Crop mapping from image time series: deep learning with multi-scale label hierarchies

Mehmet Ozgur Turkoglu[a], Stefano D'Aronco[a], Gregor Perich[b], Frank Liebisch[b,c], Constantin Streit[d], Konrad Schindler[a], Jan Dirk Wegner[a,e]

[a]*EcoVision Lab, Photogrammetry and Remote Sensing, ETH Zurich, Switzerland*
[b]*Crop Science, ETH Zurich, Switzerland*
[c]*Agroecology and Environment, Agroscope, Switzerland*
[d]*Federal Office for Agriculture, Switzerland*
[e]*Institute for Computational Science, University of Zurich, Switzerland*

Abstract

The aim of this paper is to map agricultural crops by classifying satellite image time series. Domain experts in agriculture work with crop type labels that are organised in a hierarchical tree structure, where coarse classes (like *orchards*) are subdivided into finer ones (like *apples*, *pears*, *vines*, etc.). We develop a crop classification method that exploits this expert knowledge and significantly improves the mapping of rare crop types. The three-level label hierarchy is encoded in a convolutional, recurrent neural network (convRNN), such that for each pixel the model predicts three labels at different level of granularity. This end-to-end trainable, hierarchical network architecture allows the model to learn joint feature representations of rare classes (e.g., *apples*, *pears*) at a coarser level (e.g., *orchard*), thereby boosting classification performance at the fine-grained level. Additionally, labelling at different granularity also makes it possible to adjust the output according to the classification scores; as coarser labels with high confidence are sometimes more useful for agricultural practice than fine-grained but very uncertain labels. We validate the proposed method on a new, large dataset that we make public. *ZueriCrop* covers an area of 50 km × 48 km in the Swiss cantons of Zurich and Thurgau with a total of 116'000 individual fields spanning 48 crop classes, and 28,000 (multi-temporal) image patches from Sentinel-2. We compare our proposed hierarchical convRNN model with several baselines, including methods designed for imbalanced class distributions. The hierarchical approach performs superior by at least 9.9 percentage points in F1-score.

Keywords: deep learning, recurrent neural network (RNN), convolutional RNN, hierarchical classification, multi-stage, crop classification, multi-temporal, time series

1. Introduction

Monitoring agricultural land use is of high importance for food production, biodiversity, and forestry (Gómez et al., 2016). An increasing world population, climate change, and changes in food consumption habits put yet uncultivated areas under pressure, while leading to intensification in existing agricultural areas (Laurance et al.,

2014). Cropland expansion and intensive use of agricultural areas are often connected with negative ecological impacts like deforestation and biodiversity loss, but also degradation of ecosystem services like ground and surface water quality (Herzog et al., 2008; Dise et al., 2011). Therefore, dense, accurate monitoring of agricultural lands plays an essential role for their optimal and sustainable management. Knowledge of crop areas and certain land uses is of importance for many political programs that aim to reduce and alleviate the environmental impacts of intensive agriculture, too (Gómez et al., 2016). Policy driven incentives, for instance, foster a particular share of a farm area to remain extensively used grassland to promote biodiversity, or give subsidies to promote a certain crop mix in the rotation (Finger and Lehmann, 2012). Collecting information is traditionally based on farmer self-reporting and spot checking by the authorities in the field, which is laborious, costly, and prone to errors.

Modern machine learning methods in combination with publicly available satellite imagery provide new possibilities for more accurate spatially dense monitoring of agricultural sites at high temporal resolution and low cost. One particularly promising recent sensor is Sentinel-2, due to its low ground sampling distance (10 m) at a revisit rate of 3-5 days. In general, the spectral signal of the vegetation as captured by the satellite has specific characteristics as a function of *(i)* soil structure and composition (e.g., soil brightness, soil water content, soil type, etc.), *(ii)* vegetation structure (e.g., canopy cover, Leaf Area Index (LAI), plant height, leaf angle, etc.) and *(iii)* leaf biochemistry (e.g., chlorophyll, water content, nitrogen content, etc.) (Thenkabail et al., 2013). Not only each plant species has its own spectral signature, but spectral characteristics are also highly dependent on the phenological stage of the plant (Walter et al., 2015; Anderegg et al., 2020). Instead of merely analysing images at a single point in time, time-series (sequences) analysis of satellite images thus provides significant additional evidence about crop species.

Supervised machine learning – recently in particular deep learning (Rußwurm and Körner, 2017, 2018b; Rustowicz et al., 2019; Zhong et al., 2019; Pelletier et al., 2019; Rußwurm et al., 2019; Sainte Fare Garnot et al., 2019, 2020) has shown good performance as a tool for multi-temporal vegetation mapping, on different datasets (Rußwurm and Körner, 2017, 2018b; Rustowicz et al., 2019; Zhong et al., 2019; Pelletier et al., 2019; Rußwurm et al., 2019; Sainte Fare Garnot et al., 2020). However, these existing datasets usually contain only a small number of relatively well-balanced crop classes (e.g., 9 classes in Rußwurm et al. (2019), 10 classes in Pelletier et al. (2019), 13 in Zhong et al. (2019)). In practice, large-scale datasets that cover all existing plots in some geographic region have highly skewed and long tail class distributions with a large number of different classes.

When dealing with imbalanced data, a large number of (rare) classes comes with very few labels, which makes training any data-driven method challenging. A viable way to alleviate this problem consists in using hierarchical classification strategies (Srivastava and Salakhutdinov, 2013; Zhu and Bain, 2017; Wehrmann et al., 2018; Koo et al., 2018; Roy et al., 2020). Although many classes are rare, they may belong to the same super-class at a coarser level and share common features, e.g., leopards, tigers, lions, and cheetahs all share the visual properties of cats. This observation can be used to regularize the training of a model and improve the generalization error, especially for the rare classes. Imposing prior knowledge about the class structure adds (soft)

constraints to the model and encourages it to pool shared information from related classes. Such "coarse-granularity features" are easier to learn due to the larger training set, while the fine-grained classification can focus on discriminating fewer sub-classes, thus using the rare training examples more efficiently. Crop classification is a task for which we can define such a label hierarchy. For example, apple orchards, pear orchards and chestnut orchards (which rarely appear individually) all belong to the same *orchards* subclass and share many visual features. We point out another advantage of a hierarchical scheme: one can use the class scores to determine how reliable predictions are at different hierarchy levels. In case output scores at the most fine-grained level (e.g., apple orchard, pear orchard, chestnut orchard) are low, we can always use coarser outputs (orchard), which typically receive significantly higher scores because they aggregate evidence across more data. For many applications (e.g., summary statistics, calculation of subsidies) that coarser granularity of annotation is good enough. Moreover, only passing the fine-grained decisions between few, rare classes to human experts greatly reduces their manual cleaning and relabeling workload.

In this work, we propose a deep learning network architecture for crop mapping that is *hierarchical*, to exploit a tree-structured label hierarchy built by domain experts; *convolutional* to encode image data; and *recursive* to represent time series. The proposed architecture has multiple levels of representation that consist of stacked, convolutional recurrent neural networks. The different network levels predict successively finer label resolution in the hierarchical tree. Moreover, we add a label refinement module, which takes the predictions as input and refines them with a Convolutional Neural Network (CNN) that exploits the correlations between labels across the hierarchy.

In order to test our model, we introduce a new dataset called *ZueriCrop*. This dataset is based on farm census data from the Swiss Federal Office for Agriculture (FOAG) and consists of annotated field polygons from the Cantons of Zurich and Thurgau from the year 2019. This dataset contains 48 different classes with a realistic, highly imbalanced class distribution. The dataset comes with a label hierarchical tree, built using expert knowledge, that can be leveraged during training. Our experiments show that the proposed model outperforms the state-of-the-art methods on our *ZueriCrop* dataset; in addition, it is more effective than widely used techniques for coping with imbalanced data distribution, such as data augmentation or class-balanced loss functions. To summarize, our contributions are:

- We propose a new, multi-temporal crop classification method that encodes a domain-specific label hierarchy directly inside an end-to-end trainable model architecture. It outputs labels at multiple granularity levels for each pixel and significantly improves classification accuracy.

- We provide a new, publicly available crop classification dataset *ZueriCrop*, equipped with a tree-structured label hierarchy. *ZueriCrop* covers a 50 km $\times$ 48 km area in the Swiss cantons of Zurich and Thurgau. It contains 28,000 Sentinel-2 image patches of size 24 pixels $\times$ 24 pixels, each observed 71 times over a period of 52 weeks; 48 agricultural land cover classes; and 116,000 individual agricultural fields.

2. Related Work

Crop Classification with multi-temporal satellite data has been widely studied in remote sensing. Traditional machine learning approaches with handcrafted features (Inglada et al., 2015; Wardlow and Egbert, 2008; Vuolo et al., 2018) predominantly rely on vegetation indices like the Normalized Difference Vegetation Index (NDVI) (Foerster et al., 2012; Ustuner et al., 2014; Peña-Barragán et al., 2011; Conrad et al., 2010). Different strategies have been explored to better model the temporal evolution as further evidence for classification, such as temporal windows (Conrad et al., 2014), hidden Markov models and dynamic time warping (Siachalou et al., 2015; Belgiu and Csillik, 2018), and conditional random fields (Bailly et al., 2018). For instance, Wardlow and Egbert (2008) extract features, which are time series of NDVI, from MODIS data collected over the growing season of crops; and they perform classification with a decision tree. Similarly Conrad et al. (2014) investigate the optimum number of acquisition dates and most suitable temporal windows for the discrimination of crops from RapidEye satellite time-series data. These traditional methods have in common that their performances are constrained by the limited discriminativeness and robustness of the hand-crafted features, as well as by the limited expressive power of conventional classifiers.

More recently deep learning methods have shown their ability to effectively solve many pattern recognition task. Their main advantage is twofold: *(i)* they no longer rely on hand-engineered features to encode spectral, spatial, or temporal patterns; *(ii)* their large capacity makes them able to learn very complex, highly non-linear relationships, if given sufficient labeled training data and computational resources. Rußwurm and Körner (2017) use a recurrent neural network with Long Short-Term Memory (LSTM) to encode temporal dependencies in the data, while Rußwurm and Körner (2018b) improve the result on the same dataset by encoding both, temporal *and* spatial dependencies via convolutional LSTM and Gated Recurrent Units (GRUs). In (Rustowicz et al., 2019; Sainte Fare Garnot et al., 2019), satellite images are first processed individually with a CNN to obtain per-image features; then temporal dependencies between these features are modeled with a separate Recurrent Neural Networks (RNNs). Further options are temporal CNNs that combine features also across time with convolutions (Pelletier et al., 2019), or models that use the attention principle (Vaswani et al., 2017) to aggregate information across time (Rußwurm et al., 2019; Rußwurm and Körner, 2020). Sainte Fare Garnot et al. (2020) combine pixel-set encoder and transformer (Vaswani et al., 2017) and show improved performance over RNN-based approaches. Finally, in our previous work (Turkoglu et al., 2021) we build a deep RNN with a new cell structure termed STAR that trains better than LSTM- and GRU-type models while being more parameter-efficient. This makes it possible to train deeper models, which translates to improved performance across a range of sequence modelling tasks, including crop classification.

Handling imbalanced datasets is generally an issue in supervised classification. Modern deep learning methods are data-hungry and prone to overfit. In order to generalize well on the test data, a large amount of labeled training data is usually required. In practice, however, some classes occur more often than others (e.g., animal species, crop types) or some labels are simply easier to collect. This leads to long-tailed class distri-

butions being the norm, rather than the exception, for large, real-world datasets (Xiao et al., 2010). Under standard training regimes, machine learning models tend to ignore rare, under-represented classes and focus on the dominant classes to maximize cumulative performance across the entire dataset (Wang et al., 2017; Ren et al., 2018; Dong et al., 2018). While those shortcomings are easily overlooked when evaluating with global performance metrics like overall accuracy, they become obvious with class-balanced metrics like average class precision or F1-score. In fact, for many practical applications a class-balanced evaluation is essential, as rare classes have the same (or even higher) importance as frequent ones. This is also true for our agricultural mapping problem where crops that have high financial or ecological value (for instance orchards and vegetables) are rare compared to pastoral grasslands or wheat fields.

