Leveraging Bioclimatic Context for Supervised and Self-Supervised Land Cover Classification*

Johannes Leonhardt[1 [0000−0002−4505−5086]], Lukas Drees[1 [0000−0003−2052−1914]], Jürgen Gall[1,2 [0000−0002−9447−3399]], and Ribana Roscher[3,1 [0000−0003−0094−6210]]

[1] University of Bonn, Germany
{jleonhardt,ldrees,jgall,ribana.roscher}@uni-bonn.de
[2] Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
[3] Forschungszentrum Jülich GmbH, Germany

Abstract. Modern neural networks achieve state-of-the-art results on land cover classification from satellite imagery, as is the case for almost all vision tasks. One of the main challenges in this context is dealing with geographic variability in both image and label distributions. To tackle this problem, we study the effectiveness of incorporating bioclimatic information into neural network training and prediction. Such auxiliary data can easily be extracted from freely available rasters at satellite images' georeferenced locations. We compare two methods of incorporation, learned embeddings and conditional batch normalization, to a bioclimate-agnostic baseline ResNet18. In our experiments on the EuroSAT and BigEarthNet datasets, we find that especially the use of conditional batch normalization improves the network's overall accuracy, generalizability, as well as training efficiency, in both a supervised and a self-supervised learning setup. Code and data are publicly available at https://t.ly/NDQFF.

Keywords: Remote Sensing · Land Cover Classification · Multi-Modal Learning · Data Shift · Conditional Batch Normalization

1 Introduction

Land cover data has diverse applications, including the study of climate change, resource and disaster management, as well as spatial planning, stressing its importance for both science and practice [41,48]. Therefore, the extraction of land cover data from satellite imagery is one of the most extensively studied tasks in remote sensing and Earth observation. Recent developments in satellite technology and deep learning methods have further amplified this area of study,

* This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1502/1-2022 - Projektnummer: 450058266 and partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 – 390732324

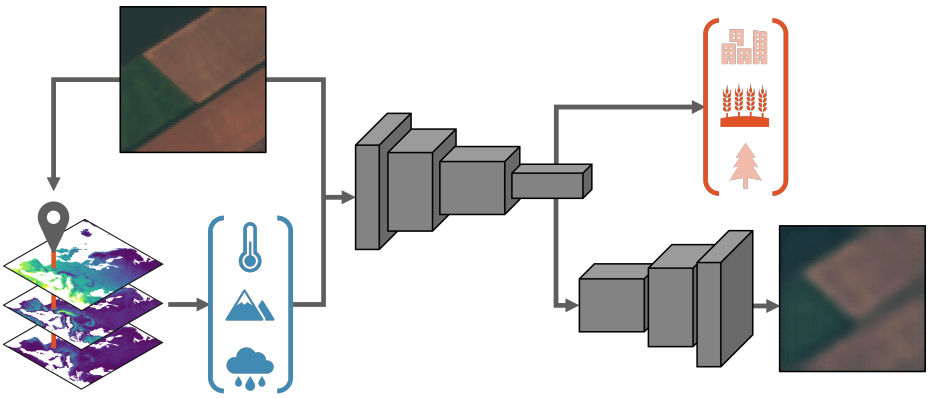


Fig. 1: Schematic summary of the presented approach: Satellite images' georeferencing are used to extract bioclimatic information like temperature, elevation, and percipitation (blue) from geographical rasters for each image. This is provided to the neural networks as an additional input modality. In the supervised setting (top), an encoder and a classification head directly predict the land cover class label (red). In the self-supervised setting (bottom), the encoder is pretrained using a symmetric decoder for image reconstruction.

boosting the accuracy, as well as the spatial and temporal resolution of current land cover products.

One of the main challenges of land cover mapping remains the great geographic variability in both the distributions of images X and class labels Y. This raises questions about both the accuracy and generalizability of the resulting models [16,26,44]. To tackle this issue, we draw a connection between land cover and bioclimatology – the study of the physical environment's effects on life on Earth [43] – and leverage bioclimatic auxiliary information in a multi-modal learning approach to support land cover classification models. We argue that by explicitly incorporating bioclimatic context, models learn representations, which are more invariant to bioclimatic context, and thus more discriminative with regard to the land cover-relevant image content.

