Improved Semisupervised UNet Deep Learning Model for Forest Height Mapping With Satellite SAR and Optical Data

Shaojia Ge, Hong Gu, Weimin Su, Jaan Praks, *Member, IEEE*, and Oleg Antropov, *Member, IEEE*

***Abstract*—In this study, we introduce an improved semisupervised deep learning approach, and demonstrate its suitability for modeling the relationship between forest structural parameters and satellite remote sensing imagery and producing forest maps. The improved approach is based on a popular UNet model, modified and fine-tuned to improve the forest parameter prediction performance. Within the improved model, squeeze-and-excitation blocks are embedded to recalibrate the multisource features via retrieved channel-wise self-attention and a novel cross-pseudo regression strategy is implemented to train the model in a semisupervised way. The improvement imposes consistency learning on two perturbed network branches: 1) generating regression pseudo-reference; 2) expanding the dataset size. For demonstration, we used satellite synthetic aperture radar (SAR) Sentinel-1 and multispectral optical Sentinel-2 images as remote sensing data, complemented with reference data represented by forest tree height as one of the key forest structural variables. The study area is located in a boreal forestland in Central Finland. Proposed approach showed larger accuracy compared to traditional machine learning methods such as random forests and boosting trees, and baseline UNet model. Best accuracy figures for forest tree height were achieved with combined SAR and optical imagery and were as small as 24.1% root-mean-square error (RMSE) on pixel-level and 15.4% RMSE on forest stand level.**

***Index Terms*—Deep learning (DL), regression, semisupervised, Sentinel-1, Sentinel-2, synthetic aperture radar (SAR), UNet.**

I. Introduction

INCREASINGLY important climate change monitoring requires precise methods for forest carbon assessment using earth observation (EO) sensors [1]. Forests are essential in maintaining healthy ecosystem interaction and biodiversity on earth and have an important role in tackling climate change since forests can quickly restore carbon stock on ground. Presently, the forest carbon stock and its change are typically estimated through inventories of forest structural variables [2], which still involves large amount of manual labour. Thus, accurate, timely, and reliable approaches for retrieval of forest structural variables from EO data (such as forest tree height, growing stock volume, tree age and density, tree species, and forest above ground biomass) are needed. Updated information on forest resources is also necessary for forest management purposes.

Satellite remote sensing combined with *in situ* forest measurements can be considered an effective means for producing forest attribute maps and forest estimates on various areal levels [3]. Suitable EO datasets include satellite optical and imaging radar data, augmented by forest plot or forest stand-level reference measurements. Airborne laser scanning data are costly and rarely available over large areas, while usability of remotely sensed optical datasets are often affected by cloudiness and haze. Synthetic aperture radar (SAR) data can be considered suitable over boreal regions and other regions with frequent cloud cover or poor light conditions. If both SAR and optical datasets are available, it seems useful to combine data sources within modeling approaches to take advantage of their recognized synergistic potential [4]. Traditionally, physics model based approaches and machine learning models were found suitable for modeling relationships between EO measurements and forest structural variables. SAR data approaches particularly using L, C, and X-band satellite data were well elaborated for forest structural variable prediction [5]. Depending on data source used, forest mapping has been done using inversion of semiempirical and physics-based models, statistical, and machine learning approaches [6]–[12]. Utility of various data sources, potential of combined use of SAR and optical data, as well as different regression approaches were studied. However, forest parameter prediction accuracy is still relatively low to meet forest users' needs. Common remaining challenges include lack of high quality or representative models for training, poor performance of pixel-based models that do not take into account spatial context, lack of model flexibility. This requires introducing hand-engineered features to capture spatial dependencies. A possible alternative is using deep learning (DL) models like convolutional

Manuscript received 31 March 2022; revised 4 June 2022; accepted 16 June 2022. Date of publication 4 July 2022; date of current version 27 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62001229, Grant 62101264 and Grant 62101260, and in part by China Postdoctoral Science Foundation under Grant 2020M681604. The work of Oleg Antropov was supported by Multico project funded by Business Finland and Forest Carbon Monitoring project funded by European Space Agency. *(Corresponding author: Oleg Antropov.)*

Shaojia Ge, Hong Gu, and Weimin Su are with the School of Electronic and Optical Engineering, Department of Electronic Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: geshaojia@njust.edu.cn; guhong666@126.com; suweimin1959@outlook.com).

Jaan Praks is with the Department of Electronics and Nanoengineering, Aalto University, 02150 Espoo, Finland (e-mail: jaan.praks@aalto.fi).

Oleg Antropov is with the VTT Technical Research Centre of Finland, 02044 Espoo, Finland (e-mail: oleg.antropov@vtt.fi).

Digital Object Identifier 10.1109/JSTARS.2022.3188201

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

neural network (CNN) that can capture both spectral and spatial dependencies within the EO images.

Since recently, DL models have been increasingly used in EO applications [13], [14], indicating potential for analyzing complex structured vegetation such as boreal forests. Despite promising results for classification tasks, DL has not yet been used extensively for regression modeling tasks in environmental studies [15]. Several prior studies focusing on prediction of continuous forest variables were reported [16]–[22]. One example is the use of CNN with Sentinel-2 data for prediction of forest tree height in Switzerland and Gabon [22], with average root-mean-squared errors (RMSEs) of 3.4 m and 5.6 m, respectively. Stacked sparse autoencoders were used to predict above ground-biomass in mixed broadleaved and coniferous forests of China, outperforming several traditional regression approaches [20], [21]. In another study, the Chimera predictor suitable for forest mapping also outperformed random forest and support vector machines in all regression tasks.

To perform well, DL models need to be trained using large amount of reference (ground truth) data, that is either commonly not available or imperfect/inaccurate in forest studies. In this context, transfer learning and semisupervised learning have become appealing options [23]–[27]. Transfer learning entails adapting a pretrained model and fine-tuning it using available sample of training data. One of such approaches has been recently demonstrated with optical Sentinel-2 data over boreal forest in Finland [19]. Semisupervised learning, in order to improve modeling, concentrates on improving the use of unlabeled training data. In the context of semisupervised learning, two primary approaches are considered suitable: 1) *self-training* [28]–[30], that a model is firstly pretrained, followed by using the model to generate so-called pseudo-labels, and finally retrain the model with all data including true labels and pseudo-labels; 2) *consistency learning* [31], [32], that the idea is to apply different transformations on unlabeled training data, encourage the model to produce similar outputs in spite of the different transformation, and in this way to lead the model learning a more compact data representation.

Inspired by the approach of Chen et al. [33], we adopt in this study a new strategy called cross-pseudo regression (CPR), which brings the two options together and allows to train the model with unlabeled data in a semisupervised way. To our knowledge, this is the first attempt to introduce pseudo ground truth and consistency learning into regression modeling in the context of satellite EO based forest inventory.

In this study, we focus on forest tree height as one of the key forest structural variables, and demonstrate the utility of developed approaches in boreal forest environment. We expect the model can be suitable for capturing relationships between EO data and other forest variables (e.g., growing stock volume, above ground biomass, and stem volume) as well, also in different forest biomes. For demonstrating our approach, we will use free European Copernicus programme Sentinel-1 imaging radar satellite and Sentinel-2 optical satellite datasets over a boreal forest test site in Finland.

Our primary contributions can be summarized as follows.

1) We introduce an improved DL model based on UNet [34], which is suitable for predicting forest attributes and producing forest maps.
2) The proposed model introduces two key modifications: 1) feature recalibration based on modified squeeze-and-excitation (SE) channel-attention mechanism; 2) CPR strategy.
3) For validation purposes, performance of the developed model is compared to basic UNet model, several popular machine learning approaches as well as alternative semisupervised DL models in predicting forest tree height using SAR Sentinel-1 and optical Sentinel-2 data.
4) To the best of our knowledge, this is the first study utilizing DL models for wall-to-wall prediction of forest structural variables using time series of Sentinel-1 SAR images, or combined optical and SAR imagery.

II. Improved Semisupervised UNet Framework

We selected UNet [34], a variant of fully convolutional network, as a baseline DL model. The UNet model was originally proposed for biomedical image segmentation, and is presently often used in various semantic segmentation tasks [35]–[38].

A. Basic UNet Model

The basic UNet (also known as Vanilla UNet) uses convolutional network to extract features [34]. Unlike CNN [39], the fully convolutional and skip-connection structures allow UNet to extract deeper features of input data, maintain good fusion ability at all levels, while keeping the feature map size unchanged [34]. This suggests it can be naturally applied for pixel-level classification and regression tasks.