Two major strategies have been explored to counter class imbalance: *(i)* algorithm-level approaches and *(ii)* data-level approaches. A typical algorithm-level approach is cost-sensitive learning where the loss function is re-weighted by a factor inversely proportional to the class frequencies (Ling and Sheng, 2008; Huang et al., 2016; Khan et al., 2017, 2019), where training samples of rare classes receive higher weight. An inherent consequence of resampling or reweighting training samples according to rarity is a model bias towards the rare classes. Data-level approaches try to balance the dataset either by oversampling minority classes (Chawla et al., 2002; Douzas and Bacao, 2018; Cui et al., 2019) or by under-sampling the majority classes (He and Garcia, 2009). Undersampling dominant classes runs the risk to miss large parts of the data distribution, thus hurting model performance. On the other hand, oversampling rare classes reaches its limit if the number of available samples in a class is too low. We propose to leverage the hierarchical structure of the labels to counter class imbalance and to improve performance for rare classes.

Hierarchical classification in remote sensing has been investigated before. Melgani and Bruzzone (2004) develop hierarchical tree-based classification strategies using binary classifiers like support vector machines (SVM) for hyper-spectral remote sensing data. Chen et al. (2009) propose a rule-based method that hierarchically classifies land cover and land use from LIDAR and WorldView-2 data. Similarly, Wu et al. (2016) apply a rule-based classification of LIDAR data for *building* classification followed by a classification of *road*, *vegetation*, and *bare soil* with SVMs by additionally incorporating WorldView-2. Another rule-based, hierarchical method is proposed in (Heupel et al., 2018) for crop classification from four different satellite sensors: Landsat-7 and Landsat-8, Sentinel-2A and Rapid-Eye. The first-level classifier decides whether it is *winter crop* or *summer crop* and the second-level classifiers predict into eight fine-grained classes, e.g., *potato, corn*. Their method exhaustively relies on hand-crafted features and expert knowledge. Jiao et al. (2019) apply rule-based decision-making to classify land covers of coastal wetlands into four coarse classes which are subdivided into more fine-grained classes with SVMs. Goel et al. (2018) study hierarchical metric learning for classification of remote sensing data. They use iterative max-margin clustering to organize the classes in a hierarchical fashion and subsequently learn different distance metric transformations for the classes present at the non-leaf nodes of the tree. Another idea is using individual Random Forests at different hierarchy levels for land cover and land use classification (Sulla-Menashe et al., 2011, 2019) from MODIS data. More recently, Demirkan et al. (2020) investigated the benefit of hierarchical classifi-

cation with SVMs and Random Forests for land cover and land use classification from Sentinel-2 data. Although hierarchical classification in remote sensing has been studied before, all methods use traditional, hand-crafted features and decision trees designed for specific scenes and datasets by domain experts. Complex workflows that employ multiple independent classifiers at different stages are costly to compute, need much manual tuning for each new dataset, and erroneous decisions at early stages can hardly be compensated for later on in the pipeline. Additionally, none of the existing workflows dealt with a hierarchical approach for a realistic, large-scale, imbalanced dataset with a large number of classes.

Hierarchical classification has already been studied in deep learning literature. Srivastava and Salakhutdinov (2013) introduce a tree-like hierarchy in CNNs for image classification. Their method learns to organize the classes into a tree hierarchy that imposes a prior over the classifier's parameters, which improves performance for minority classes. Yan et al. (2015) embed deep CNNs into a two-level hierarchy. Easily distinguishable classes are separated with a coarser classifier, while another classifier separates the other, more difficult cases at a more fine-grained level. Xiao et al. (2014); Roy et al. (2020) study hierarchical networks composed of deep CNNs in the context of incremental learning. Chen et al. (2019) propose a training strategy that leverages the information from a label hierarchy. It maximizes the probability of the ground truth class, and at the same time, neutralizes the probabilities of the other classes in a hierarchical fashion, making the model take advantage of the label hierarchy explicitly. Another interesting recent work, in the field of (hierarchical) text classification, is (Mao et al., 2019), where the hierarchy is explored at both training and inference time with a Markov decision process and deep reinforcement learning. Zhu and Bain (2017) and Wehrmann et al. (2018) propose multi-stage deep CNNs. Like ours, their models have multiple outputs of different granularity as well as multiple objectives. Koo et al. (2018) further improve on that idea by combining a CNN that extracts a hierarchical image representation with an RNN to capture the hierarchical tree of labels.

To the best of our knowledge, our method is the first that explicitly encodes the inherent (and for domain experts well-known) label hierarchy for crop classification in satellite image sequences in the deep learning setting. Our method differs from all existing literature in that it is based on an integrated, convolutional and recurrent model (convRNN) that is able to capture all relevant spatio-temporal correlations. To demonstrate these features, and to enable further work in this direction, we also provide a new dataset which, compared to existing ones (Rustowicz et al., 2019; Rußwurm et al., 2019; Rußwurm and Körner, 2017), has many more classes and a realistic, much less balanced class distribution.

3. Method

Formally, our objective is to predict a crop type map $\mathbf{Y} \in \mathbb{R}^{H \times W \times C}$ from a sequence of input images $\{\mathbf{X}_1, \mathbf{X}_2, .., \mathbf{X}_T\} \in \mathbb{R}^{H \times W \times B}$. H and W are height and width of the input images, respectively, B is the number of input bands, T is number of time stamps in the input sequence, and C is the number of crop types. See Fig. 4. We assume that multiple labels are assigned to a single pixel. Each of those labels belongs to a different level in a hierarchical structure that encodes agricultural crop types at a different granularity,

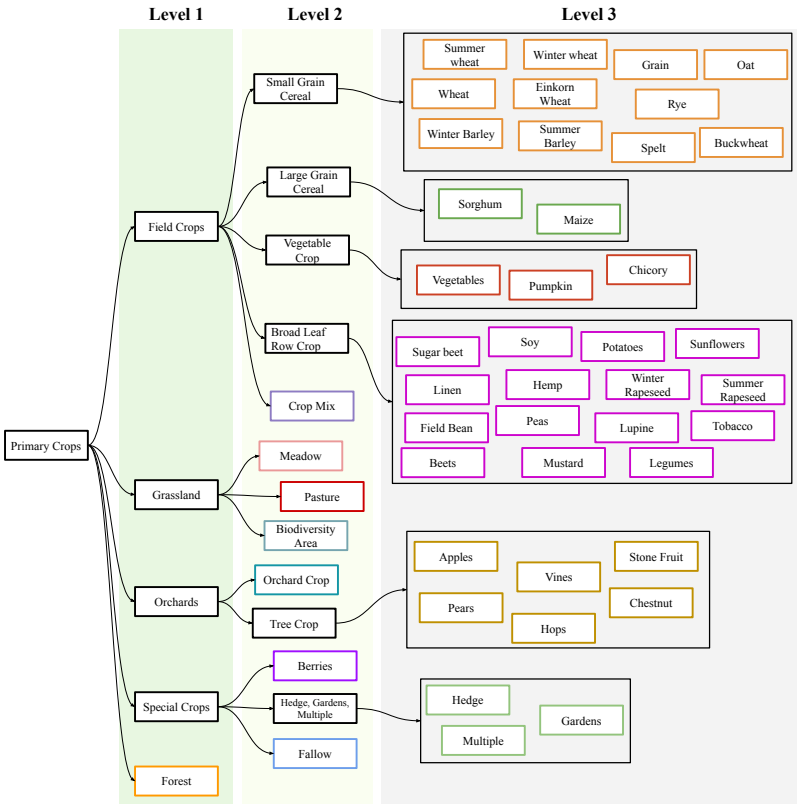


Figure 1: The hierarchy of all crop classes of the *ZueriCrop* dataset. Black box indicates intermediate label levels while color boxes indicates the finest granularity.

from coarse to fine. This label hierarchy is created by human experts, the levels have an intrinsic semantic meaning. Fig. 1 shows our label tree for all classes used in our dataset.

Labels are denoted as $\mathbf{Y}^n$, where $n \in 1, ..., N$ represents the level inside the label hierarchy. Even though there are multiple labels for each crop type, the ultimate goal is to predict (as much as possible) the finest granularity $\mathbf{Y}^N$, corresponding to detailed species labels. In our dataset, we distinguish three hierarchy levels, with 1 the coarsest and 3 the finest. We note that although, for clarity, we stick to the case $N = 3$ for the rest of the paper, our method is generic and be used for other values of N.

3.1. Convolutional Recurrent Neural Networks

Convolutional recurrent neural networks (convRNN) are the convolutional version of RNNs, designed to represent spatio-temporal data. ConvRNNs have been used for different spatio-temporal modeling tasks like weather forecasting (Xingjian et al., 2015), video action recognition (Li et al., 2018), video forecasting (Su et al., 2020), and the prediction of heat diffusion Saha et al. (2020). They differ from standard RNNs in

that the matrix multiplications are replaced with the convolution operator. In general it is straightforward to convert any recurrent cell with its convolutional version, see for instance (Xingjian et al., 2015; Siam et al., 2017; Turkoglu et al., 2021). Constructing a network with multiple layers helps learning more discriminative evidence via a richer set of features. However, training networks with many layers is hard if using widely known LSTM and GRU cells as basic recurrent units as shown in (Turkoglu et al., 2021). A computationally more efficient cell type with less parameters that allows training deeper models is convSTAR (Turkoglu et al., 2021). We thus construct a network using convSTAR units and demonstrate its superior performance over versions using GRU and LSTM units in the experiments section (Section 6), Table 3). In the following, we will briefly recap the design of a convSTAR cell before describing the construction of our hierarchical approach.

More formally, the convSTAR cell at convolution layer l and at time t, takes as input the hidden state tensor $\mathbf{H}_t^{l-1}$ of the previous layer $(l-1)$, where the input $\mathbf{H}_t^0$ corresponds to the channels of the multi-spectral input image $\mathbf{X}_t$. That tensor is first non-linearly transformed and then linearly combined with the previous hidden state $\mathbf{H}_{t-1}^l$, with the weights of the linear combination modulated by a gating variable $\mathbf{K}_t^l$ that depends on both inputs and controls the information flow. Formally, the computation is given by

$$\mathbf{K}_t^l = \sigma(\mathbf{W}_x * \mathbf{H}_t^{l-1} + \mathbf{W}_h * \mathbf{H}_{t-1}^l + \mathbf{B}_K) \tag{1}$$

$$\mathbf{Z}_t^l = \tanh(\mathbf{W}_z * \mathbf{H}_t^{l-1} + \mathbf{B}_z) \tag{2}$$

$$\mathbf{H}_t^l = \tanh(\mathbf{H}_{t-1}^l + \mathbf{K}_t^l \circ (\mathbf{Z}_t^l - \mathbf{H}_{t-1}^l)) \tag{3}$$

where σ is the sigmoid non-linearity, $*$ denotes convolution, $\circ$ is the Hadamard (element-wise) product, and $\mathbf{W}$ and $\mathbf{B}$ are the trainable weight and bias tensors, l and t denote the cell layer and time step of the sequence, respectively. The hidden state $\mathbf{H}_T^L$ of the deepest (L-th) layer serves as latent encoding of the input sequence up to time T and can be used for classification, regression or forecasting, by passing it through an appropriate decoder.

3.2. Hierarchical Convolutional Recurrent Network

In order to construct a hierarchical representation similar to multi-stage CNN classification networks, but for image sequences instead of single images (Wehrmann et al., 2018; Zhu and Bain, 2017), we propose a network that has N stages. Each stage is made of a 2-layer convSTAR architecture (Fig. 2). The output tensors (hidden states) at stage n are fed as inputs to the next stage $(n+1)$. The final hidden state $(\mathbf{H}_T^2)^n$ of each hierarchical level is fed to a conventional CNN classifier to obtain label scores $\hat{\mathbf{Y}}^n$.