In particular, we point out two connected, yet distinct interactions between the two: First, some land cover classes are more prevalent in some bioclimatic regimes: In Boreal climates, for instance, one is much more likely to observe the class *Coniferous Forest* than in Atlantic climates, where the class *Pastures* may be much more prevalent, instead. We call this effect bioclimatic prior shift $P(Y|A = a_1) \neq P(Y|A = a_2)$, as it can be described by an inequality of prior class probabilities depending on the regional bioclimatic conditions A.

Second, we observe that land cover classes can appear differently depending on the bioclimatic circumstances: As an example, the kinds of crops cultivated in different bioclimates have a big effect on the appearance of classes like *Annual Crop* and *Perennial Crop*. As this effect describes a change in the distribution

of the inputs with respect to A, we call this effect bioclimatic covariate shift $P(X|A = a_1) \neq P(X|A = a_2)$.

To evaluate bioclimate-aware neural networks' abilities to counteract these effects, we first conduct extensive experiments on EuroSAT, a Europe-wide benchmark for land cover classification [15]. We find that models which leverage bioclimatic data by means of conditional batch normalization [6] reliably outperform both the model with added learned embeddings and the bioclimate-agnostic model with respect to overall accuracy, generalizability, and training efficiency.

Besides the supervised case, we also consider a self-supervised learning setup, where the encoder is trained on the pretext task of image reconstruction [21]. We find that the resulting representations are better separable with respect to land cover, if bioclimatic data are incorporated into the training through conditional batch normalization. Furthermore, we achieve additional improvements for the supervised setting by initializing the model with the pre-trained weights.

At the same time, we show that the bioclimate-aware models do not suffer from the recently described pitfall of shortcut learning [34], which we examine using the Grad-CAM tool for model interpretability [33].

We finally provide a brief outlook regarding the approach's applicability in more large-scale and complex settings. To this end, we show that classification accuracy is also improved for supervised training on the much larger, multi-label BigEarthNet [37,38] dataset.

A graphical outline of our approach is presented in Fig. 1.

2 Related Work

Improving neural networks by learning from multiple data modalities at once is an intensely studied topic in the field of deep learning [31]. While many pioneering works of the field have focused on image or video understanding tasks based on additional text or audio [6,28,30,35], the framework has since been applied to many other applications and modalities like medicine [4,25] or robotics [19,22].

A popular method specifically designed for fusing complex structured data with context data from other modalities is conditional batch normalization [6]. The method has since proven its effectiveness for many tasks such as style transfer [8,17], super-resolution [45], domain adaptation [24], and image synthesis [29,50].

Although most land cover classification approaches rely on images as the only input modality [7,15,37], multi-modal learning approaches are of great interest to the domain of remote sensing. Most importantly, RGB- or multispectral data have been combined with other types of imagery like SAR [1,38], Lidar data [2], or street-level imagery [36]. Our work differs from such approaches as we do not fuse satellite imagery with another image sensor modality, but with bioclimatic auxiliary information, which is easily obtainable and universally applicable to a wide range of remote sensing data and applications.

More recently, co-georeferenced, multi-modal datasets have been used explicitly for self-supervised neural network pretraining [14,32]. In analogy to our

approach, the additional modality can sometimes be derived either directly or indirectly from the images' georeferencing. For instance, latitude and longitude information have been used to not only improve classification accuracy of geotagged images [40], but also to design a geography-aware self-supervised contrastive training objective, which was also applied to remote sensing images [3]. Similarly, satellite images have been enriched with land cover statistics derived from existing products to aid general representation learning [23]. The difference between these works and our approach, however, is that we do not use the bioclimatic context information for deriving an entirely new objective, but explicitly provide it to the network as an additional input modality, while relying on standard, non-geographic loss functions in both the supervised and self-supervised setting. This makes our approach less specific to certain realizations of self-supervised learning and thus, more universally applicable to a wide range of training schemes. In addition, it is safe to assume that georeferencing information is available for most satellite remote sensing, but not natural image datasets.

Closely related to our work, conditional regularization approaches have been adapted for some applications in Earth observation: For predicting wildfires from various static and dynamic geographical variables, for example, location-aware denormalization layers are used in conjunction with a multi-branch neural network [9]. It is also used to extract geometric information from auxiliary GIS data for individual building segmentation in SAR images [39] and for generating high-resolution satellite images from a set of semantic descriptors [27].