The overall structure of UNet is symmetric, similar to encoder–decoder [34]. The encoder is responsible for feature extraction, and the decoder restores the feature map to the original size. The structure of UNet is identical to representation in Fig. 1 if SE blocks are removed. Each box in the UNet indicates a feature map, where the corresponding size is denoted near the boxes. The blue arrow indicates a double-convolution structure, each one is consisted by cascading a 2-D convolution, batch-normalization and ReLu activation. It is the core unit of UNet. The 2-D convolution captures features at current level and an activation layer projects the obtained feature map to a nonlinear feature space.

The pink-color arrows indicate pooling operations. A 2×2 pooling would downscale the original feature map to half of its spatial size, as a result expanding the receptive field for the subsequent double-convolution. As the model goes deeper, the larger receptive field means more global/implicit information of the input data can be captured.

In decoder, the green arrow indicates the upsampling operation to restore the size of feature maps. It is often achieved by bilinear interpolation or deconvolution. Although the above-mentioned pooling operation is beneficial to obtain different levels of features, it will also discard some detailed information. Here, by applying skip-connection, represented by gray arrows, the shallow feature maps are concatenated to deep features

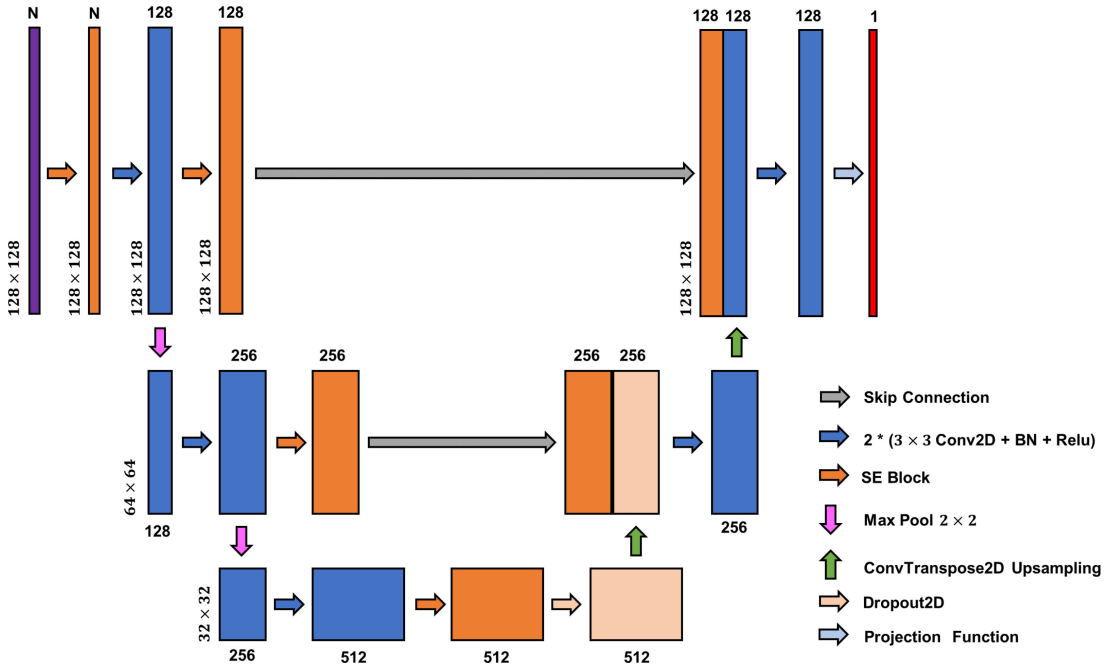

Fig. 1. SeUNet model for regression task with embedded SE blocks.

recovered from upsampling. The final light blue arrow represents a 1×1 convolution projection function, which maps the last feature map to the target space. The 1×1 convolution kernel size preserves the spatial size and enables pixel-level prediction.

B. Modified Models

The basic UNet model has good image modeling capability [35], [36], but due to the characteristics of remote sensing task, especially the forest height regression, certain modifications are necessary. Here we improve the basic model by introducing two key modifications: 1) feature recalibration based on modified SE channel-attention block [40]; 2) semisupervised CPR strategy. We denote these modified models as SeUNet and CPrSeUNet, respectively.

1) SeUNet Model: The convolution operation can acquire and integrate spatial-wise and channel-wise features at the same time. However, the channel dependency of input data needs more exploitation. Firstly, the observation span of satellite image time series often covers a whole year or even longer, the quality of satellite images varies a lot due to the influence of weather or other observation conditions. There are both importance differences and information redundancy among timestamps, which contribute differently to the prediction of forest height. Secondly, although multisensor data increase the feature space, their sensitivity to forest height varies depending on specific wavelength and scattering mechanism.

For these two reasons, we modify the basic UNet model by embedding a modified SE block [40] within SeUNet as shown in Fig. 1. The SE block learns channel-wise self-attention by training [40]. Then, channels of original tensor are scaled and recalibrated based on the attention weights.

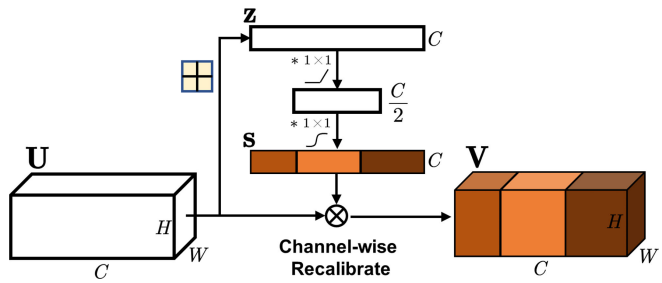

Fig. 2. SE block for channel recalibration within the SeUNet model.

The SE block shown in detail in Fig. 2 operates as follows. Let the dimension of input tensor be $C \times H \times W$, where C is the number of channels, and H and W denote the height and width of a certain channel, respectively. SE block operates in two steps: 1) squeeze 2) excitation.

Squeeze process compresses the original tensor $\boldsymbol{U}$ spatially from $C \times H \times W$ to $C \times 1 \times 1$. It is often accomplished by global average pooling (GAP). The global spatial feature is distilled for each channel along the channel dimension. In details, let $\boldsymbol{U}_c$ denote the feature map of Channel c, and $\mathrm{F}_{\mathrm{gap}}$ denote the GAP function, the squeeze process can be described in

$$z_c = \mathrm{F}_{\mathrm{gap}}\left(\boldsymbol{U}_c\right) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j). \tag{1}$$

Excitation process, on the other hand, introduces the concept of gates to obtain the dependence between channels [40]. Instead of two fully-connected layers in classic SE block, here we use two projection functions (1×1 convolution) to implicitly obtain interchannel self-attention. These two layers of projection

functions are essentially a bottleneck-structure, in which the nonlinear dependencies are expressed by dimensionality reduction. Compared to the original fully connected layers, projection function greatly reduces the amount of parameters and computations without sacrificing the model capability. Excitation process can be expressed mathematically by

$$\boldsymbol{s} = \mathrm{F}_{\mathrm{ex}}(\boldsymbol{z}) = \sigma\left(\boldsymbol{W}_2 \delta\left(\boldsymbol{W}_1 \boldsymbol{z}\right)\right) \tag{2}$$

where $\boldsymbol{z}$ and $\boldsymbol{s}$ are the vectors before and after excitation, respectively, $\boldsymbol{W}_1$, and $\boldsymbol{W}_2$ are two projection functions, δ is the relu activation function, and σ is sigmoid function. The dimension of channel $\boldsymbol{z}$ is first downscaled to $(C/2) \times 1 \times 1$ by projection function $\boldsymbol{W}_1$, followed by a relu layer δ introducing the nonlinearity to the channel-attention. The second $\boldsymbol{W}_2$ then recovers the channel-attention to its original channel size $C \times 1 \times 1$. At last, a sigmoid layer σ projects the channel-attention to the range (0,1) and makes them as the weights of channels.

