Bae, S. R. Levick, L. Heidrich, P. Magdon, B. F. Leutner, S. Wöllauer,

A. Serebryanyk, T. Nauss, P. Krzystek, M. M. Gossner, P. Schall, C. Heibl, C. Bässler, I. Doerfler, E. D. Schulze, F. S. Krah, H. Culmsee, K. Jung, M. Heurich, M. Fischer, S. Seibold, S. Thorn, T. Gerlach, T. Hothorn, W. W. Weisser, and J. Müller, "Radar vision in the mapping of forest biodiversity from space," *Nature Communications*, vol. 10, no. 1, 2019.

[18] W. Qi, S. K. Lee, S. Hancock, S. Luthcke, H. Tang, J. Armston, and R. Dubayah, "Improved forest height estimation by fusion of simulated GEDI Lidar data and TanDEM-X InSAR data," *Remote Sensing of Environment*, vol. 221, 2019.

[19] T. Markus, T. Neumann, A. Martino, W. Abdalati, K. Brunt, B. Csatho, S. Farrell, H. Fricker, A. Gardner, D. Harding, M. Jasinski, R. Kwok, L. Magruder, D. Lubin, S. Luthcke, J. Morison, R. Nelson, A. Neuenschwander, S. Palm, S. Popescu, C. K. Shum, B. E. Schutz, B. Smith, Y. Yang, and J. Zwally, "The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): Science requirements, concept, and implementation," *Remote Sensing of Environment*, vol. 190, 2017.

[20] M. García, S. Saatchi, S. Ustin, and H. Balzter, "Modelling forest canopy height by integrating airborne LiDAR samples with satellite Radar and multispectral imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 66, 2018.

[21] D. Fawcett, W. Verhoef, D. Schläpfer, F. Schneider, M. Schaepman, and A. Damm, "Advancing retrievals of surface reflectance and vegetation indices over forest ecosystems by combining imaging spectroscopy, digital object models, and 3d canopy modelling," *Remote Sensing of Environment*, vol. 204, pp. 583 – 595, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425717304637

[22] Q. Wang and P. Li, "Canopy vertical heterogeneity plays a critical role in reflectance simulation," *Agricultural and Forest Meteorology*, vol. 169, pp. 111 – 121, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168192312002961

[23] G. Camps-Valls, L. Martino, D. H. Svendsen, M. Campos-Taberner, J. Muñoz-Marí, V. Laparra, D. Luengo, and F. J. García-Haro, "Physics-

aware Gaussian processes in remote sensing," *Applied Soft Computing Journal*, vol. 68, 2018.

[24] A. I. Logvin, L. P. Ligthart, and A. I. Kozlov, "Methods for solving inverse problems in radar remote sensing," in *14th International Conference on Microwaves, Radar and Wireless Communications, MIKON 2002*, vol. 2, 2002.

[25] H. Ren, R. Stewart, J. Song, V. Kuleshov, and S. Ermon, "Adversarial Constraint Learning for Structured Prediction," 2018.

[26] R. Stewart and S. Ermon, "Label-Free Supervision of Neural Networks with Physics and Domain Knowledge," 1 2016.

[27] N. Muralidhar, M. R. Islam, M. Marwah, A. Karpatne, and N. Ramakrishnan, "Incorporating Prior Domain Knowledge into Deep Neural Networks," in *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 2019.

[28] A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (pgnn): An application in lake temperature modeling," 2018.

[29] T. Li and V. Srikumar, "Augmenting neural networks with first-order logic," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020.

[30] D. Fawcett, W. Verhoef, D. Schläpfer, F. D. Schneider, M. E. Schaepman, and A. Damm, "Advancing retrievals of surface reflectance and vegetation indices over forest ecosystems by combining imaging spectroscopy, digital object models, and 3D canopy modelling," *Remote Sensing of Environment*, vol. 204, pp. 583 – 595, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425717304637

[31] C. Dong, G. Zhao, Y. Meng, B. Li, and B. Peng, "The effect of topographic correction on forest tree species classification accuracy," *Remote Sensing*, vol. 12, no. 5, 2020.

[32] D. Kukenbrink, A. Hueni, F. D. Schneider, A. Damm, J. P. Gastellu-Etchegorry, M. E. Schaepman, and F. Morsdorf, "Mapping the Irradiance Field of a Single Tree: Quantifying Vegetation-Induced Adjacency Effects," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, 2019.

[33] F. D. Schneider, R. Leiterer, F. Morsdorf, J.-P. Gastellu-Etchegorry, N. Lauret, N. Pfeifer, and M. E. Schaepman, "Simulating imaging spectrometer data: 3D forest modeling based on LiDAR and in situ data," *Remote Sensing of Environment*, vol. 152, pp. 235 – 250, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425714002284

[34] L. Wang, W. Cho, and K. J. Yoon, "Deceiving Image-to-Image Translation Networks for Autonomous Driving with Adversarial Perturbations," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, 2020.