All convolutional kernels have size 3×3. Each convSTAR layer has 64 filters, a shallow 1-layer convolutional neural network (CNN) is used to convert the final hidden state to a patch of labels. The network is trained with the cross-entropy (CE) loss, leading to the objective function

$$L = \sum_{n=1}^{N} \lambda_n CE(\mathbf{Y}^n, \hat{\mathbf{Y}}^n) = -\sum_{n=1}^{N} \sum_{c=1}^{C} \lambda_n \mathbf{Y}_c^n \log(\hat{\mathbf{Y}}_c^n) \tag{4}$$

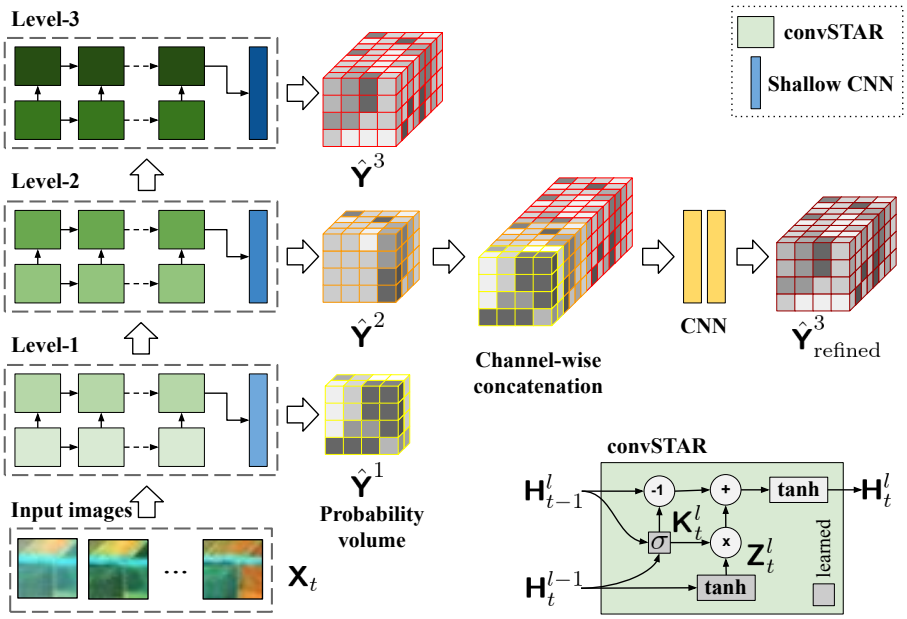


Figure 2: Our proposed hierarchical, multi-stage, convolutional STAR network (ms-convSTAR).

where λ_n represents hyper-parameters that determine the relative influence of different hierarchy levels on the tree, with $\sum \lambda_n = 1$. By taking into account the loss at all levels of granularity, the network imposes the label hierarchy as a (soft) prior to guide the feature encoding. For example, the features of *apple orchards* and *pear orchards* should support an assignment to the coarser *orchard* label, too.

3.3. Label Refinement

The model described so far embeds the label hierarchy in the network. However, by itself this does not guarantee consistency between the predictions at different stages. As an extreme example, the network could learn to simply ignore the input from the coarser hierarchy level and build independent classifiers for different levels of granularity. As a consequence, labels predicted at test time could possibly violate the parent-child relations of the hierarchy. A pixel could receive labels *sunflower* and at the same time *orchard*, for example. To imprint the preference for coherent labels across the hierarchy levels, we add a *label refinement network*, which is a CNN after the multi-level ms-convSTAR network. The refinement stage takes all the predictions $\{\hat{\mathbf{Y}}^1 \dots \hat{\mathbf{Y}}^N\}$ from the individual network stages and produce a final prediction for the finest label granularity, $\hat{\mathbf{Y}}^N_{\text{refined}}$, while modelling their interactions. More specifically, 3-dimensional probability volumes from three stages are concatenated in the channel dimension and fed to the CNN. See Fig. 2. Formally, it is defined as

$$\hat{\mathbf{Y}}^N_{\text{refined}} = \hat{\mathbf{Y}}^N + F(\text{concat}[\hat{\mathbf{Y}}^1, \hat{\mathbf{Y}}^2, ..., \hat{\mathbf{Y}}^N]) \tag{5}$$

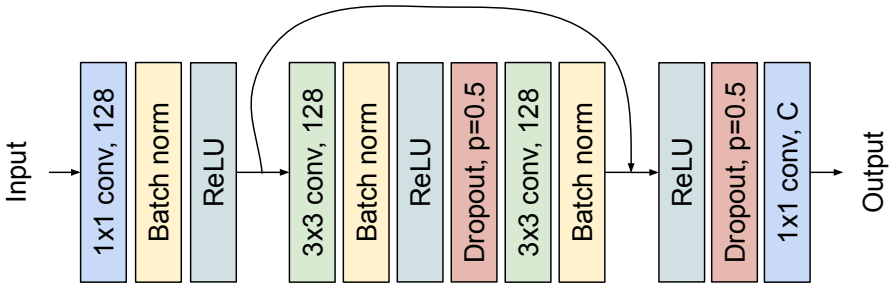


Figure 3: Label refinement CNN architecture.

where F represents a CNN, see Fig. 3. The final loss function of the network is the weighted sum of the ms-convSTAR losses and the refinement loss, with a hyper-parameter γ for the relative influence of the refinement loss:

$$L = \sum_{n=1}^{N} \lambda_n CE(\mathbf{Y}^n, \hat{\mathbf{Y}}^n) + \gamma CE(\mathbf{Y}^N, \hat{\mathbf{Y}}^N_{\text{refined}}) \tag{6}$$

where $\hat{\mathbf{Y}}^N_{\text{refined}}$ represent the output of the label refinement module and has exactly the same dimension as $\hat{\mathbf{Y}}^N$, see Fig. 2. Note that, empirically, we observed that refining coarser predictions does not affect the final model performance significantly. For experimental evaluation we thus run the refinement module exclusively for the finest label granularity.

3.4. Implementation details

We have implemented our network architecture in PyTorch. Input image sizes are $H = W = 24$, $B = 4$ and sequence length is $T = 71$. For all experiments, we use Adam (Kingma and Ba, 2014) as optimiser with batch size 4 and run the training for 30 epochs. The learning rate is set to 0.001 at the start of the training and divided by 10 every 10 epochs. The model is regularised with weight decay of 0.0001. Gradient magnitudes are clipped to 5 to prevent exploding gradients. The hyper-parameters of the loss function (4) are empirically determined and set to $\lambda_1 = 0.1$, $\lambda_2 = 0.3$, and $\lambda_3 = \gamma = 0.6$. Input image patches are flipped randomly with 66% chance during training for data augmentation. All source code, trained models and the dataset are available online at https://github.com/0zgur0/ms-convSTAR.

4. Dataset

The *ZueriCrop* dataset contains ground truth labels of 116,000 field instances. Each field instance consists of a polygon representing the borders of the field, and its dominant crop label in 2019. The ground truth labels of all 48 crop classes are provided by the Swiss Federal Office for Agriculture (FOAG) and correspond to the primary crop grown per field during the year. No information is provided about intermediate or cover

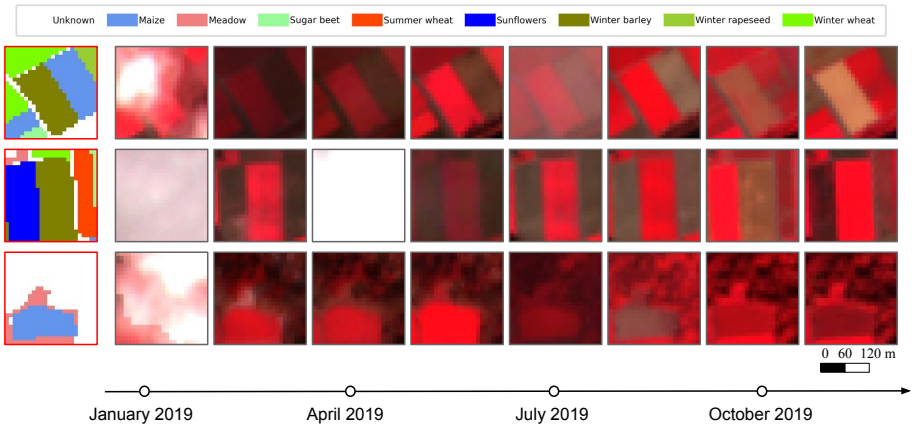


Figure 4: Example Sentinel-2 satellite images of the *ZueriCrop* dataset. Each row shows randomly sampled images (false color composite: NIR-Red-Green) from a satellite image time-series. The first column shows the ground-truth where different colors correspond to different crop types.

crops, i.e., crops planted after harvesting the primary crop to cover the field over the winter in order to improve soil fertility and reduce disease pressure. The input data is a time series of 71 multi-spectral Sentinel-2 Level-2A bottom-of-atmosphere reflectance images with a ground sampling distance (GSD) of 10 meters (Fig. 4). All input images are atmospherically corrected using the Sen2Cor v2.8 software package. The dataset is collected over a 50 km $\times$ 48 km area (Fig. 5) in the Swiss Cantons of Zurich and Thurgau between January 2019 and December 2019.

We subdivide the entire scene into smaller patches of 24 px$\times$24 px. Patches without any ground-truth information are discarded. In the remaining patches the fraction of pixels without reference label is $\approx$48%. Only those four spectral channels available at the highest, 10 m resolution (Red, Green, Blue, and Near-Infrared) are used because we observed that adding more channels did not significantly improve performance while increasing the computational cost, we refer the reader to Appendix B for an empirical evaluation how additional bands impact the model performance. We do not use any cloud detection method to discard patches with a high cloud cover because RNN architectures are robust to uninformative inputs. See Section 6.4.

Switzerland has a small-structured agricultural system, where farmers are not allowed to grow crop after crop, but are required to adhere to a diverse crop rotation scheme. The average farm and field sizes in Switzerland are 21 hectares and 1.5 hectares, respectively with approximately 70% of all Swiss agricultural land being grassland (Bundesamt für Statistik, 2020), of which 12% are temporary grasslands used in rotation with other crops (Stumpf et al., 2020). This situation leads to a diverse set of crop classes with a highly skewed, imbalanced class distribution (Fig. 6).

4.1. Crop class hierarchy

We organize the 48 crop classes of the *ZueriCrop* dataset into a hierarchy based on expert knowledge about the Swiss agricultural system (Fig. 1).

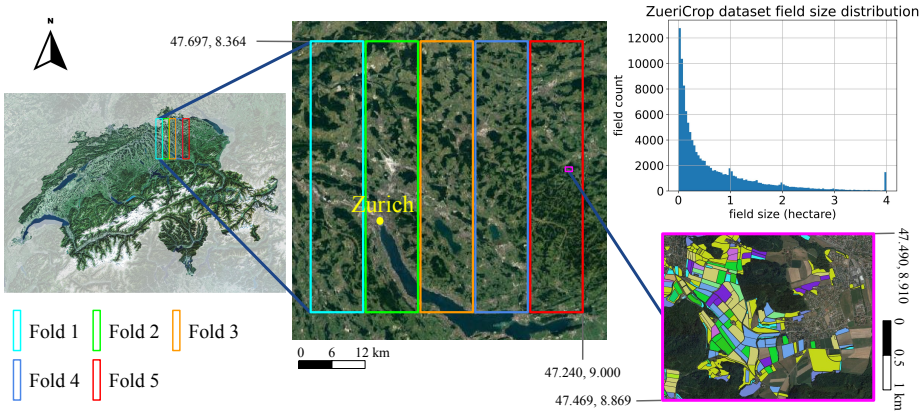


Figure 5: Overview of the *ZueriCrop* dataset collected in 2019: Location inside Switzerland (left), Sentinel-2 image of the area of interest, overlaid with the geographical split used for cross-validation (center), example of GIS reference data for the main crop per field (bottom right), and distribution of field sizes (fields > 4 hectares are pooled into one bin for visualization). The average field size is 0.72 hectares.

The 1st level of the class hierarchy was chosen with two goals in mind: *(i)* separate the main categories found in the Swiss agricultural landscape and *(ii)* group crop types according to their visual appearance in satellite images. For example, *field crops* are grown in crop rows on fields, a feature that can be picked up by remote sensing. Another example is *grassland*, the largest class in the dataset, which is a very heterogeneous class containing many different grassland types, mixtures and other land use scenarios. Their main feature is that they are permanently green with high biomass and generally low plant height. Other classes on the 1st level are: *orchards*, *special crops* and *forest*. Orchards can usually be identified by their specific planting patterns. Special crops are a pool of marginally grown specialist crops (e.g., asparagus, different berries and herbs) and have a very diverse set of sub-classes.

The 2nd level of the class hierarchy contains more refined versions of the preceding classes. Classes in the second level were selected to represent the plant family and agronomic use and practices in cultivation. All crops of the *small grain cereal* class are cultivated in a very similar manner in rows with little row spacing, similar plant seed density per square meter and similar erectile canopy structure with ears appearing at the top of the plant habitus (small grain as fruits). Similarly, the *broad leaf row crop* class contains dicotyledonous plant species that are cultivated in rows and possess, in contrast to grain crops, a horizontal leaf surface pattern.

The 3rd level distinguishes different crop species and is the finest level in our hierarchy. For some crops the ground truth was not reported at that granularity. In such cases the 2nd-level labels were copied. For the *forest* class, 1st level label is copied to 2nd and 3rd levels.