For a certain Channel c, the original feature map $\boldsymbol{U}_c$ is scaled by channel weight s_c, as shown in (3). In this manner, the whole input tensor is recalibrated according to the channel-attention

$$\boldsymbol{V}_c = \mathrm{F}_{\mathrm{scale}}\left(\boldsymbol{U}_c, s_c\right) = s_c \boldsymbol{U}_c. \tag{3}$$

Within each SE block, the bottleneck ratio of excitation is set to half of the original channel number. SE block would perform feature recalibration at not only the input level but also at each skip connection. When SE block is embedded into input level, channel-attention essentially indicates the significance of input images, which gives us another insight of the feature selection. Considering the relatively small scale of the training dataset, SeUNet is stacked with only three layers. As the model goes deeper, the numbers of convolution kernels of double-convolution are set to 128, 256, and 512, respectively, with the convolution kernel size of 3×3, stride of 1, and padding of 1. Before each skip connection, another SE module is used to recalibrate the feature map obtained by convolution. We also introduce a 2-D-dropout to the deepest layer to bring more regularization to the feature map. At last, the deconvolution is selected as our upsampling operation. In order to reduce the tessellation effect caused by the deconvolution, the deconvolution kernel size is 2, the same as the stride size.

2) CPrSeUNet Model: DL models require significant amount of data for successful training, while lack of ground truth data typical for EO-based forest inventory can negatively affect prediction performance of modeling approaches. Here, we propose the CPR strategy to address this issue. Both labeled and unlabeled data are utilized to train the model. Cross-pseudo supervision (CPS) has been proposed by [33] for semantic segmentation task. To adopt CPS into pixel-level prediction of a continuous forest structural attribute, we use a regression head instead of the segmentation head. Another difference lays within the training batch formation. Considering the mixed-up labeled and unlabeled patches for training, the unlabeled patches are designed to have a higher probability of appearance within early training epochs. Similar to two-step training, where unlabeled patches are used for pretraining and the labeled ones for fine-tuning, this improves the network convergence, especially within final training epochs. Compared to using only labeled training data, the suggested method essentially augments the original dataset and further improves the performance of the model. We call this hybrid model a CPrSeUNet model. Further, we describe the approach in more detail.

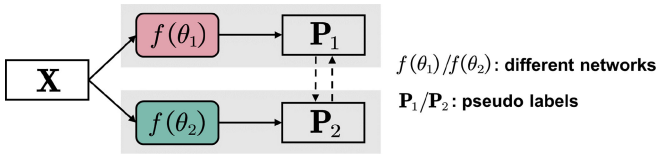

Fig. 3. CPR strategy.

Firstly, we apply the concept of consistency learning on two perturbed network branches. Both branches have the same structure but vary by initialization. We encourage the model to give similar predictions, even though some perturbations are introduced, which means the representative capability of the model is compact enough to hold the perturbations. Given the two network branches as M_1 and M_2, when fed into the same input $\boldsymbol{X}$, the two branches can be denoted as

$$\begin{aligned} \boldsymbol{P}_1 &= f\left(\boldsymbol{X}; \boldsymbol{\theta}_1\right) \\ \boldsymbol{P}_2 &= f\left(\boldsymbol{X}; \boldsymbol{\theta}_2\right) \end{aligned} \tag{4}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ denote the initial weights of M_1 and M_2, respectively, $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ denote the regression results, which in our case are pixel-level forest height predicted by SeUNets.

The flowchart is shown in Fig. 3, where solid arrow represents the forward computation, and dash arrows represent the flow of supervision information. We take all the data as the inputs, including both labeled and unlabeled. When the same input data are fed into two perturbed branches, each branch yields its own regression results $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$.

We first ignore all label information, and treat all data as unlabeled. We take the regression result of one branch as the pseudo-reference of the other. That is, $\boldsymbol{P}_1$ serves as the pseudo-reference of branch M_2, and vice versa. In this way, both network branches can be updated by back-propagation. We select pixel-wise the mean-squared error (MSE) between the regression results and reference as our basic loss

$$\mathrm{Loss}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \tag{5}$$

where $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ are the prediction and reference, respectively, y_i is the ith element in $\boldsymbol{Y}$, and n is the total number of the elements. Considering $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ have the same number of pixels, the joint loss of both branches, which is named as cross-pseudo loss, can be represented in

$$\ell_c = \mathrm{Loss}\left(\boldsymbol{P}_2, \boldsymbol{P}_1\right) + \mathrm{Loss}\left(\boldsymbol{P}_1, \boldsymbol{P}_2\right). \tag{6}$$

Then, the label information is taken into consideration. For labeled part, we normally use its own ground truth $\boldsymbol{R}$ as the reference. The supervised joint loss of this part of data is formulated as

$$\ell_s = \mathrm{Loss}\left(\boldsymbol{P}_1', \boldsymbol{R}\right) + \mathrm{Loss}\left(\boldsymbol{P}_2', \boldsymbol{R}\right). \tag{7}$$

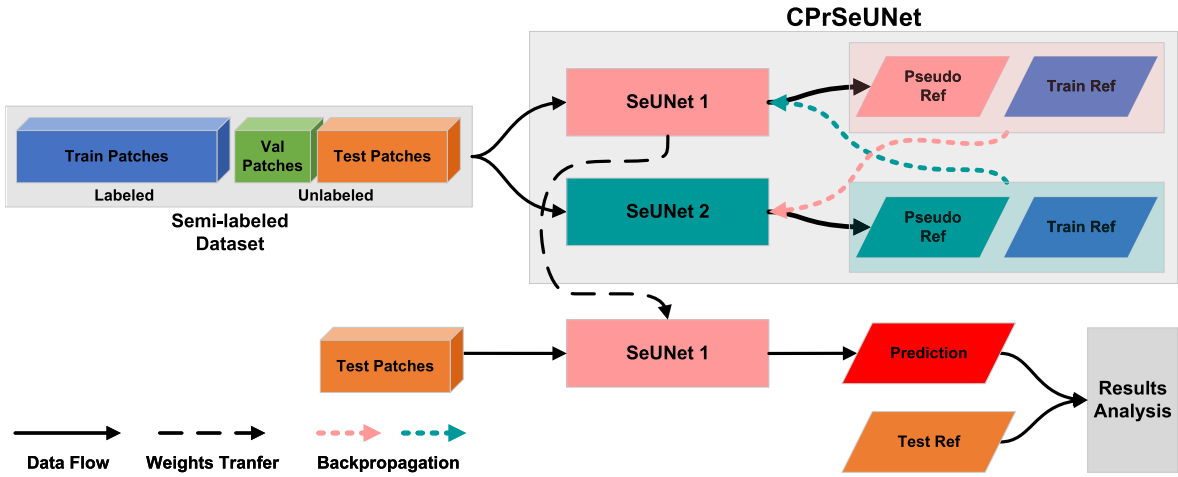

Fig. 4. CPrSeUNet model showing two parallel SeUNet models in the context of semisupervised regression.

Finally, we summarize all the losses with both include- and exclude-label information as

$$\ell = \ell_s + \lambda_c \ell_c + \lambda_w \frac{1}{n_w} \sum_{j=1}^{n_w} (w_j)^2 \tag{8}$$

where λ_c is defined as the tradeoff weight to control the impact of cross-pseudo loss. According to the cross-validation, we set λ_c as 0.5 in our model. To mitigate the overfitting, a weight decay (l_2 penalty) item is also introduced as a regularization to the model weights, where w_j is the jth weight, n_w is the total number of weights, and the strength of regularization is controlled by λ_w.

CPR strategy helps train the model in a semisupervised way, unlabeled data is successfully mixed up with labeled on training stage, as a result improving the capability of the model. As shown in Fig. 4, our entire model is composed of two perturbed SeUNet branches with the same structure but differing only in initialization. On the prediction stage, we only need to extract one branch as our main model for the regression task without any further fine-tuning.

III. Materials and Methods

In this article, various combinations of EO images represented by time series of Sentinel-1 SAR images and Sentinel-2 image bands are used for forest height prediction with baseline UNet and modified DL models, as well as with other established machine learning approaches. The baseline and developed models are described in Section III-B. The main goal of the study was to develop and fine-tune a reliable and highly accurate DL model for predicting forest tree height using multidimensional EO dataset, and validate its performance against other modeling approaches.

The overall approach is illustrated in Fig. 5. SAR and optical images and reference data are described in Section III-A. Reference data are split into geographically nonoverlaping subsets for model training and validation, and testing (accuracy assessment). Details of experimental approach are described in Section III-C.