[35] Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. C. Duchi, "Unlabeled data improves adversarial robustness," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[37] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang, "Adversarial training can hurt generalization," 2019.

[38] J. Uesato, J. B. Alayrac, P. S. Huang, R. Stanforth, A. Fawzi, and P. Kohli, "Are labels required for improving adversarial robustness?" in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[39] R. Mangal, A. V. Nori, and A. Orso, "Robustness of neural networks: A probabilistic and practical approach," in *Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results, ICSE-NIER 2019*, 2019.

[40] A. Laugros, A. Caplier, and M. Ospici, "Are adversarial robustness and common perturbation robustness independent attributes?" in *Proceed-*

ings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019, 2019.

[41] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini, “Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services,” *Remote Sensing of Environment*, vol. 120, 2012.

[42] M. E. Schaepman, M. Jehle, A. Hueni, P. D’Odorico, A. Damma, J. Weyermann, F. D. Schneider, V. Laurent, C. Popp, F. C. Seidel, K. Lenhard, P. Gege, C. Küchler, J. Brazile, P. Kohler, L. De Vos, K. Meuleman, R. Meynart, D. Schläpfer, M. Kneubühler, and K. I. Itten, “Advanced radiometry measurements and Earth science applicationswith the Airborne Prism Experiment (APEX),” *Remote Sensing of Environment*, vol. 158, no. 1, 2015.

[43] G. Büttner and B. Kosztra, “Updated CLC illustrated nomenclature guidelines, European Topic Centre on Urban, land and soil systems (ETC/ULS),” Tech. Rep., 2017.

[44] M. Hancher, I. Housman, and G. Donchyts, “Cloud Score Algorithm and Shadow Masking.” [Online]. Available: https://code.earthengine.google.com/0f0a7c6b152d4d5909b4dcb7f1af7d7b

[45] R. Richter and D. Schläpfer, “Atmospheric and Topographic Correction (ATCOR Theoretical Background Document),” Tech. Rep., 2019.

[46] Etat de Fribourg, “ALS coverage Fribourg,” 2017.

[47] M. Bruggisser, M. Hollaus, D. Kükenbrink, and N. Pfeifer, “COMPARISON of FOREST STRUCTURE METRICS DERIVED from UAV LIDAR and ALS DATA,” in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, no. 2/W5, 2019.

[48] rapidlasso GmbH, “LAStools - efficient LiDAR processing software,” 2020.

[49] R. Stewart and S. Ermon, “Label-free supervision of neural networks

with physics and domain knowledge," 09 2016.

[50] S. Ermon, R. L. Bras, S. K. Suram, J. M. Gregoire, C. Gomes, B. Selman, and R. B. van Dover, "Pattern decomposition with complex combinatorial constraints: Application to materials discovery," 2014.

[51] R. Aabeyir, S. Adu-Bredu, W. A. Agyare, and M. J. Weir, "Allometric models for estimating aboveground biomass in the tropical woodlands of Ghana, West Africa," *Forest Ecosystems*, vol. 7, no. 1, 2020.

[52] W. A. Mugasha, E. E. Mwakalukwa, E. Luoga, R. E. Malimbwi, E. Zahabu, D. S. Silayo, G. Sola, P. Crete, M. Henry, and A. Kashindye, "Allometric Models for Estimating Tree Volume and Aboveground Biomass in Lowland Forests of Tanzania," *International Journal of Forestry Research*, vol. 2016, 2016.

[53] R. I. Barbosa, P. N. Ramírez-Narváez, P. M. Fearnside, C. D. A. Villacorta, and L. C. d. S. Carvalho, "Allometric models to estimate tree height in northern amazonian ecotone forests," *Acta Amazonica*, vol. 49, no. 2, 2019.

[54] M. Henry, A. Bombelli, C. Trotta, A. Alessandrini, L. Birigazzi, G. Sola, G. Vieilledent, P. Santenoise, F. Longuetaud, R. Valentini, N. Picard, and L. Saint-André, "GlobAllomeTree: International platform for tree allometric equations to support volume, biomass and carbon assessment," *IForest*, vol. 6, no. 6, 2013.

[55] F. E. Fassnacht, J. Poblete-Olivares, L. Rivero, J. Lopatin, A. Ceballos-Comisso, and M. Galleguillos, "Using Sentinel-2 and canopy height models to derive a landscape-level biomass map covering multiple vegetation types," *International Journal of Applied Earth Observation and Geoinformation*, vol. 94, 2021.

[56] P. Köhler and A. Huth, "Towards ground-truthing of spaceborne estimates of above-ground life biomass and leaf area index in tropical rain forests," *Biogeosciences*, vol. 7, no. 8, 2010.

[57] C. Pascual, A. García-Abril, W. B. Cohen, and S. Martín-Fernández, "Relationship between LiDAR-derived forest canopy height and Landsat

images," *International Journal of Remote Sensing*, vol. 31, no. 5, 2010.