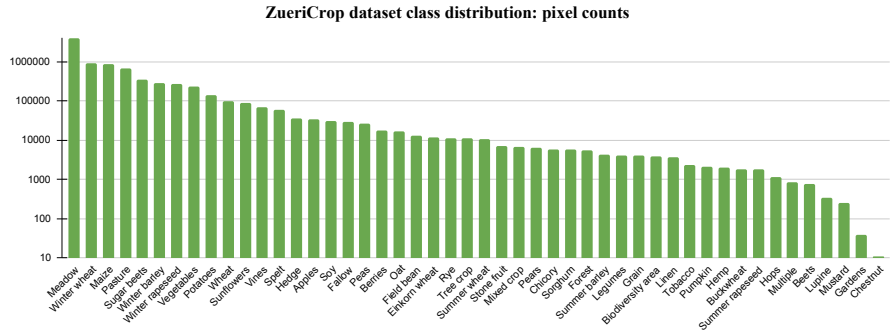


Figure 6: Class distribution in the *ZueriCrop* dataset: Histogram of pixels. Note the logarithmic scale of the y-axis.

5. Experiments

We compare the performance of the proposed hierarchical scheme to several baselines, and to competing state-of-the-art methods. In order to avoid any biases due to regional differences within the area of interest, we always perform 5-fold leave-one-out cross-validation and average the performance metrics across all five train/test splits. We divide the dataset into 5 geographically disjoint strips of equal size as shown in Fig. 5, and use 4 strips as training set and the 5^{th} one as test set. In the ground truth, a single label is assigned to each field polygon, i.e., differences between administrative field boundaries and actual crop boundaries are not represented. To match that format, we assign the same label to all pixels within a field polygon by majority voting (See Appendix for results without majority voting). Classification performance is evaluated pixel-wise, with four different metrics: overall accuracy, average per-class precision, average per-class recall, and average per-class F1-score (Table 1). Note that overall accuracy, unlike the other metrics, does not compensate for class frequencies. It measures correctness *per pixel*, respectively area, but not correctness *per label*, respectively crop type; and is thus dominated by the performance for (few) frequent crops. On the contrary, the other metrics are better indicators of class-wise performance, as they are computed separately per class (independent of the absolute pixel count) and averaged.

As baseline for the hierarchical approach we also run the standard (single-level) convSTAR network (Turkoglu et al., 2021) without intermediate outputs, losses, and label refinement. Hyper-parameters, like network depth, hidden state size and learning rate schedule are all set the same for the hierarchical method. As further baseline experiments we add data augmentation and class-balanced losses to the standard convSTAR network, which are alternative, widely used techniques to improve the performance on imbalanced datasets. When balancing the loss, the contribution of a training example is weighted inversely proportional to the class frequency, such that every class (in principle) contributes an equal share of the total loss. For data augmentation, training patches are sampled inversely proportional to the class frequency to achieve the same effect. The class frequencies are found by counting pixels over the entire training set.

The use of a class-balanced loss functions causes a subtle difference, as it affects the gradient magnitude reducing the overall learning rate, which can potentially harm the training. Therefore, we also test another variant, denoted as Class-balance loss-2, where, after including the weight of the loss function for each class we compute the median effective learning rate and set it to match the default learning rate. Moreover, we also evaluate the recently proposed, more robust re-weighting scheme of Cui et al. (2019). That state-of-the-art technique uses the effective number of samples for each class to re-balance the loss. As the method has an additional, empirically chosen hyper-parameter β, we test different settings $\beta \in \{0.99, 0.999, 0.9999\}$.

The baselines described so far serve to isolate the impact of the proposed hierarchical labeling scheme, and to that end use the same convSTAR backbone. As an additional, external reference we also compare against Random Forest and other state-of-the-art deep learning methods that have been developed for multi-temporal crop classification. Rußwurm and Körner (2017) use LSTM, Pelletier et al. (2019) propose a temporal convolutional network, and Rußwurm et al. (2019); Rußwurm and Körner (2020) design a Transformer network. All four methods process pixels individually, unlike our approach that aggregates context information from image patches with convolutions. Convolutions and recurrence are used separately by Rustowicz et al. (2019), who extract features from the images with a convolutional architecture in the style of U-Net (Ronneberger et al., 2015) and apply convolutional LSTM to the resulting feature vectors to represent their temporal evolution. Also, U-Net (Ronneberger et al., 2015) itself is compared since it has become a standard architecture for image segmentation and has been used in remote sensing (e.g., Stoian et al. (2019); Flood et al. (2019)). Early temporal fusion, spectral channels of different timestamps are concatenated in the channel direction, is applied to deal with multi-temporal data. An integrated, convolutional variant of GRU is already used by Rußwurm and Körner (2018b). This work is, from a technical point of view, similar to our convSTAR baseline, except that it uses a different form of recurrence and employs a bi-directional approach.

6. Results

In this section, we first compare the performance of our proposed ms-convSTAR against baseline methods (Table 1), as well as other state-of-the-art methods (Table 2) on the *ZueriCrop* dataset. In the ablation study (Section 6.2), we evaluate the effectiveness of our label refinement component (Table 1), and compare different convRNN types (Table 3). We also discuss how the proposed method can be used in combination with the hierarchical label tree to provide more certain prediction by adjusting the label coarseness at prediction time. Finally, we also discuss the robustness of ms-convSTAR against clouds.

6.1. Performance Comparison

We find that among the baselines, simple data augmentation performs best, but all baselines are clearly outperformed by the proposed, hierarchical ms-convSTAR, on all performance metrics. See Table 1. Most significantly, there are significant improvements on those metrics that compensate for class frequencies and average measure

Method	Prec (%)	Rec (%)	F1 (%)	Acc (%)
convSTAR	40.2	37.3	37.2	87.3
+ Data augmentation	48.3	39.3	41.1	85.0
+ Class-balanced loss	26.9	32.7	28.2	75.6
+ Class-balanced loss-2	26.5	31.7	27.3	74.0
+ Cui 2019, $\beta = 0.99$	42.2	37.5	36.5	87.3
+ Cui 2019, $\beta = 0.999$	39.4	35.7	35.6	87.4
+ Cui 2019, $\beta = 0.9999$	43.3	39.1	39.4	87.1
ms-convSTAR w/o LR	59.3	47.1	49.8	87.6
ms-convSTAR	**60.1**	**49.8**	**52.4**	**88.0**

Table 1: Performance comparison between ms-convSTAR (bottom row) and non-hierarchical baseline methods. We compare against standard (1-level) convSTAR (top row) as well as further baselines that extend convSTAR with different techniques intended to compensate class imbalance. Precision, recall and F1-score are mean values over all classes. All numbers are averaged over 5 cross-validation folds. The best score for each metric is shown with **bold**.

per-class performance. Our proposed ms-convSTAR increases mean class precision by > 11 percentage points, and mean class recall by > 10 percentage points. Accordingly, the F1-score (their harmonic mean) increases by > 11 percentage points. These results indicate that our method improves in particular the classification of less frequent classes, which was the initial motivation for using the crop label hierarchy and developing ms-convSTAR. Data augmentation, where we over-sample the minority classes during training (i.e., patches with rare classes are sampled inversely proportional to their frequencies), improves the F1-score by 2.9 percentage points compared to the baseline convSTAR. However, it degrades the overall accuracy significantly by 2.3 percentage points, because over-sampling rare classes induces a global bias towards those classes and degrades performance for the dominant classes. Adding a standard class-balanced loss to the baseline convSTAR approach significantly decreases performance across all measures, whereas the state-of-the-art class-balanced loss technique (Cui et al., 2019) can improve the F1-score a little (by 2.2 percentage points with $\beta = 0.9999$) but slightly reduces the overall accuracy (by 0.2 percentage points). In summary, even though a class-balanced loss or data augmentation bring a mild improvement for minority classes, the gains are not very big, and they tend to reduce the overall performance in return. In contrast, our proposed ms-convSTAR greatly boosts performance for mean precision, mean recall and F1-score, while improving overall accuracy, too.

Fig. 7 illustrates the difference between the confusion matrix of results achieved with proposed ms-convSTAR (Fig. 8) and with its baseline counterpart convSTAR. ms-convSTAR improves the performance for many less frequent classes like *stone fruit*, *legumes*, *tobacco* while it does not harm the performance for frequent classes like *meadow* or *maize*. We note that for exceedingly rare classes the performance does not improve. These classes have too few pixels (typically < 1000) to be represented well, moreover they are often only present in some of the five cross-validation stripes,

Method	Prec(%)	Rec(%)	F1(%)	Acc(%)
Random Forest*	46.4	<u>40.7</u>	38.9	78.8
LSTM (Rußwurm and Körner, 2017)	37.7	27.9	29.2	84.1
TCN (Pelletier et al., 2019)	39.2	27.7	29.3	83.5
Transformer (Rußwurm et al., 2019)	<u>56.8</u>	38.4	42.3	85.4
2D-CNN (U-Net)	34.6	25.7	26.7	82.2
U-Net+convLSTM (Rustowicz et al., 2019)	47.7	32.8	35.2	85.0
Bi-convGRU (Rußwurm and Körner, 2018b)	55.0	39.6	<u>42.5</u>	<u>86.4</u>
ms-convSTAR	**60.1**	**49.8**	**52.4**	**88.0**

Table 2: Performance comparison of ms-convSTAR (bottom row) with state-of-the-art methods. Precision, recall and F1-score are mean values over all classes. All numbers are averaged over 5 cross-validation folds. The best score for each metric is shown with **bold** and the second best is <u>underlined</u>. *Random Forest is trained with a balanced dataset by under-sampling the majority classes which leads to improved class-wise performance.

such that they do not appear in either the training set or the test set. For completeness, we nevertheless leave those classes in the dataset.

In Table 2, we compare the proposed ms-convSTAR (bottom row) to a number of state-of-the-art methods for crop classification from image time series. ms-convSTAR significantly improves performance across all measures. Again the gains are mostly due to better classification of rare classes, as indicated by an improvement of > 9.9 percentage points in F1-score. Model implementations are taken from official codebases, and similar hyperparameters with the proposed method are used as the hidden size, the initial learning rate, the number of epochs, etc.; for other model-specific hyperparameters (e.g., number of layers in TCN (Pelletier et al., 2019), number of heads in Transformer (Rußwurm et al., 2019)), values from original papers are used. See Appendix for more details. Random Forest parameter settings follow Rußwurm and Körner (2020). To achieve better classification performance for the Random Forest, we augmented the raw input reflectance with the NDVI and the training dataset is randomly under-sampled such a way that number of per class samples is limited to 10K ($\approx$ the median of the number of per class samples) like done in Rußwurm and Körner (2020).

We show qualitative comparisons for several output samples in Fig. 9. Our new method does what it is designed for: it correctly predicts rare classes like *linen* or *sorghum*, where all other methods fail. Another qualitative comparison to the closest competitor (Rußwurm and Körner, 2018b) is shown in Fig. 10 for a larger region. We point out that both approaches, ms-convSTAR and (Rußwurm and Körner, 2018b), typically mis-classify the same fields. But the number of mistakes is significantly smaller with ms-convSTAR. For qualitative results without polygon aggregation, e.g., for mapping in the absence of field boundaries, see Appendix (Fig. B.15).