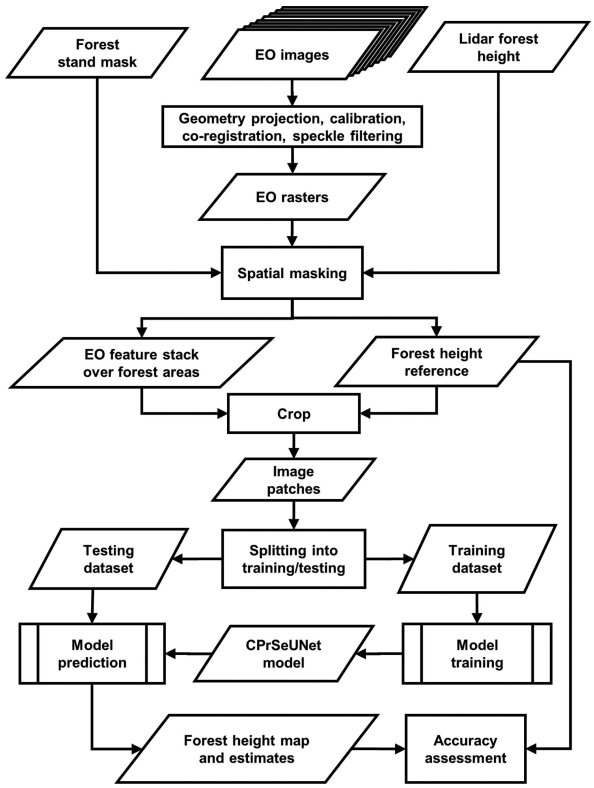

Fig. 5. Image processing chain for producing forest height map using EO images.

A. Study Site, Satellite SAR, Optical and Reference Data

The study site is located in Central Finland in the vicinity of Hyytiälä forestry station and occupies an area of $2500\,\mathrm{km}^2$. The area represents a typical mixed boreal forestland with forest growing stock volume up to $500\,\mathrm{m}^3/\mathrm{ha}$ and $170\,\mathrm{m}^3/\mathrm{ha}$ on average. The location of our study site and natural color composite of Sentinel-2 image is shown in Fig. 6

SAR data are represented by a time series of Sentinel-1 images acquired during 2015. Altogether 27 dual-pol Sentinel-1 A images available as ground range detected products were used.

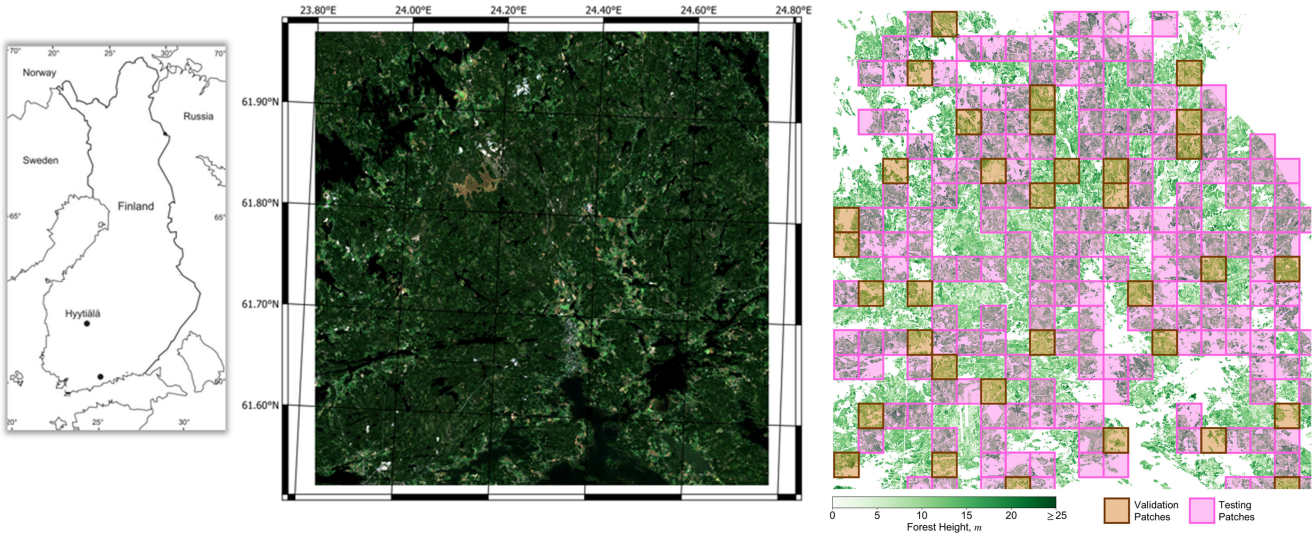

Fig. 6. Study site location in Finland (left image), optical Sentinel-2 scene (centre), and reference ALS data with sampling design (left image).

The images were radiometrically terrain-flattened and orthorectified with in-house software using local digital elevation model available from National Land Survey of Finland [41], [42]. Final preprocessed images were in gamma-naught format [43].

Satellite optical data were represented by an ESA (European Space Agency) Sentinel-2 image acquired in August 2015. The Level-2A surface reflectance product systematically generated by ESA and distributed in tiles of 100×100 km^2 was used. The multispectral instrument on board Sentinel-2 satellites has 13 spectral bands with 10-m (four bands), 20-m (six bands), and 60-m (three bands) spatial resolutions. Only 10 m bands were used in the experiments, altogether four bands.

Airborne laser scanning (ALS) data collected by National Land Survey of Finland during summer of 2015 were used as a reference. Forest heights were computed from ALS point clouds as average elevation of forest classified points over ground layer within $20 \times 20\,\mathrm{m}^2$ pixel cells. In this way, a wall-to-wall coverage of the test site with the reference height information was obtained. Forest stand-mask from the Finnish Forest Centre was used to calculate stand-level estimates of tree height from the forest height map.

B. Baseline Models

The utility of developed DL models described in Section II is demonstrated using comparison with several conventional machine learning (ML) approaches popular in satellite-based forest inventory [5], as well as the baseline UNet model. Three ML models are considered: 1) Multiple Linear Regression (MLR), random forest (RF), and Light Gradient Boosting Machine (LightGBM or LGBM) [44]. MLR and RF have been widely used in forest remote sensing utilizing EO data [4]. LightGBM is a modern Gradient Boosting Decision Tree tool developed by Microsoft with demonstrated advantages in regression tasks. All these regression models operate on pixel-level.

For MLR, the data processing pipeline also includes principal component analysis (PCA) to reduce the number of independent input variables in regression models. When the size of the input feature set is large (for example 54 features of Sentinel-1 dataset), only ten primary components are extracted by PCA before MLR regression.

In RF and LightGBM, feature selection mechanism is already built in the model. Additionally, model fine-tuning based on five-fold cross validation is performed to achieve better prediction accuracy. Considering the large number of training samples, we randomly select more than 10% as its fine-tuning subset, in order to alleviate the computational workload. The exhaustive hyperparameter search is conducted for 100 trials using Optuna, an automatic hyperparameter optimization framework.

To the best of our knowledge, the proposed CPR strategy is the first example of using semisupervised learning for forest variable prediction, tree height in our study case. To estimate its potential, two well-known semisupervised strategies previously used for segmentation/change-detection [27], [45] are converted to regression task as additional baseline models: 1) *reconstruction-based two-step training;* 2) *Siamese-network-based consistency learning*. To make the comparison fair, both strategies use SeUNet as their backbone model, similar to the proposed CPrSeUNet. We further call these models Rec-SeUNet and Sia-SeUNet. They both consist of two training steps: 1) pretraining with unlabeled data; 2) fine-tuning with labeled training data. The difference mainly lies within the pretraining procedure.

In Rec-SeUNet, a reconstruction head is firstly placed on top of the backbone model, and the model gets pretrained by means of reconstructing the original unlabeled inputs. Specifically, given $\boldsymbol{X}$ denotes the input tensor and f denotes the model, the reconstruction can be denoted as $\hat{\boldsymbol{X}} = f(\boldsymbol{X}; \boldsymbol{\theta})$, where θ is the model parameter. The input data itself are used as the reference, and the pixel-wise MSE$(\boldsymbol{X}, \hat{\boldsymbol{X}})$ is calculated as the loss during model optimization. In this unsupervised manner,

the pretrained backbone is assumed to learn the representations of the input data [27]. Then on the second step, we fine-tune the model by replacing the reconstruction head with a regression head. Not only the original labeled, but also the pseudo-labeled data are used in the fine-tuning. The pseudolabels are generated from unlabeled datasets iteratively in each epoch.