3 Methodology

3.1 Data Preparation

We conduct the majority of our experiments on the EuroSAT benchmark for satellite image land cover classification [15]. In this dataset, there are 27000 georeferenced images x from the Sentinel-2 mission, which are distributed across the European continent. The images have a size of 64×64 pixels and 13 spectral bands in the visible and infrared domains. The bands are originally at different spatial resolutions between 10 m and 60 m, but have been sampled to a common 10 m grid using bicubic interpolation. Each of the images is associated with one land cover label y, distinguishing between the 10 classes *Annual Crop*, *Forest*, *Herbaceous Vegetation*, *Highway*, *Industrial*, *Pasture*, *Perennial Crop*, *Residential*, *River*, and *Sea & Lake*.

In addition, we use BigEarthNet [37,38] as an example of a larger-scale and more complex dataset. Like EuroSAT, it contains pairs of Sentinel-2 images and land cover class labels across Europe. BigEarthNet is, however, much larger at 519341 120×120 pixel images. Furthermore, it is a multi-label dataset with 19 land cover classes with the possibility of an image being assigned to multiple labels. For a list of classes, we refer to Fig. 2.

For both EuroSAT and BigEarthNet, bioclimatic auxiliary vectors a are derived from the WorldClim-BIO dataset [12]. It contains worldwide rasters of

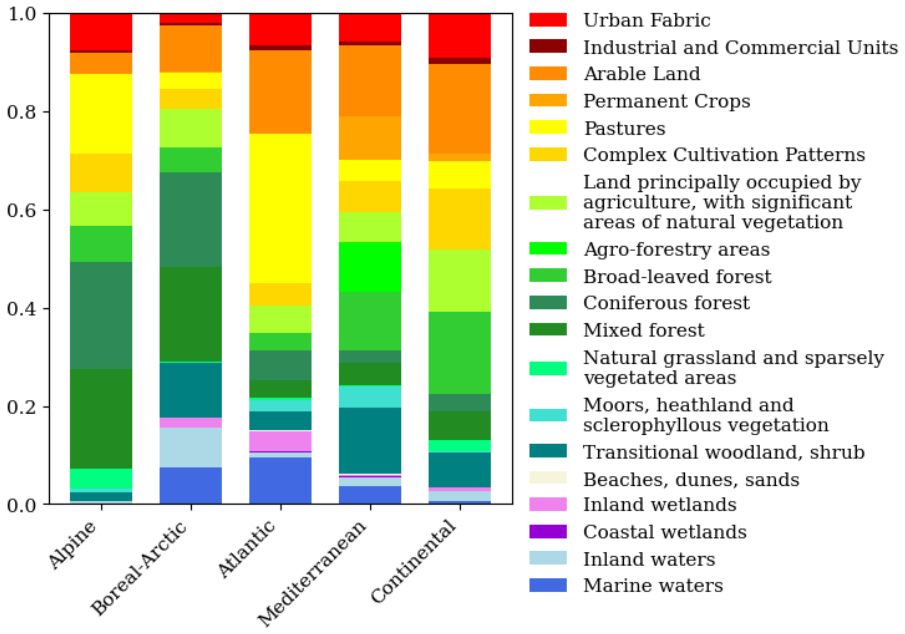
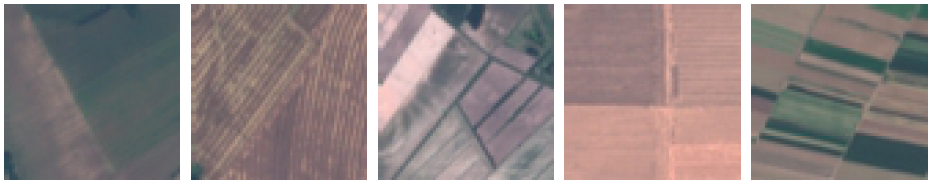


Fig. 2: The variability of land cover class label distributions across the different biogeographic regions in BigEarthNet provides an indication of prior shift.

a total of 19 bioclimatic variables and one additional elevation raster from the shuttle radar topography mission [11] at a resolution of 10 arcmin, which roughly equates to 18.5 km in north-south direction, and between 8 km and 15 km in west-east direction for our region of interest. Among the bioclimatic variables are e.g. the mean annual temperature, temperature seasonality, and annual precipitation. For a full list, we refer to the link in the corresponding reference [12]. To assign a set of values to a specific image, the rasters are sampled at the center location of the image using nearest neighbor interpolation. As the images are much higher-resolved than the auxiliary rasters, variability of climatic variables across the footprint area of single images is negligible.