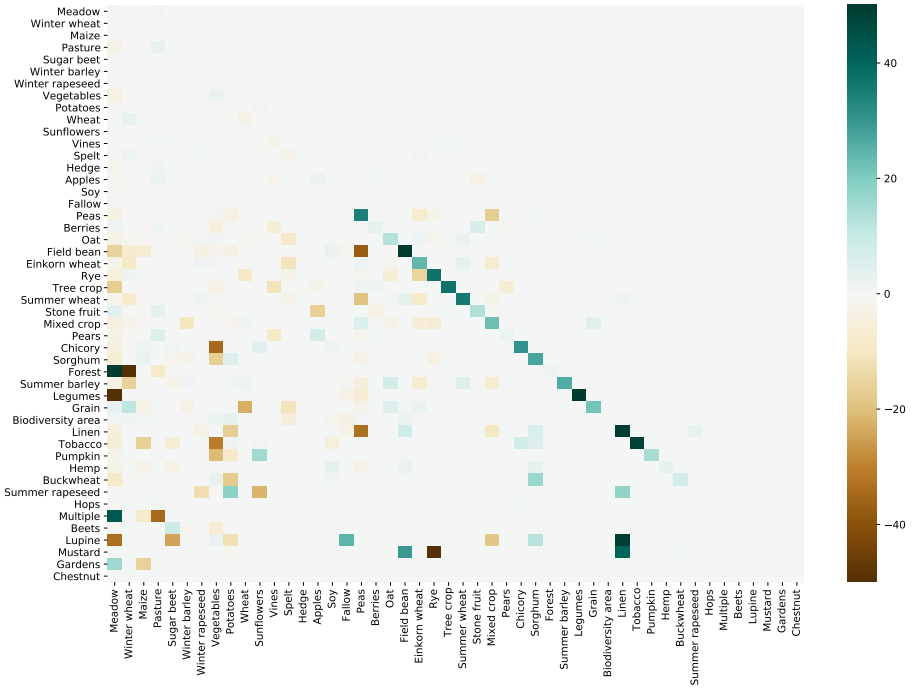


Figure 7: Benefit of ms-convSTAR: Difference between normalized confusion matrices. Averaged over 5 cross-validation folds. Green denotes margins in favour of ms-convSTAR (higher correctness on the diagonal, respectively lower confusion off the diagonal), brown denotes margins in favour of the baseline counterpart (convSTAR).

6.2. Ablation Study

As an ablation study, we first evaluate the performance gain due to the label refinement module by also running our hierarchical ms-convSTAR model without the label refinement module (denoted as ms-convSTAR w/o LR). Results in Table 1 show that the CNN-based label refinement consistently improves all performance metrics, in particular per-class performance.

We then investigate the importance of the RNN cell type used in the network (Table 3). We construct the hierarchical multi-stage network (ms-convX) and its non-hierarchical baseline counterparts (convX) using the most popular RNN cells: LSTM and GRU. We experimentally evaluate 3-layer and 6-layer convRNN versions and corresponding ms-convRNN versions (Table 3). While convLSTM and convGRU produce acceptable results for the 3-layer version without hierarchy, they perform very poorly for the deeper, 6-layer versions. As we have shown in (Turkoglu et al., 2021), both cell types LSTM and GRU suffer from gradient vanishing problems, which get more severe with deeper architectures. Using our proposed hierarchical approach fixes these gradient problems and leads to substantial improvements with respect to their baseline versions without hierarchy. Unlike convLSTM and convGRU, convSTAR does not suffer from gradient issues at any point and performs well for all settings, outperform-

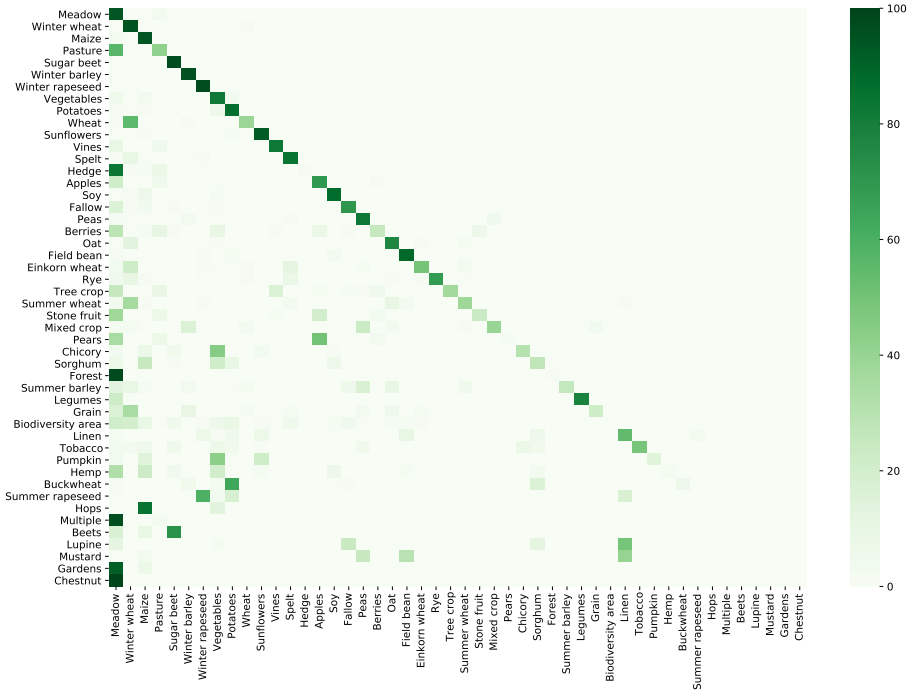


Figure 8: Confusion matrix for the proposed method. Averaged over 5 cross-validation folds. Rows show true labels and columns show predicted labels. The sum of each row is normalized to 1.

ing all existing approaches in overall accuracy (Table 2). Combining convSTAR with the proposed hierarchical approach leads to superior performance with a clear margin above all other methods on the *ZueriCrop* dataset.

6.3. Simultaneous multi-level classification

An attractive property of ms-convSTAR is that it simultaneously predicts per-pixel labels at three different levels in the class hierarchy. Performance of proposed ms-convSTAR at different levels of granularity is given in Table 4. As we expected, the accuracy of the model improves while going to the coarser levels.

Such feature makes also possible to choose the granularity of the labels according to the output (confidence) scores. In this way one can produce reliable maps where most pixels are assigned a label and all labels have sufficiently high confidence – at the cost of only assigning a coarse-grained label to some pixels were the fine-grained answer is too uncertain. For instance if the model is uncertain about deciding for either *apple* or *pear* orchard, the coarser label *orchard* can be assigned with much higher confidence. In a number of applications this coarser level of annotation is good enough. An example are the summary statistics computed by the Swiss federal administration, where a large degree of coverage with coarse labels is critical. Moreover, coarse but correct and (nearly) complete answers are a lot more useful for downstream GIS processing: for

	Method	Prec (%)	Rec (%)	F1 (%)	Acc (%)
3 layers	convLSTM	31.5	27.9	28.1	84.6
	ms-convLSTM	37.7	30.1	31.1	83.9
	convGRU	43.1	38.2	38.6	87.2
	ms-convGRU	49.8	38.6	40.8	85.8
	convSTAR	48.3	42.6	43.5	<u>87.8</u>
	ms-convSTAR	<u>54.6</u>	<u>42.7</u>	<u>45.3</u>	86.8
6 layers	convLSTM	1.1	2.3	1.4	47.2
	ms-convLSTM	40.5	33.6	34.7	85.2
	convGRU	15.8	15.7	15.0	71.1
	ms-convGRU	52.2	42.9	44.7	86.9
	convSTAR	40.2	37.3	37.2	87.3
	ms-convSTAR	**60.1**	**49.8**	**52.4**	**88.0**

Table 3: Performance comparison of multi-stage convRNNs (ms-convRNNs). All numbers are averaged over 5 cross-validation folds. The best score for each metric is shown with **bold** and the second best is <u>underlined</u>.

Level	Prec (%)	Rec (%)	F1 (%)	Acc (%)
1st	80.3	52.9	57.0	96.3
2nd	73.0	51.0	54.5	89.2
3rd	60.1	49.8	52.4	88.0

Table 4: Performance of proposed ms-convSTAR at different levels of granularity. All numbers are averaged over 5 cross-validation folds.

instance, mapping orchards is much easier if the user knows which polygons have not yet received a fine-grained label and must be checked.

To quantify this effect, we measure the area that is classified above a given confidence score, and compare ms-convSTAR models with different numbers of hierarchy levels. We set a prediction confidence value p, and, for each pixel switch to the next-coarser label if the confidence level p is not achieved. Fig. 11 shows the overall accuracy and the coverage rate (the proportion of pixels with a confident label at any granularity) for two confidence levels, $p = 0.6$ and $p = 0.9$. As can be seen if we wish to retain only confident predictions and set $p = 0.9$, the 3-levels hierarchy brings a large benefit. While accuracy (computed over covered areas) stays the same, coverage improves by $\approx 0.20 - 0.25$, although, obviously, some areas are only predicted at a coarse label resolution.

Fig. 12a shows how the coverage rate and the model performance change as functions of the confidence value p. One can see that, as expected, when decreasing p the coverage increases whereas the accuracy decreases. Fig. 12b shows how the model performance improve with increasing p in the full coverage case (every pixel is classified

aat the coarsest available level regardless of the p value). The full, 3-level model boosts performance significantly with increasing p; in contrast, a basic 1-level model does not show this behaviour. This analysis suggests that the imposed label hierarchy is actually meaningful for the problem, in the sense that the coarser classes are indeed easier to discriminate than their finer sub-classes. Moreover, the results once more support our hierarchical multi-level scheme: additional hierarchy levels consistently help to predict more pixels correctly and confidently, for any value of p.

6.4. Cloud Robustness

RNNs normally exhibit good robustness against missing data due to clouds, in the context of crop mapping see, e.g., Rußwurm and Körner (2018a). We have conducted two experiments to further analyse robustness against clouds. For Table 5 we have trained the proposed model with the cloud mask as an additional input channel, so as to explicitly give the model the information which pixels are obscured by clouds. The performance is practically the same, suggesting that the model learns to detect the presence of clouds, such that the additional input brings no benefit. In Figure 13 we look at the issue from another angle: we thresholding the Sen2Cor cloud score to determine how many images are cloudy in the time series of every pixel. Then, we quantify how the varying number of cloud-free observations influences the performance. To that end we rank the classified pixels according to the number of cloudy observations in the time series. As can be seen, the accuracy remains almost constant as pixels are affected increasingly by cloud cover.

Method	**Prec (%)**	**Rec (%)**	**F1 (%)**	**Acc (%)**
ms-convSTAR	59.8	**49.7**	52.1	**88.0**
ms-convSTAR w/ CM	**60.2**	48.8	**52.2**	87.9

Table 5: Performance comparison: the proposed model vs. the same model with the cloud masks (CM) as additional input. Numbers are averaged over 5 cross-validation folds. The best score for each metric is shown in **bold**.

7. Discussion

We go on to discuss limitations of the proposed approach, and highlight cases where prediction is particularly difficult in the *ZueriCrop* dataset.

Although the hierarchical approach significantly reduces the misclassification of underrepresented classes, exceedingly rare classes (e.g., *beets*, *lupine*, *grain*) still often get confused with more frequent ones. The full confusion matrix is shown in Fig. 8. In most cases, very rare crops are misclassified as *meadow*, which is the largest class in the dataset, with almost half of the labeled pixels. The misclassification of field crops like *grains* and *lupine* may in part be attributed to a strong presence of weeds, which is especially common in organic or extensive cropping systems (Giller et al., 2009; Gomiero et al., 2011). This is also reflected by the *mixed crop* class that represents a mixture of grain and legume crops used for Nitrogen efficient production of

animal forage. Another source of errors are mixed pixels that span several crop types. Because the Sentinel-2 ground sampling distance of 10 meters is still relatively large relative to the small field sizes of Switzerland (see Fig. 5), mixed pixels can have a noticeable impact at field boundaries, where stripes of meadow (grassland) and hedges are common (see Fig. B.14). Other examples of misclassifications are apple and pear, which belong to the same genus (family of Rosaceae). Similarly, sugar beets and beets (forage or vegetable) are the same species and just differentiated in the cultivar or variety. Finally, intermediate crops, cover crops and secondary crops are not part of the reference dataset, but are of course visible in the satellite imagery throughout the year. It is required by Swiss law to grow cover crops during winter on bare soils as part of sustainable practice to reduce environmental degradation, such as erosion caused by strong rainfall events or nutrient leaching during winter (Prasuhn, 2012). These intermediate or secondary crops are often, but not always, one of the sub-classes of the *grassland* class. We hypothesize that this might be another reason for very rare classes being wrongly associated with *grassland*.

Of course our proposed approach also has limitations, which we discuss in the following, together with ideas how to mitigate them in future work. Like many deep neural networks, the model requires a significant amount of training labels for optimal performance. One possible way to alleviate the consequences of label scarcity during training is to use an active learning scheme. I.e., iterate between model training and the selection of maximally informative samples to extend the training set (which must then be annotated by an expert). Active learning can significantly reduce the amount of training labels needed to reach a defined performance level, the price to pay is that the overall lower annotation effort becomes more protracted and less projectable. Also, data-driven models are intrinsically tied to the data used during training. A domain shift, caused for instance by a change of target crop types, will require a re-training or fine-tuning of the model. At least fine-tuning may also be required to maintain performance if the model is applied in an area with different ecological background conditions. Pairing our method with a meta-learning scheme, as recently explored in (Rußwurm et al., 2020), could ease the adaptation effort to different crop types and different geographic regions. Finally, the label hierarchy in this work is specific to Switzerland and different hierarchies will be needed for other agricultural regimes. In that context we reiterate that hard-coding a specific label hierarchy into the network architecture is an integral part of the current design, but lacks flexibility when moving to other environments. A technically challenging, but potentially useful extension to the current method could be to learn the label hierarchy from data rather than impose it a priori, using for instance unsupervised learning techniques. It can be expected that constructing the hierarchy in a data-driven manner, rather than with expert knowledge, will require fairly large amounts of data.