In Sia-SeUNet, we leverage the pretraining by using Siamese network [46]. It consists of two subnets that share the same structure and parameters. Unlabeled input data $\boldsymbol{X}$ firstly undergo heavy pixel-wise augmented (e.g., *GaussianNoise*, *Blur*). Then, $\boldsymbol{X}$ and the augmented data $\boldsymbol{X}_{\text{aug}}$ are fed into different subsets. Let $\boldsymbol{y_1} = f(\boldsymbol{X}; \boldsymbol{\theta})$ and $\boldsymbol{y_2} = f(\boldsymbol{X_{aug}}; \boldsymbol{\theta})$ denote the predictions of the corresponding subnets. We pretrain the subnets by minimizing the pixel-wise MSE$(\boldsymbol{y_1}, \hat{\boldsymbol{y_2}})$, which is also known as consistency loss according to consistency learning concept. In this way, a more compact representation of the data is obtained. On the following fine-tuning step, we extract one of the subnets as the prediction model, and use the (labeled) training data to fine-tune it.

C. Experimental Setup

In this section, we describe the preprocessing of EO data for training, validation, and testing, and give detail. Preprocessed Sentinel-1 images are stacked together with Sentinel-2 optical image bands to form the input EO raster image stack.

1) Training, Validation, and Testing Datasets: The original SAR and optical images were firstly cropped into $128\,\text{px} \times 128\,\text{px}$ image patches. Nonforested regions were masked out by setting both feature values and predictor values to zero. The cropping is performed using a sliding window with the stride of $128\,\text{px}$ so that cropped image patches have no spatial overlap. To avoid excessive proportion of nonforested area within each patch, a threshold value of 0.2 was set for forest cover proportion, with less-forested areas removed from further analysis.

In such a way, altogether 340 nonoverlapping patches were cropped from the original image data stack. Further, these patches were split into three subsets using random sampling, as shown in Fig. 6 The areas shown in pink (170 image patches) were set aside for testing, the yellow colored areas (34 patches) were reserved for validation, while remaining areas (136 patches) were targeted for training. Thus, the ratio of training and validation areas to testing areas is approximately 1:1.

However, to make better use of the spatial context at the borders of adjacent training patches, we did not use the mentioned 136 training patches directly, but performed dense cropping over areas these patches occupy. All other areas were masked out. The dense cropping was performed using the same $128\,\text{px} \times 128\,\text{px}$ sliding window but with a smaller stride of $64\,\text{px}$. This recropping can be in fact considered as a kind of data augmentation: spatial shifting. After applying forest-cover threshold of 0.2 and additional augmentations, such as image rotation and flipping, the final number of training patches was 716.

Denoting the dimensions of a dataset as $N \times C \times H \times W$, where C is the quantity of channels, H and W are the height and width of each patch, respectively, and N is the number of patches, we have $C = 58$, $H = W = 128$, and $N \in \{716, 34, 170\}$ for training, validation, and testing subsets, respectively.

For pixel-based approaches, such as MLR, RF, and LGBM, pixel-wise feature vectors were compiled from areas corresponding to each subset. Specifically, denoting the dimensions of a subset as $N \times C$, we obtained $C = 58$, $N = 1,882,330$ for training and validation subset, $N = 1,778,777$ for testing.

2) Experimental Implementation: Our experimental platform was a 64-bit windows 2012 server, with a single GPU (NVIDIA Tesla P40) and 32 GB RAM. Our model was built in Python 3.8 with PyTorch 1.9.0 as its backend. The general processing flowchart is shown in Fig. 5. We firstly train the model in a semisupervised way. Regarding the loss calculation, the tradeoff weight λ_c was set as 0.5 and the weight decay was $1e^{-4}$. The whole model is trained by using two separate Adam optimizers on each branch. The initial learning rate was set as $5e^{-4}$, while a cosine annealing schedule was used to adjust the learning rate during the training. With the batch size of ten, we ran the training stage for 500 epochs, updating the model checkpoint to a lower validation loss. On the testing stage, we extracted one branch (that is M_1) from the final checkpoint as our final model, and applied it for the forest height prediction on testing data set followed by accuracy assessment of produced forest height map.

D. Accuracy Assessment

Accuracy assessment of the model predictions is performed on pixel level as well as on stand level between the predictions and the reference, using the following equations:

$$\text{RMSE} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \tag{9}$$

$$\text{RMSE}\% = \frac{RMSE}{\bar{y}} \cdot 100\% \tag{10}$$

$$\text{MAE} = \frac{\sum_i |y_i - \hat{y}_i|}{n} \tag{11}$$

$$\text{BIAS} = \frac{\sum_i (\hat{y}_i - y_i)}{n} \tag{12}$$

$$\mathrm{R}^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y_i})^2} \tag{13}$$

where y_i and $\hat{y}_i$ are model predictions of forest height and measured values from the reference forest inventory data, respectively, and n is the total number of measurements.

IV. Results and Discussion

Here, we perform assessment of relative performance of various forest tree height prediction methods and models, as well as on comparing added value of imaging radar and optical datasets. Further, we report and analyze obtained results with respect to these two primary aspects.

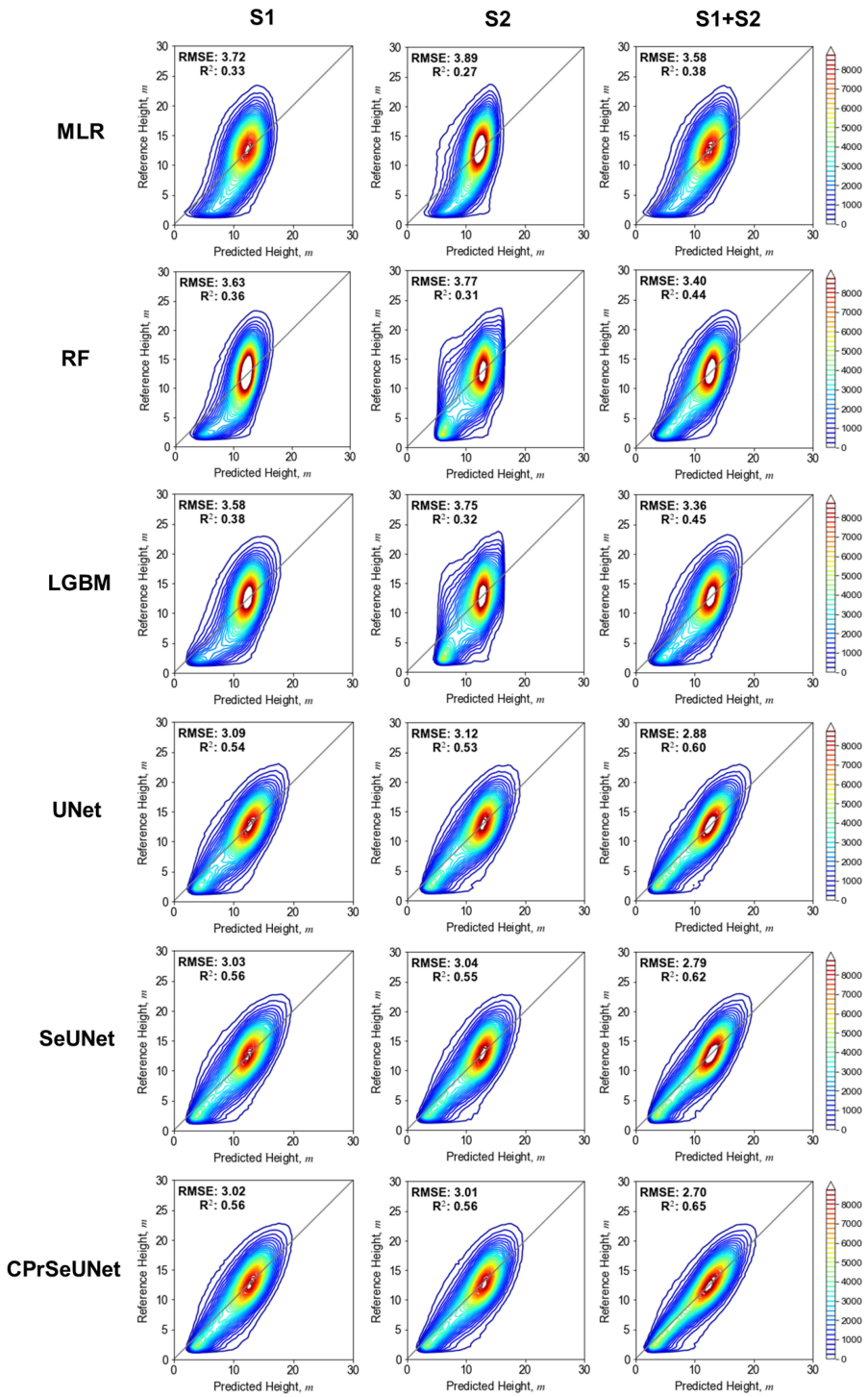

Fig. 7. Dependence of predicted versus reference forest heights on map unit (pixel) level for various EO datasets (S1 = ESA Sentinel-1 time series, S2 = ESA Sentinel-2 image).