[58] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," 2017.

[59] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, 2017.

[60] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems*, 2016.

[61] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017.

[62] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," in *Advances in Neural Information Processing Systems*, vol. 2017-December, 2017.

[63] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, 2017.

[64] A. Daw, R. Q. Thomas, C. C. Carey, J. S. Read, A. P. Appling, and A. Karpatne, "Physics-guided architecture (pga) of neural networks for quantifying uncertainty in lake temperature modeling," in *Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020*, 2020.

[65] A. T. Mohan, N. Lubbers, D. Livescu, and M. Chertkov, "Embedding hard physical constraints in neural network coarse-graining of 3D turbulence," 2020.

[66] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, 2017.

[67] N. Lang, K. Schindler, and J. D. Wegner, "Country-wide high-resolution vegetation height mapping with sentinel-2," *ArXiv*, vol. abs/1904.13270, 2019.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, 2016.

[69] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, and T. Liu, "Residual Networks of Residual Networks: Multilevel Residual Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, 2018.

[70] S. Ioffe and C. Szegedy, "Batch norm," *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, 2015.

[71] S. Wu, G. Li, L. Deng, L. Liu, D. Wu, Y. Xie, and L. Shi, "L1 -Norm Batch Normalization for Efficient Training of Deep Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 7, 2019.

[72] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation learning with Deep Convolutional GANs," *International Conference on Learning Representations*, 2016.

[73] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, 2019.

[74] C. Bourgoin, J. Betbeder, P. Couteron, L. Blanc, H. Dessard, J. Oszwald, R. Le Roux, G. Cornu, L. Reymondin, L. Mazzei, P. Sist, P. Läderach, and V. Gond, "UAV-based canopy textures assess changes in forest structure from long-term degradation," *Ecological Indicators*, vol. 115, 2020.

[75] P. Ploton, R. Pélissier, N. Barbier, C. Proisy, B. R. Ramesh, and P. Couteron, "Canopy texture analysis for large-scale assessments of tropical forest stand structure and biomass," in *Treetops at Risk: Chal-*

lenges of Global Canopy Ecology and Conservation, 2013.

[76] A. P. Sviridov, Z. Ulissi, V. V. Chernomordik, M. Hassan, and A. H. Gandjbakhche, "Visualization of biological texture using correlation coefficient images," *Journal of Biomedical Optics*, vol. 11, no. 6, pp. 1 – 3, 2006. [Online]. Available: https://doi.org/10.1117/1.2400248

[77] J. R. Mathiassen, A. Skavhaug, and K. Bø, "Texture similarity measure using kullback-leibler divergence between gamma distributions," in *Computer Vision — ECCV 2002*, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 133–147.

[78] A. D. E. Maliani, M. E. Hassouni, N. Lasmar, and Y. Berthoumieu, "Texture classification based on the generalized gamma distribution and the dual tree complex wavelet transform," in *2010 5th International Symposium On I/V Communications and Mobile Network*, 2010, pp. 1–4.

[79] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, 2002.

[80] J. Lin, "Divergence Measures Based on the Shannon Entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, 1991.

[81] L. Weng, "From gan to wgan," 2019.

[82] A. Kadir, L. E. Nugroho, A. Susanto, and P. I. Santosa, "Experiments of distance measurements in a foliage plant retrieval system," 2013.

[83] D. Freedman and P. Diaconis, "On the histogram as a density estimator:l2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 57, pp. 453–476, 1981.

[84] R. Leiterer, R. Furrer, M. E. Schaepman, and F. Morsdorf, "Forest canopy-structure characterization : A data-driven approach," *Forest Ecology and Management*, vol. 358, pp. 48–61, 9 2015. [Online]. Available: https://www.zora.uzh.ch/id/eprint/116685/

[85] R. Leiterer, H. Torabzadeh, R. Furrer, M. E. Schaepman, and F. Morsdorf, “Towards automated characterization of canopy layering in mixed temperate forests using airborne laser scanning,” *Forests*, vol. 6, no. 11, 2015.

[86] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 63, no. 2, 2001.

[87] R. Richter and D. Schläpfer, “Geo-atmospheric processing of airborne imaging spectrometry data. part 2: Atmospheric/topographic correction,” *International Journal of Remote Sensing*, vol. 23, pp. 2631–2649, 07 2002.

[88] J. Weyermann, D. Schläpfer, A. Hueni, M. Kneubühler, and M. Schaepman, “Spectral Angle Mapper (SAM) for anisotropy class indexing in imaging spectrometry data,” in *Imaging Spectrometry XIV*, vol. 7457, 2009.

[89] E. Halme, P. Pellikka, and M. Mõttus, “Utility of hyperspectral compared to multispectral remote sensing data in estimating forest biomass and structure variables in Finnish boreal forest,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 83, 2019.

[90] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.