8. Conclusion

We have proposed a novel, hierarchical classification approach for multi-temporal crop classification from satellite images – in our case Sentinel-2, but the scheme is generic and could be applied to other sensors. Our ms-convSTAR method is a multi-stage convolutional recurrent neural network that leverages an explicit, hierarchical tree

structure of the labels. Besides classifying the input data simultaneously at multiple interdependent hierarchy levels of granularity, the method also features an CNN-based label-refinement component to favour consistency across the hierarchy.

In our study, the labels are based on the classes of the Swiss governmental reporting scheme, and the hierarchy was defined by domain experts according to agronomic knowledge and best practices. Based on that hierarchy we also have collected a new dataset of Sentinel-2 time series images, *ZueriCrop*, that densely covers a large agricultural region in Central Europe. The dataset is larger, more imbalanced, and more representative of real applications than earlier science datasets. In the experiments on *ZueriCrop*, ms-convSTAR has shown improved per-class performance and outperforms competing state-of-the-art methods. In particular, it greatly improves the classification of many rare classes.

Acknowledgments

We thank the Swiss Federal Office for Agriculture (FOAG) for partially funding this Research project through the DeepField Project.

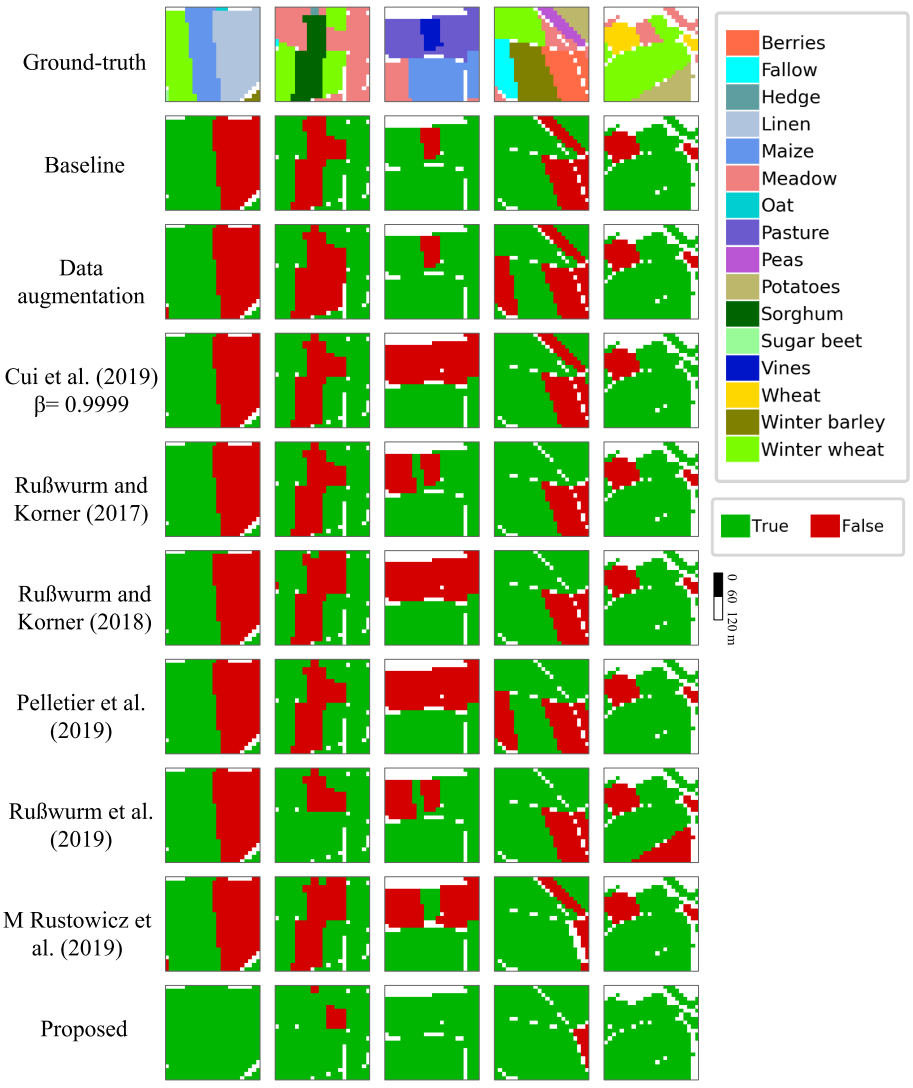


Figure 9: Visual results for five different samples (train-test fold 2). Green color indicates correct classification, red are mis-classified pixels (respectively, fields).

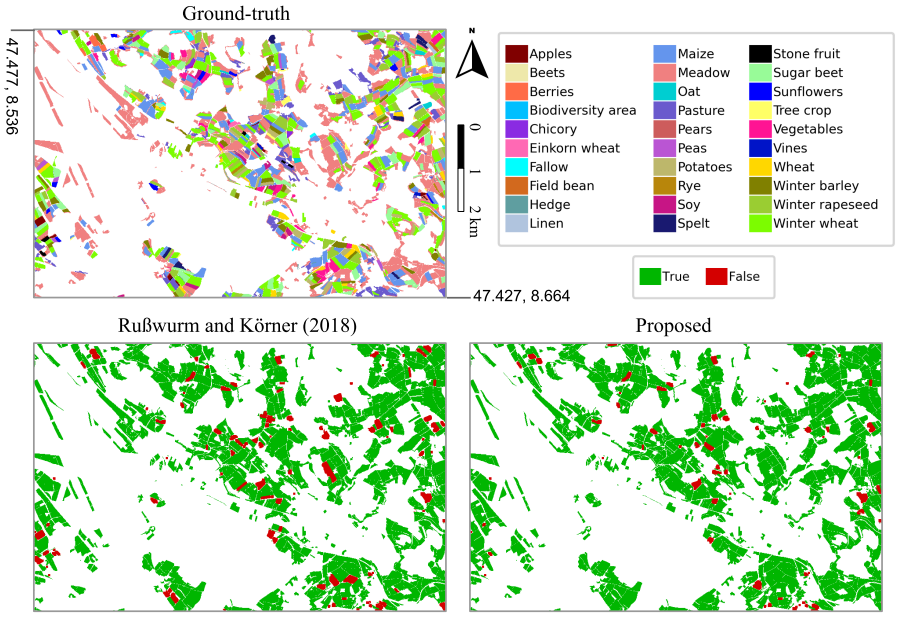
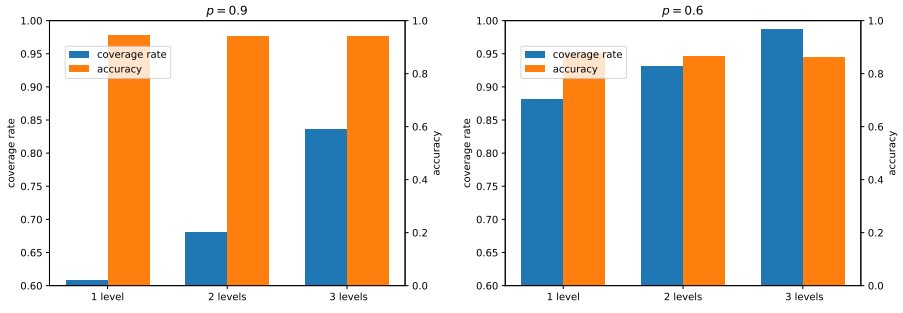
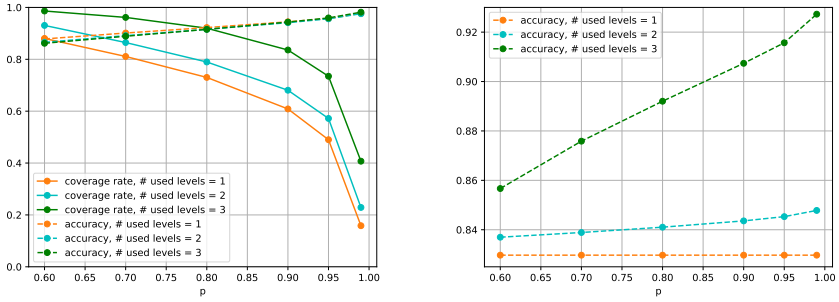


Figure 10: Qualitative comparison: ms-convSTAR vs. the best-performing alternative on the *ZueriCrop* data.



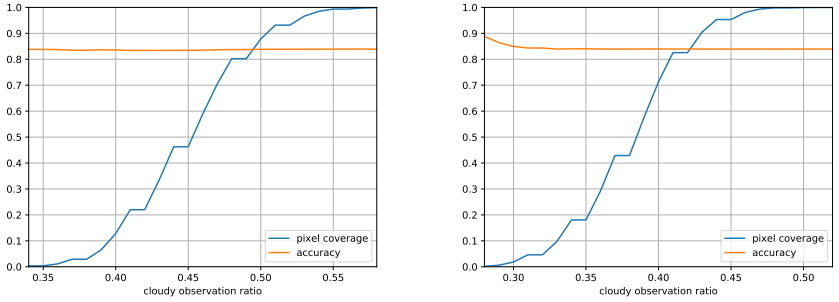
(a) $p = 0.9$ (b) $p = 0.6$

Figure 11: Performance of models with different numbers of hierarchical levels at fixed confidence values: (a) $p = 0.9$, (b) $p = 0.6$. The fraction of classified pixels (coverage rate, blue) grows with an increasing number N of hierarchy levels (computed with a single train-test fold). See Section 6.3.



(a) (b)

Figure 12: Performance of models with different numbers of hierarchical levels: (a) Coverage rate and model performance vs. confidence value p, (b) performance vs. confidence value p in full coverage case for models with different number of hierarchy levels (computed with a single train-test fold). See Section 6.3.



(a) $c = 0.1$ (b) $c = 0.5$

Figure 13: Robustness against clouds. We plot classification accuracy against the proportion of cloud-free observations. Along the x-axis, pixels are ordered according to how often they are obscured by clouds. The blue curve depicts the cumulative number of pixels with at most a certain fraction of cloudy views. The orange curve shows the accuracy of the model evaluated over those pixels. The test was repeated with two different thresholds c for the Sen2Cor cloud scores.

References

Anderegg, J., Yu, K., Aasen, H., Walter, A., Liebisch, F., Hund, A., 2020. Spectral vegetation indices to track senescence dynamics in diverse wheat germplasm. Frontiers in Plant Science 10, 1749.

Bailly, S., Giordano, S., Landrieu, L., Chehata, N., 2018. Crop-rotation structured classification using multi-source sentinel images and lpis for crop type mapping, in: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 1950–1953.

Belgiu, M., Csillik, O., 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. Remote sensing of environment 204, 509–523.

Bundesamt für Statistik, 2020. Landwirtschaft und Ernährung - Taschenstatistik 2020. Bern.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Chen, H.Y., Tsai, L.H., Chang, S.C., Pan, J.Y., Chen, Y.T., Wei, W., Juan, D.C., 2019. Learning with hierarchical complement objective. arXiv preprint arXiv:1911.07257 .

Chen, Y., Su, W., Li, J., Sun, Z., 2009. Hierarchical object oriented classification using very high resolution imagery and lidar data over urban areas. Advances in Space Research 43, 1101–1110.

Conrad, C., Dech, S., Dubovyk, O., Fritsch, S., Klein, D., Löw, F., Schorcht, G., Zeidler, J., 2014. Derivation of temporal windows for accurate crop discrimination in heterogeneous croplands of uzbekistan using multitemporal rapideye images. Computers and Electronics in Agriculture 103, 63–74.

Conrad, C., Fritsch, S., Zeidler, J., Rücker, G., Dech, S., 2010. Per-field irrigated crop classification in arid central asia using spot and aster data. Remote Sensing 2, 1035–1056.

Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S., 2019. Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Demirkan, D.Ç., Koz, A., Düzgün, H.S., 2020. Hierarchical classification of sentinel 2-a images for land use and land cover mapping and its use for the corine system. Journal of Applied Remote Sensing 14, 026524.