A. *Model Prediction Performance*

Performance of studied models with both optical Sentinel-2 data and multitemporal Sentinel-1 data is illustrated with scatterplots shown in Fig. 7 for pixel-level predictions and in Fig. 8 for stand-level predictions. The quantitative performance is summarized in Tables I and II, respectively. An example of patch-level prediction within 2.56 km×2.56 km area for various models is shown along with the reference data in Fig. 9.

1) Relative Performance of Different Datasets in Forest Variable Prediction and Optimal Data Fusion: For Sentinel-1-based predictions, only time series results are shown, that is when all 27 images are used. Performance of methods with single SAR images were checked only with selected methods (MLR and RF), and was poorer than with the time series, ranging from 38.8 to 40.7% when MLR was applied for tree height prediction using individual SAR images, and 38.3–40.9% with RF. Best prediction was achieved with Sentinel-1 image acquired on

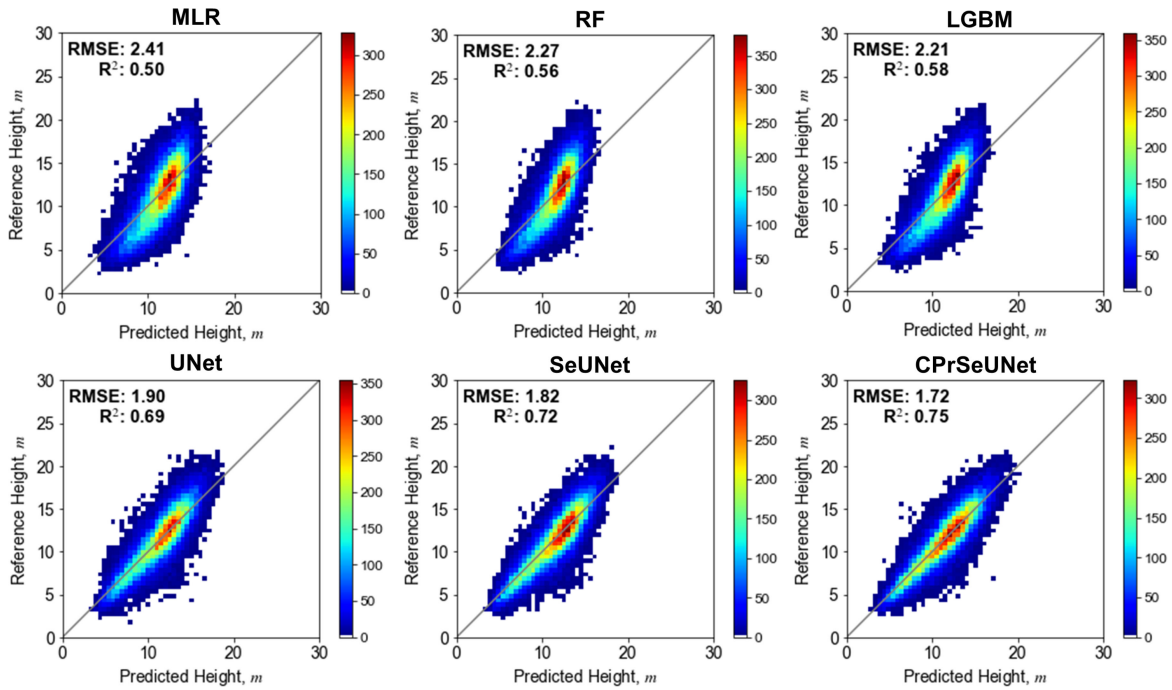

Fig. 8. Dependence of predicted versus reference forest heights on forest stand-level for various ML and DL models using combined SAR and optical datasets (multitemporal Sentinel-1 and Sentinel-2 images).

TABLE I
ACCURACY ASSESSMENT USING PIXEL-LEVEL DATA

	RMSE, m	RMSE%	MAE	Bias	R^2
Sentinel-1 time series					
MLR	3.72	33.28	2.95	0.00	0.33
RF	3.63	32.50	2.89	0.01	0.36
LGBM	3.58	32.08	2.81	-0.01	0.38
UNet	3.09	27.68	2.36	0.06	0.54
SeUNet	3.03	27.14	2.29	-0.13	0.56
CPrSeUNet	3.02	27.03	2.29	0.05	0.56
Sentinel-2					
MLR	3.89	34.80	3.08	0.01	0.27
RF	3.77	33.79	2.99	0.03	0.31
LGBM	3.75	33.61	2.98	0.02	0.32
UNet	3.12	27.90	2.36	-0.05	0.53
SeUNet	3.04	27.21	2.29	-0.12	0.55
CPrSeUNet	3.01	26.98	2.26	0.11	0.56
Sentinel-1 time series and Sentinel-2					
MLR	3.58	32.04	2.79	-0.03	0.38
RF	3.40	30.44	2.66	0.01	0.44
LGBM	3.36	30.09	2.61	-0.02	0.45
UNet	2.88	25.76	2.14	-0.17	0.60
SeUNet	2.79	25.00	2.07	-0.02	0.62
CPrSeUNet	2.70	24.14	1.96	-0.07	0.65

TABLE II
ACCURACY ASSESSMENT USING STAND-LEVEL DATA

	RMSE, m	RMSE%	MAE	Bias	R^2
Sentinel-1 time series					
MLR	2.53	22.66	2.01	0.06	0.45
RF	2.51	22.49	2.01	0.05	0.46
LGBM	2.40	21.49	1.90	0.04	0.51
UNet	2.09	18.76	1.60	0.07	0.62
SeUNet	2.03	18.23	1.53	-0.10	0.64
CPrSeUNet	2.02	18.10	1.53	0.03	0.65
Sentinel-2					
MLR	2.72	24.35	2.16	-0.03	0.37
RF	2.55	22.80	2.02	-0.02	0.44
LGBM	2.55	22.81	2.02	-0.03	0.44
UNet	2.13	19.09	1.62	-0.01	0.61
SeUNet	2.07	18.50	1.56	-0.09	0.63
CPrSeUNet	2.03	18.15	1.52	0.09	0.65
Sentinel-1 time series and Sentinel-2					
MLR	2.41	21.60	1.88	-0.00	0.50
RF	2.27	20.33	1.78	0.01	0.56
LGBM	2.21	19.78	1.72	-0.00	0.58
UNet	1.90	17.06	1.41	-0.13	0.69
SeUNet	1.82	16.30	1.34	0.01	0.72
CPrSeUNet	1.72	15.38	1.25	-0.03	0.75

2015/03/02, and the worst prediction with image acquired on 2015/05/13. Performance improved as more images were used. This agrees well with published literature, e.g., [5] and [47].