Dise, N.B., Ashmore, M., Belyazid, S., Bleeker, A., Bobbink, R., Vries, W.D., Erisman, J.W., Spranger, T., Stevens, C.J., Berg, L.V.D., 2011. Nitrogen as a threat to European terrestrial biodiversity. Cambridge University Press.

Dong, Q., Gong, S., Zhu, X., 2018. Imbalanced deep learning by minority class incremental rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 1367–1381.

Douzas, G., Bacao, F., 2018. Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Systems with applications 91, 464–471.

Finger, R., Lehmann, N., 2012. Policy reforms to promote efficient and sustainable water use in swiss agriculture. Water Policy 14, 887–901. doi:10.2166/wp.2012.152.

Flood, N., Watson, F., Collett, L., 2019. Using a u-net convolutional neural network to map woody vegetation extent from high resolution satellite imagery across queensland, australia. International Journal of Applied Earth Observation and Geoinformation 82, 101897.

Foerster, S., Kaden, K., Foerster, M., Itzerott, S., 2012. Crop type mapping using spectral–temporal profiles and phenological information. Computers and Electronics in Agriculture 89, 30–40.

Giller, K.E., Witter, E., Corbeels, M., Tittonell, P., 2009. Conservation agriculture and smallholder farming in Africa: The heretics' view. Field Crops Research 114, 23–34. doi:10.1016/j.fcr.2009.06.017.

Goel, A., Banerjee, B., Pižurica, A., 2018. Hierarchical metric learning for optical remote sensing scene categorization. IEEE Geoscience and Remote Sensing Letters 16, 952–956.

Gómez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: A review. ISPRS Journal of Photogrammetry and Remote Sensing 116, 55–72. doi:10.1016/j.isprsjprs.2016.03.008.

Gomiero, T., Pimentel, D., Paoletti, M.G., 2011. Environmental impact of different agricultural management practices: Conventional vs. Organic agriculture. Critical Reviews in Plant Sciences 30, 95–124. doi:10.1080/07352689.2011.554355.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering 21, 1263–1284.

Herzog, F., Prasuhn, V., Spiess, E., Richner, W., 2008. Environmental cross-compliance mitigates nitrogen and phosphorus pollution from Swiss agriculture. Environmental Science and Policy 11, 655–668. doi:10.1016/j.envsci.2008.06.003.

Heupel, K., Spengler, D., Itzerott, S., 2018. A progressive crop-type classification using multitemporal remote sensing data and phenological information. PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science 86, 53–69.

Huang, C., Li, Y., Change Loy, C., Tang, X., 2016. Learning deep representation for imbalanced classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5375–5384.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P., et al., 2015. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. Remote Sensing 7, 12356–12379.

Jiao, L., Sun, W., Yang, G., Ren, G., Liu, Y., 2019. A hierarchical classification framework of satellite multispectral/hyperspectral images for mapping coastal wetlands. Remote Sensing 11, 2238.

Khan, S., Hayat, M., Zamir, S.W., Shen, J., Shao, L., 2019. Striking the right balance with uncertainty, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 103–112.

Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R., 2017. Cost-sensitive learning of deep feature representations from imbalanced data. IEEE transactions on neural networks and learning systems 29, 3573–3587.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization, in: ICLR.

Koo, J., Klabjan, D., Utke, J., 2018. Combined convolutional and recurrent neural networks for hierarchical classification of images. arXiv preprint arXiv:1809.09574 .

Laurance, W.F., Sayer, J., Cassman, K.G., 2014. Agricultural expansion and its impacts on tropical nature. Trends in Ecology and Evolution 29, 107–116. URL: http://dx.doi.org/10.1016/j.tree.2013.12.001, doi:10.1016/j.tree.2013.12.001.

Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G., 2018. Videolstm convolves, attends and flows for action recognition. Computer Vision and Image Understanding 166, 41–50.

Ling, C.X., Sheng, V.S., 2008. Cost-sensitive learning and the class imbalance problem. Encyclopedia of machine learning 2011, 231–235.

Mao, Y., Tian, J., Han, J., Ren, X., 2019. Hierarchical text classification with reinforced label assignment. arXiv preprint arXiv:1908.10419 .

Melgani, F., Bruzzone, L., 2004. Classification of hyperspectral remote sensing images with support vector machines. IEEE Transactions on geoscience and remote sensing 42, 1778–1790.

Pelletier, C., Webb, G.I., Petitjean, F., 2019. Temporal convolutional neural network for the classification of satellite image time series. Remote Sensing 11, 523.

Peña-Barragán, J.M., Ngugi, M.K., Plant, R.E., Six, J., 2011. Object-based crop identification using multiple vegetation indices, textural features and crop phenology. Remote Sensing of Environment 115, 1301–1316.

Prasuhn, V., 2012. On-farm effects of tillage and crops on soil erosion measured over 10 years in Switzerland. Soil and Tillage Research 120, 137–146. URL: http://dx.doi.org/10.1016/j.still.2012.01.002, doi:10.1016/j.still.2012.01.002.

Ren, M., Zeng, W., Yang, B., Urtasun, R., 2018. Learning to reweight examples for robust deep learning, in: ICML.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Roy, D., Panda, P., Roy, K., 2020. Tree-cnn: a hierarchical deep convolutional neural network for incremental learning. Neural Networks 121, 148–160.

Rußwurm, M., Körner, M., 2017. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Rußwurm, M., Körner, M., 2018a. Convolutional lstms for cloud-robust segmentation of remote sensing imagery, in: NIPS Workshop.

Rußwurm, M., Körner, M., 2018b. Multi-temporal land cover classification with sequential recurrent encoders. ISPRS International Journal of Geo-Information 7, 129.

Rußwurm, M., Körner, M., 2020. Self-attention for raw optical satellite time series classification. ISPRS Journal of Photogrammetry and Remote Sensing 169, 421–435.

Rußwurm, M., Lefèvre, S., Körner, M., 2019. Breizhcrops: A satellite time series dataset for crop type identification, in: ICML Workshop.

Rußwurm, M., Wang, S., Korner, M., Lobell, D., 2020. Meta-learning for few-shot land cover classification, in: Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops, pp. 200–201.

Rustowicz, R., Cheong, R., Wang, L., Ermon, S., Burke, M., Lobell, D., 2019. Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Saha, P., Dash, S., Mukhopadhyay, S., 2020. Phicnet: Physics-incorporated convolutional recurrent neural networks for modeling dynamical systems. arXiv preprint arXiv:2004.06243 .

Sainte Fare Garnot, V., Landrieu, L., Giordano, S., Chehata, N., 2019. Time-space tradeoff in deep learning models for crop classification on satellite multi-spectral image time series, in: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 6247–6250.

Sainte Fare Garnot, V., Landrieu, L., Giordano, S., Chehata, N., 2020. Satellite image time series classification with pixel-set encoders and temporal self-attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12325–12334.

Siachalou, S., Mallinis, G., Tsakiri-Strati, M., 2015. A hidden markov models approach for crop classification: Linking crop phenology to time series of multi-sensor remote sensing data. Remote Sensing 7, 3633–3650.

Siam, M., Valipour, S., Jagersand, M., Ray, N., 2017. Convolutional gated recurrent networks for video segmentation, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 3090–3094.

Srivastava, N., Salakhutdinov, R.R., 2013. Discriminative transfer learning with tree-based priors, in: Advances in neural information processing systems, pp. 2094–2102.

Stoian, A., Poulain, V., Inglada, J., Poughon, V., Derksen, D., 2019. Land cover maps production with high resolution satellite image time series and convolutional neural networks: Adaptations and limits for operational systems. Remote Sensing 11, 1986.

Stumpf, F., Schneider, M.K., Keller, A., Mayr, A., Rentschler, T., Meuli, R.G., Schaepman, M., Liebisch, F., 2020. Spatial monitoring of grassland management using multi-temporal satellite imagery. Ecological Indicators 113, 106201. URL: http://weekly.cnbnews.com/news/article.html?no=124000https://linkinghub.elsevier.com/retrieve/pii/S1470160X20301382, doi:10.1016/j.ecolind.2020.106201.

Su, J., Byeon, W., Huang, F., Kautz, J., Anandkumar, A., 2020. Convolutional tensor-train lstm for spatio-temporal learning. arXiv preprint arXiv:2002.09131 .

Sulla-Menashe, D., Friedl, M.A., Krankina, O.N., Baccini, A., Woodcock, C.E., Sibley, A., Sun, G., Kharuk, V., Elsakov, V., 2011. Hierarchical mapping of northern eurasian land cover using modis data. Remote Sensing of Environment 115, 392–403.

Sulla-Menashe, D., Gray, J.M., Abercrombie, S.P., Friedl, M.A., 2019. Hierarchical mapping of annual global land cover 2001 to present: The modis collection 6 land cover product. Remote Sensing of Environment 222, 183–194.

Thenkabail, P.S., Mariotto, I., Gumma, M.K., Middleton, E.M., Landis, D.R., Huemmrich, K.F., 2013. Selection of hyperspectral narrowbands (hnbs) and composition of hyperspectral twoband vegetation indices (HVIS) for biophysical characterization and discrimination of crop types using field reflectance and hyperion/EO-1 data.

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6, 427–439. doi:10.1109/JSTARS.2013.2252601.

Turkoglu, M.O., D'Aronco, S., Wegner, J., Schindler, K., 2021. Gating revisited: Deep multi-layer rnns that can be trained. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Ustuner, M., Sanli, F.B., Abdikan, S., Esetlili, M., Kurucu, Y., 2014. Crop type classification using vegetation indices of rapideye imagery. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 40, 195.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

Vuolo, F., Neuwirth, M., Immitzer, M., Atzberger, C., Ng, W.T., 2018. How much does multi-temporal sentinel-2 data improve crop type classification? International journal of applied earth observation and geoinformation 72, 122–130.

Walter, A., Liebisch, F., Hund, A., 2015. Plant phenotyping: from bean weighing to image analysis. Plant methods 11, 1–11.

Wang, Y.X., Ramanan, D., Hebert, M., 2017. Learning to model the tail, in: Advances in Neural Information Processing Systems, pp. 7029–7039.

Wardlow, B.D., Egbert, S.L., 2008. Large-area crop mapping using time-series modis 250 m ndvi data: An assessment for the us central great plains. Remote sensing of environment 112, 1096–1116.

Wehrmann, J., Cerri, R., Barros, R., 2018. Hierarchical multi-label classification networks, in: International Conference on Machine Learning, pp. 5075–5084.

Wu, M., Sun, Z., Yang, B., Yu, S., 2016. A hierarchical object-oriented urban land cover classification using worldview-2 imagery and airborne lidar data, in: IOP Conference Series: Earth and Environmental Science, IOP Publishing. p. 012016.

Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A., 2010. Sun database: Large-scale scene recognition from abbey to zoo, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3485–3492.

Xiao, T., Zhang, J., Yang, K., Peng, Y., Zhang, Z., 2014. Error-driven incremental learning in deep convolutional neural network for large-scale image classification, in: Proceedings of the ACM International Conference on Multimedia, pp. 177–186.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in: Advances in neural information processing systems.

Yan, Z., Zhang, H., Piramuthu, R., Jagadeesh, V., DeCoste, D., Di, W., Yu, Y., 2015. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2740–2748.

Zhong, L., Hu, L., Zhou, H., 2019. Deep learning based multi-temporal crop classification. Remote Sensing of Environment 221, 430–443.

Zhu, X., Bain, M., 2017. B-cnn: branch convolutional neural network for hierarchical classification. arXiv preprint arXiv:1709.09890 .

Appendix A. Performance comparison: Without polygon aggregation

We provide models' performances when polygon aggregation, i.e., majority voting is not performed as post-processing, in Table A.6. Compared to results in Table 2, scores for each method are decreased in a similar proportion especially in terms of overall accuracy and F1-score.

Method	**Prec (%)**	**Rec (%)**	**F1 (%)**	**Acc (%)**
Random Forest*	43.6	<u>36.2</u>	32.6	69.8
LSTM (Rußwurm and Körner, 2017)	33.9	25.2	26.5	80.2
TCN (Pelletier et al., 2019)	42.1	25.4	27.2	80.5
Transformer (Rußwurm et al., 2019)	**53.7**	34.7	<u>38.7</u>	82.0
2D-CNN (U-Net)	33.8	23.8	24.6	79.6
U-Net+convLSTM (Rustowicz et al., 2019)	44.1	31.2	31.7	81.1
Bi-convGRU (Rußwurm and Körner, 2018b)	49.1	35.7	38.4	<u>82.4</u>
ms-convSTAR	<u>53.0</u>	**44.2**	**46.2**	**83.9**

Table A.6: **Without polygon aggregation:** Performance comparison of ms-convSTAR (bottom row) with state-of-the-art methods when polygon aggregation, i.e., majority voting, is not performed as post-processing. Precision, recall and F1-score are mean values over all classes. All numbers are averaged over 5 cross-validation folds. The best score for each metric is shown with **bold** and the second best is <u>underlined</u>. *Random Forest is trained with a balanced dataset by under-sampling the majority classes which leads to improved class-wise performance.

Appendix B. Number of Input Channels

In order to motivate our decision to use exclusively the 10m bands we conducted a simple experiment where we compare the performance of our method and of a Random Forest (RF) when using exclusively the 4 10m bands, and when using both the 10m and 20m bands. Results are reported in Table B.7. As it can be seen, at least for the considered dataset, the improvements are marginal for the RF, and we actually observe a small performance drop for our method when using all the 9 bands. In terms of visual results, prediction comparison between 4 and 9 bands model are shown in Fig. 3. As can be seen the main difference is that the 9 bands model have some pixels that are misclassified on the edges of the fields. It is difficult to judge whether in this case the label is not fully accurate on the edges of the fields, or if the prediction is not accurate because of the limited resolution of the additional bands.

Appendix C. Class-wise Performance

Class-wise accuracy for each granularity is given in Table C.8.

Appendix D. More Details about Baselines

Parameters of baseline models and links for their source codes are given in Table D.9 and Table D.10, respectively.

Method	Prec (%)	Rec (%)	F1 (%)	Acc (%)
RF (9 channels)	46.5	40.9	39.0	78.9
RF (4 channels)	46.4	40.7	38.9	78.8
ms-convSTAR (9 channels)	56.3	46.6	48.7	87.8
ms-convSTAR (4 channels)	**59.8**	**49.7**	**52.1**	**88.0**

Table B.7: Performance comparison w.r.t. number of input channels. 5-fold evaluation (bi-cubic upsampling of all channels with 20 meter meter GSD).

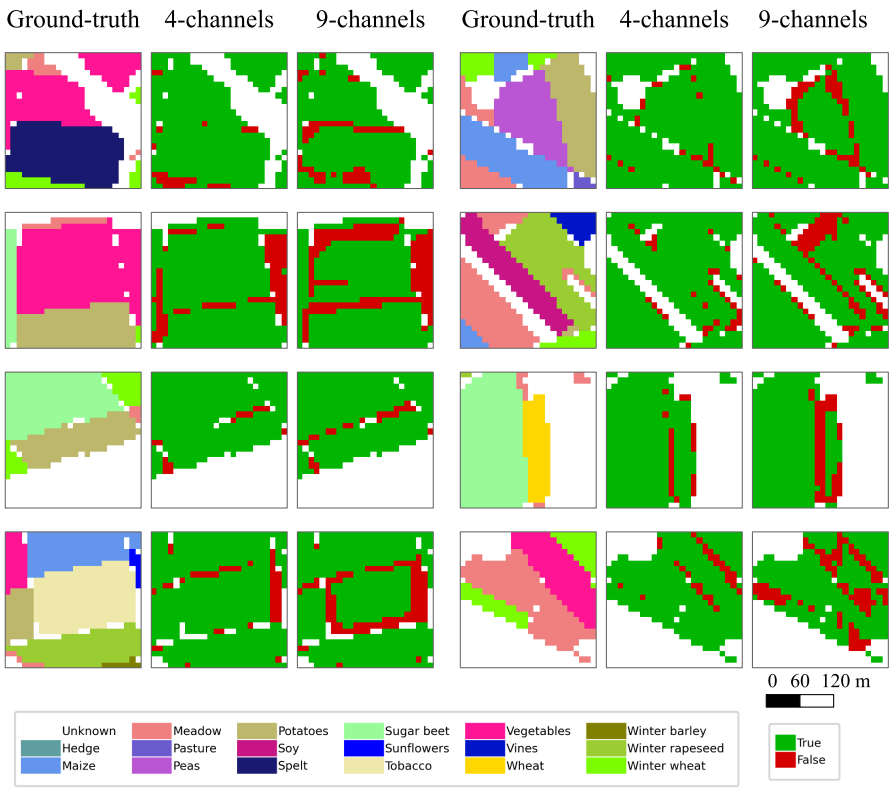


Figure B.14: Qualitative comparison: Example failure cases of the 4-channels model vs the 9-channels model. Polygon aggregation is not performed as post-processing. Both models often fail at edge cases; however, the 9-channels model makes more mistakes at field edges.

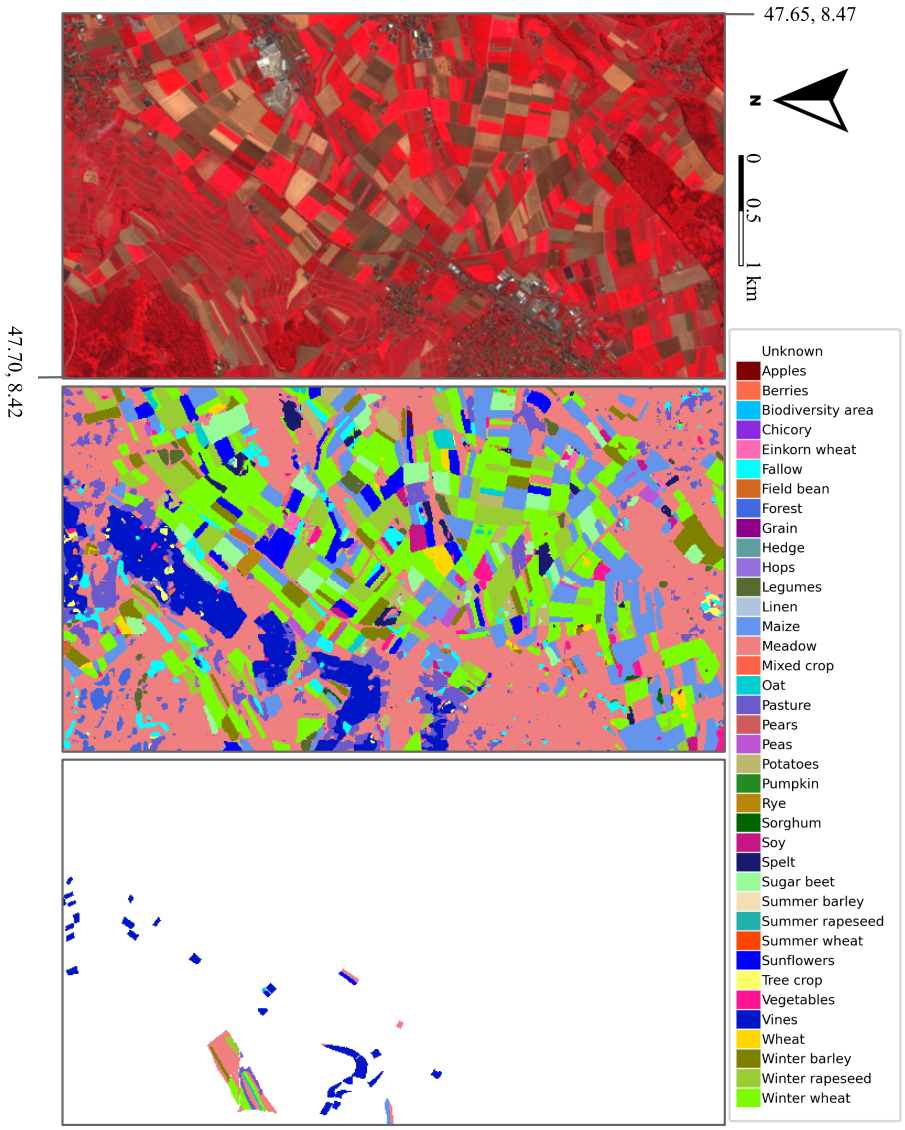


Figure B.15: Qualitative result: The first row shows the Sentinel-2 satellite image (false color composite: NIR-Red-Green) of the test area, the second row shows the crop map generated by the proposed ms-convSTAR, and the third row shows the ground truth map.

Level-1	Acc (%)	Level-2	Acc (%)	Level-3	Acc (%)
Field Crops (41.39%)	98.2	Small Grain Cereal (16.89%)	97.7	Summer Wheat (0.13%)	38.0
				Winter Wheat (10.77%)	95.7
				Wheat (1.20%)	39.4
				Einkorn Wheat (0.14%)	49.8
				Summer Barley (0.05%)	26.1
				Winter Barley (3.49%)	96.4
				Grain (0.05%)	21.7
				Rye (0.14%)	68.1
				Spelt (0.70%)	84.0
				Oat (0.21%)	76.4
				Buckwheat (0.02%)	7.0
		Large Grain Cereal (10.37%)	94.3	Maize (10.30%)	94.7
				Sorghum (0.07%)	27.0
		Vegetable Crop (2.85%)	83.4	Vegetables (2.76%)	82.3
				Pumpkin (0.03%)	14.2
				Chicory (0.07%)	30.8
		Broad Leaf Row Crop (11.20%)	95.1	Sugar Beet (4.26%)	97.2
				Beets (0.01%)	0.0
				Potatoes (1.65%)	86.6
				Sunflowers (1.05%)	92.8
				Linen (0.04%)	54.4
				Hemp (0.02%)	3.1
				Soy (0.37%)	87.4
				Winter Rapeseed (3.21%)	97.6
				Summer Rapeseed (0.02%)	0.0
				Field Bean (0.15%)	88.3
				Peas (0.32%)	82.4
				Lupine (< 0.01%)	0.0
				Tobacco (0.03%)	48.5
				Mustard (< 0.01%)	0.0
				Legumes (0.05%)	78.0
		Crop Mix (0.08%)	16.7	Crop Mix (0.08%)	39.7
Grassland (56.00%)	97.9	Meadow (47.76%)	95.3	Meadow (47.76%)	94.9
		Pasture (8.20%)	37.4	Pasture (8.20%)	42.4
		Biodiversity A. (0.04%)	0.0	Biodiversity A. (0.04%)	0.0
Orchards (1.54%)	49.6	Orchard Crop (1.41%)	76.3	Apples (0.40%)	69.3
				Pears (0.07%)	2.3
				Vines (0.83%)	82.2
				Hops (0.01%)	0.0
				Stone Fruit (0.09%)	24.1
				Chestnut (< 0.01%)	0.0
		Tree Crop (0.13%)	20.9	Tree Crop (0.13%)	37.2
Special Crops (1.00%)	8.3	Hedge, Gardens, Multiple (0.43%)	1.0	Hedge (0.42%)	1.5
				Gardens (< 0.01%)	0.0
				Multiple (0.01%)	0.0
		Berries (0.22%)	18.8	Berries (0.22%)	25.8
		Fallow (0.36%)	65.8	Fallow (0.36%)	70.3
Forest (0.07%)	0.0	Forest (0.07%)	0.0	Forest (0.07%)	1.6

Table C.8: Class-wise performance of proposed ms-convSTAR at different levels of granularity. All numbers are averaged over 5 cross-validation folds. Class frequencies are given in parenthesis.

Method	Hidden	Layer	Kernel	Batch	Note
Random Forest					max_depth:60 max_features:auto min_sample_leaf:1 min_sample_split:3 n_estimators:1000
LSTM	128	1		576	
TCN	64	3	5	576	dropout: 0.5
Transformer	64	4		576	head: 4
Bi-convGRU	64	1	3x3	4	
U-Net		12	3x3	4	filters: 64, 64, 128, 128, 256, 256
U-Net+convLSTM	256	12+1	3x3	4	filters: 64, 64, 128, 128, 256, 256

Table D.9: Parameters of baseline methods.

Method	Source
Random Forest	`https://scikit-learn.org`
LSTM	`https://pytorch.org/docs/stable/nn.html`
TCN	`https://github.com/charlotte-pel/temporalCNN`
Transformer	`https://github.com/dl4sits/BreizhCrops`
Bi-convGRU	`https://github.com/TUM-LMF/MTLCC`
U-Net	`https://github.com/roserustowicz/crop-type-mapping`
U-Net+convLSTM	`https://github.com/roserustowicz/crop-type-mapping`

Table D.10: Code sources of baseline methods.