Accuracy of Sentinel-2-based predictions suggests better suitability of Sentinel-2 image compared to a single Sentinel-1 SAR image, but proved somewhat inferior compared to predictions based on Sentinel-1 time series. Generally, collecting Sentinel-1 time series requires long observation periods, however Sentinel-2 data can be of even smaller availability, as in certain years it was hard to collect suitable optical images [48], something that can be improved in future with the help of smallsats. When both SAR and optical datasets were used, all models have shown improvement compared to using only one

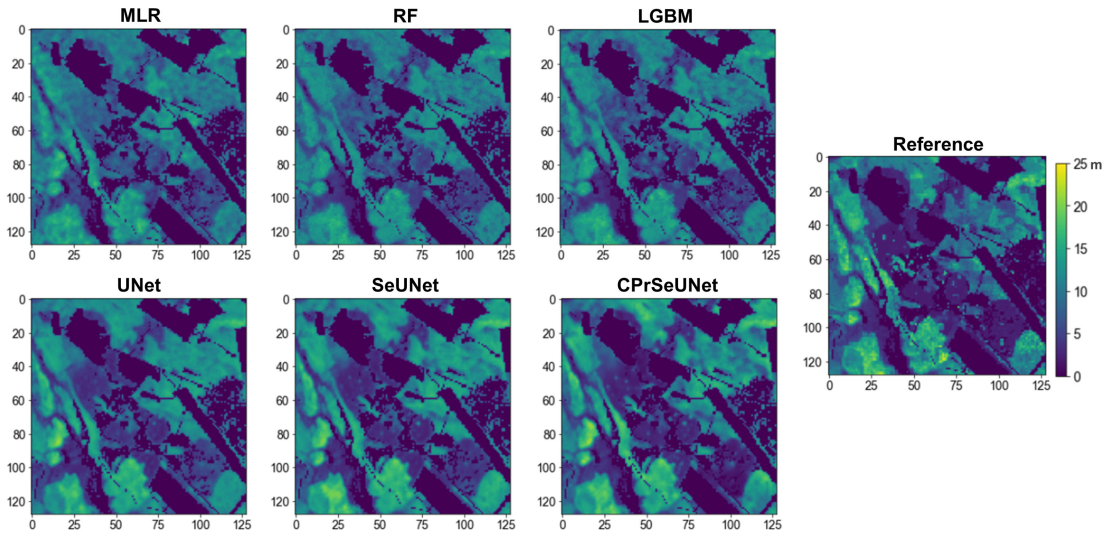

Fig. 9. Example of forest height map at the extent of one image patch produced using various models, image size 2.56 km×2.56 km.

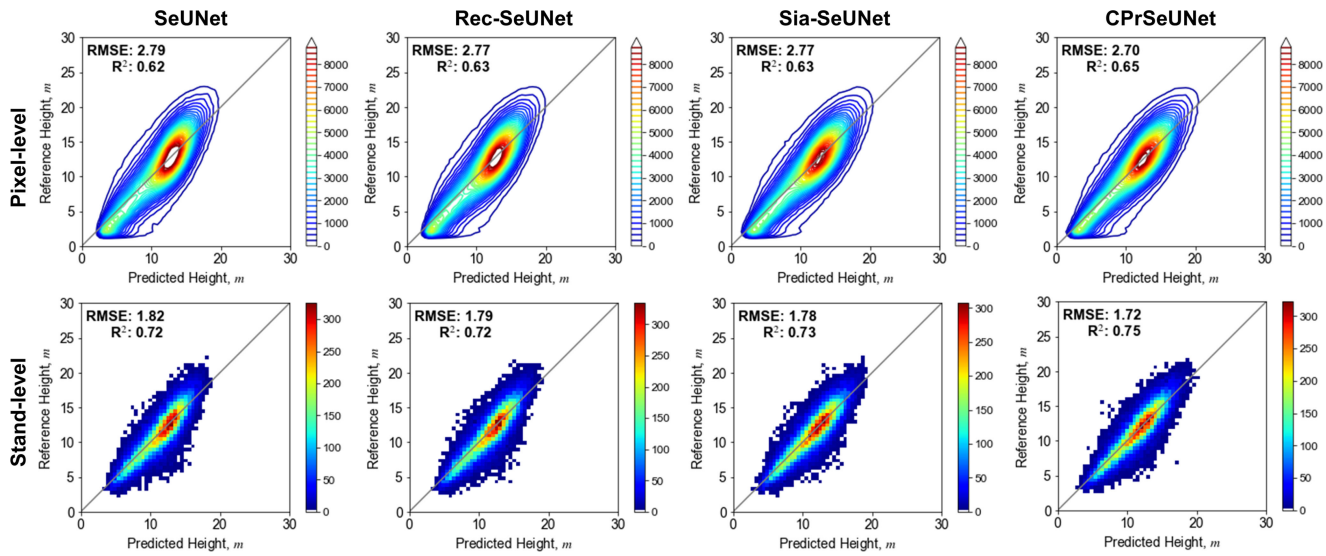

Fig. 10. Dependence of predicted versus reference forest heights on forest stand-level for semisupervised methods using combined SAR and optical datasets (multitemporal Sentinel-1 and Sentinel-2 images).

source of EO data, reaching 24% RMSE in the best case with CPrSeUNet model.

2) DL Model Versus Conventional ML: Developed CPrSeUNet model provided more accurate predictions than baseline UNet and SeUNet models. Generally DL models that account for spatial context in addition to spectral features were more accurate than pixel-based approaches, with improvement of some 6–7% in RMSE and much higher R^2. Improved performance of proposed CPrSeUNet can be attributed to channel-wise self-calibration and expanded dataset size due to CPR approach.

3) Performance Versus Semi-Supervised Methods: Here, we compare the proposed CPrSeUNet model versus two other semisupervised approaches, as well as versus basic SeUNet without self-supervision. The performance metrics are gathered in Table III and scatter plots are shown in Fig. 10. For all studied models, semisupervised methods provide better results over the backbone SeUNet, indicating that semisupervision has potential in improving prediction accuracy of pixel-wise forest variable regression. However, the improvements of Rec-SeUNet and Sia-SeUNet are relatively modest compared to basic SeUNet model. Predictions by Sia-SeUNet are slightly more accurate than Rec-SeUNet at stand level. The possible reason is the heavy augmentation before consistency learning can better handle heterogeneous pixels, thus improve the accuracy within stands. However, this perturbation depends on the choice of pixel-wise augmentation algorithm, and several trials are typically necessary to get optimal performance. Combining both pseudo-labels and consistency learning simultaneously within one model, CPrSeUNet shows the best performance of 15.38%

TABLE III
ACCURACY ASSESSMENT OF SEMISUPERVISED METHODS

	Pixel-level			Stand-level		
	RMSE, m	RMSE%	R^2	RMSE, m	RMSE%	R^2
SeUNet	2.79	25.00	0.62	1.82	16.30	0.72
Rec-SeUNet	2.77	24.77	0.63	1.79	16.07	0.72
Sia-SeUNet	2.77	24.80	0.63	1.78	15.98	0.73
CPrSeUNet	2.70	24.14	0.65	1.72	15.38	0.75

at stand-level, as well as the highest R^2, as shown in Table III and Fig. 10. Noteworthy, in CPrSeUNet the perturbation is applied for initialization of model branches, unlike manual augmentation required in the Sia-SeUNet case.

B. Comparison to Similar Work

Obtained results compare favourably to previous studies on forest height prediction. In boreal region, reported forest height accuracies with Sentinel-2 and Landsat data were 35–60% RMSE [19], [49], while proposed DL model reached as small as 24% RMSE on plot level and 15.4% on stand level. Obtained predictions with ML models and Sentinel-2 data are within the same accuracy range as in recent published studies using Sentinel-2 and Landsat [50], while our predictions using DL models are much more accurate. There is relatively limited literature using SAR data for forest height predictions, with those normally done on stand level, but our obtained stand-level tree height predictions are at the same accuracy level or even better than reported retrievals with TanDEM-X interferometric SAR data deemed much more suitable for vertical forest structure retrieval [51]–[53]. To the best of our knowledge, obtained accuracy levels of 15–17% RMSE for boreal forest using DL models and combined SAR and optical data are superior to earlier results reported in literature [4]. Obtained results naturally encourage further studies with semisupervised DL models for forest mapping using both similar EO datasets, as well as more advanced imagery, such as polarimetric, interferometric, and multifrequency SAR datasets, over boreal forests and other forest biomes.

V. CONCLUSION

Our study demonstrated the potential for applying semisupervised DL models for predicting forest height. A dedicated DL model was developed and benchmarked with a representative set of machine learning and DL methodologies over a boreal forest site in Finland. Developed model is based on UNet and can handle both SAR and optical datasets, and has shown improved prediction accuracy compared to baseline approaches. All studied approaches demonstrated advantage of combining SAR and optical datasets to achieve better accuracy of forest height retrieval. Future work will concentrate on introducing other datasets particularly suitable for retrieving vertical structure of forests, such as Sentinel-1 repeat-pass interferometric SAR and TanDEM-X bistatic inerferometric SAR datasets, as well as studying other forest variables, such as growing stock volume.

V. DATA AVAILABILITY

EO data used in the study are available via IEEE Data Port, while original Sentinel-1 and Sentinel-2 images are freely available via Copernicus Open Hub. Reference data are available from the corresponding author upon reasonable request.

REFERENCES

[1] R. E. McRoberts, E. Næsset, C. Sannier, S. V. Stehman, and E. O. Tomppo, "Remote sensing support for the gain-loss approach for greenhouse gas inventories," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1891.

[2] E. Tomppo, H. Olsson, G. Ståhl, M. Nilsson, O. Hagner, and M. Katila, "Combining national forest inventory field plots and remote sensing data for forest databases," *Remote Sens. Environ.*, vol. 112, no. 5, pp. 1982–1999, 2008.

[3] R. E. McRoberts and E. O. Tomppo, "Remote sensing support for national forest inventories," *Remote Sens. Environ.*, vol. 110, no. 4, pp. 412–419, 2007.

[4] GFOI, *Integrating Remote-Sensing and Ground-Based Observations for Estimation of Emissions and Removals of Greenhouse Gases in Forests: Methods and Guidance From the Global Forest Observations Initiative*. Group on Earth Observations, Geneva, Switzerland, 2014.

[5] S. Ge et al., "Using hypertemporal Sentinel-1 data to predict forest growing stock volume," *bioRxiv*, 2021, doi: 10.1101/2021.09.02.458789.

[6] G. V. Laurin et al., "Above-ground biomass prediction by Sentinel-1 multitemporal data in central Italy with integration of ALOS2 and Sentinel-2 data," *J. Appl. Remote Sens.*, vol. 12, 2018, Art. no. 016008.

[7] M. A. Stelmaszczuk-Górska, M. Urbazaev, C. Schmullius, and C. Thiel, "Estimation of above-ground biomass over boreal forests in Siberia using updated in situ, ALOS-2 PALSAR-2, and RADARSAT-2 data," *Remote Sens.*, vol. 10, no. 10, 2018, Art. no. 1550.

[8] Y. Li, M. Li, C. Li, and Z. Liu, "Forest aboveground biomass estimation using landsat 8 and Sentinel-1 A data with machine learning algorithms," *Sci. Rep.*, vol. 10, no. 1, 2020, Art. no. 9952.

[9] T. Hame et al., "Improved mapping of tropical forests with optical and SAR imagery, Part I: Forest cover and accuracy assessment using multiresolution data," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 6, no. 1, pp. 74–91, Feb. 2013.

[10] S. Englhart, V. Keuck, and F. Siegert, "Modeling aboveground biomass in tropical forests using multi-frequency SAR data–A comparison of methods," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 5, no. 1, pp. 298–306, Feb. 2012.

[11] C. Schmullius, C. Thiel, C. Pathe, and M. Santoro, *Radar Time Series for Land Cover and Forest Mapping*. Cham, Switzerland: Springer, 2015, pp. 323–356.

[12] J. Esteban, R. E. McRoberts, A. Fernández-Landa, J. L. Tomé, and E. Naesset, "Estimating forest volume and biomass and their changes using random forests and remotely sensed data," *Remote Sens.*, vol. 11, no. 16, 2019, Art. no. 1944.

[13] X. X. Zhu et al., "Deep learning meets SAR: Concepts, models, pitfalls, and perspectives," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 143–172, Dec. 2021.

[14] S. Scepanovic, O. Antropov, P. Laurila, Y. A. Rauste, V. Ignatenko, and J. Praks, "Wide-area land cover mapping with Sentinel-1 imagery using deep learning semantic segmentation models," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 10357–10374, 2021.

[15] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.

[16] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 59–69, 2019.

[17] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS J. Photogrammetry Remote Sens.*, vol. 173, pp. 24–49, 2021.

[18] Y. Guo et al., "A deep fusion UNet for mapping forests at tree species levels with multi-temporal high spatial resolution satellite imagery," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3613.

[19] H. Astola, L. Seitsonen, E. Halme, M. Molinier, and A. Lönnqvist, "Deep neural networks with transfer learning for forest variable estimation using Sentinel-2 imagery in boreal forest," *Remote Sens.*, vol. 13, no. 12, 2021, Art. no. 2392.
[20] L. Zhang, Z. Shao, J. Liu, and Q. Cheng, "Deep learning based retrieval of forest aboveground biomass from combined LiDAR and landsat 8 data," *Remote Sens.*, vol. 11, no. 12, 2019, Art. no. 1459.
[21] Z. Shao, L. Zhang, and L. Wang, "Stacked sparse autoencoder modeling using the synergy of airborne LiDAR and satellite optical and SAR data to map forest above-ground biomass," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 10, no. 12, pp. 5569–5582, Dec. 2017.
[22] N. Lang, K. Schindler, and J. D. Wegner, "Country-wide high-resolution vegetation height mapping with Sentinel-2," *Remote Sens. Environ.*, vol. 233, 2019, Art. no. 111347.
[23] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, pp. 1–40, 2016.
[24] Z. Huang, Z. Pan, and B. Lei, "Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data," *Remote Sens.*, vol. 9, no. 9, 2017, Art. no. 907.
[25] M. Rostami, S. Kolouri, E. Eaton, and K. Kim, "Deep transfer learning for few-shot SAR image classification," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1374.
[26] F. Gao, Y. Yang, J. Wang, J. Sun, E. Yang, and H. Zhou, "A deep convolutional generative adversarial networks (DCGANs)-based semi-supervised method for object recognition in synthetic aperture radar (SAR) images," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 846.
[27] C. Wang, H. Gu, and W. Su, "SAR image classification using contrastive learning and pseudo-labels with limited data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 4012505.
[28] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high-and low-level consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1369–1379, Apr. 2021.
[29] B. Zoph et al., "Rethinking pre-training and self-training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 3833–3845.
[30] Y. Zhu et al., "Improving semantic segmentation via efficient self-training," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2021.3138337.
[31] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6728–6736.
[32] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12674–12684.
[33] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2613–2622.
[34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[35] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.
[36] R. Li et al., "DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 11, no. 11, pp. 3954–3962, Nov. 2018.
[37] L. Jiao, L. Huo, C. Hu, and P. Tang, "Refined UNet: UNet-based refinement network for cloud and shadow precise segmentation," *Remote Sens.*, vol. 12, no. 12, 2020, Art. no. 2001.
[38] K. Trebing, T. Stanczyk, and S. Mehrkanoon, "SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture," *Pattern Recognit. Lett.*, vol. 145, pp. 178–186, 2021.
[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.
[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[41] Y. Rauste, A. Lonnqvist, M. Molinier, J.-B. Henry, and T. Hame, "Orthorectification and terrain correction of polarimetric SAR data applied in the ALOS/Palsar context," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2007, pp. 1618–1621.
[42] T. Häme et al., "Enabling intelligent copernicus services for carbon and water balance modeling of boreal forest ecosystems - North State," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 2048–2051.
[43] D. Small, "Flattening gamma: Radiometric terrain correction for SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 8, pp. 3081–3093, Aug. 2011.
[44] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3146–3154.
[45] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, "PICOCO: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 10548–10559, 2021.
[46] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 539–546.
[47] E. Tomppo, O. Antropov, and J. Praks, "Boreal forest snow damage mapping using multi-temporal Sentinel-1 data," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 384.
[48] H. Balzter, B. Cole, C. Thiel, and C. Schmullius, "Mapping CORINE land cover from Sentinel-1 A SAR and SRTM digital elevation model data using random forests," *Remote Sens.*, vol. 7, no. 11, pp. 14876–14898, 2015.
[49] J. Miettinen et al., "Demonstration of large area forest volume and primary production estimation approach based on Sentinel-2 imagery and process based ecosystem modelling," *Int. J. Remote Sens.*, vol. 42, no. 24, pp. 9467–9489, 2021.
[50] H. Astola, T. Häme, L. Sirro, M. Molinier, and J. Kilpi, "Comparison of sentinel-2 and landsat 8 imagery for forest variable prediction in boreal region," *Remote Sens. Environ.*, vol. 223, pp. 257–273, 2019.
[51] J. Praks, M. Hallikainen, O. Antropov, and D. Molina, "Boreal forest tree height estimation from interferometric TanDEM-X images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2012, pp. 1262–1265.
[52] A. Olesk, J. Praks, O. Antropov, K. Zalite, T. Arumäe, and K. Voormansik, "Interferometric SAR coherence models for characterization of hemiboreal forests using TanDEM-X data," *Remote Sens.*, vol. 8, no. 9, 2016, Art. no. 700.
[53] F. Kugler, S.-K. Lee, I. Hajnsek, and K. P. Papathanassiou, "Forest height estimation by means of Pol-InSAR data inversion: The role of the vertical wavenumber," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5294–5311, Oct. 2015.