Similarly, we also provide each image with a biogeographic label b, which is derived in a similar fashion from maps by the European Environmental Agency [10]. We differentiate between the five regions *Alpine*, *Boreal-Arctic*, *Atlantic*, *Mediterranean* (including *Macaronesia*, *Steppic* and *Black Sea*), and *Continental* (including *Pannonian*) as a proxy for bioclimatic variability. These data, however, will not be used during the training of the neural network, but only serve data split and evaluation purposes. It must be stressed, that the extraction of both a and b is straightforward, as satellite images' locations are contained in their georeferencing metadata in the majority of cases. The approach is thereby generally transferable to any georeferenced dataset.



(a) Alpine (b) Boreal-Arctic (c) Atlantic (d) Mediterranean (e) Continental

Fig. 3: The variability of example EuroSAT images for the class *Annual Crop* from different biogeographical regions provide indication of covariate shift.

For training, validating, and testing the neural network models, both datasets are split into fixed subsets of 60%, 20%, and 20%, respectively. We perform the split with respect to the biogeographic label b, instead of the labels y as is originally the case for EuroSAT [15]. This deviation is to ensure even bioclimatic diversity throughout the data splits, so that the models can be reliably evaluated regarding their geographic generalizability later on. The satellite images' pixel values, as well as the bioclimatic auxiliary vectors are normalized to the range $[0, 1]$ and we apply random horizontal and vertical flips to the images during training and validation for data augmentation purposes.

To illustrate the presence of the previously discussed bioclimatic data shift effects in the considered datasets, we provide some visual examples: First, the BigEarthNet class distributions for the different biogeographic regions are shown in Fig. 2, providing an indication for bioclimatic prior shift. Second, we show one EuroSAT image of the class *Annual Crop* for each of the biogeographic regions in Fig. 3 to highlight the presence of bioclimatic covariate shift. We note that there may also be other reasons for data shift, such as sampling bias or other unconsidered environmental variables, which are out of the scope of this work.

3.2 Neural Network Architectures and Training

Given the data quadruples $\{x, y, a, b\}$ as described above, a neural network shall predict a land cover label from the image, i.e. approximate the conditional distribution $P(Y|X)$. The most important component of prediction is a fully convolutional image encoder $\mathcal{E}$, which yields an intermediate representation $z = \mathcal{E}(x)$. We use the ResNet18 architecture [13], where the original input and pooling layer is replaced by a single layer of strided convolutions with subsequent batch normalization in order to adapt the architecture to the respective dataset's image size.

In the supervised setting, the network is completed by a one-layer classification head $\mathcal{C}$ to output land cover class label predictions $\widehat{y} = \mathcal{C}(z)$. The two model components are trained in an end-to-end fashion to minimize the Cross Entropy Loss between $\widehat{y}$ and y using the Adam optimizer [20]. As for the hyperparameters, we use a learning rate of 1×10^{-5} for both datasets and a batch

size of 64 and 512 for EuroSAT and BigEarthNet, respectively. To determine the optimal number of epochs needed for training, early stopping with a patience of 5 epochs for EuroSAT and 10 for BigEarthNet based on the validation accuracy is utilized.

For the self-supervised case, which is only employed on EuroSAT, the encoder is trained on the pretext task of image reconstruction [21]. To this end, the encoder is complemented by a ResNet18 Decoder $\mathcal{D}$, which mirrors the encoder's structure using transpose convolutions in place of convolution and pooling layers [47]. For the reconstruction loss, we use a weighted average of a base L1 loss (l_1), which only considers pixel-wise differences and the negative structural similarity index measure (l_{SSIM}), which also reflects structural differences between the images [46,49]:

$$l_{\text{rec.}}(x, \widehat{x}) = (1 - \lambda) l_1(x, \widehat{x}) + \lambda(1 - l_{\text{SSIM}}(x, \widehat{x})). \tag{1}$$

The parameters of the model are also optimized end-to-end with Adam, but the number of epochs is constantly set to 200. For the learning rate and the batch size, we use the same specifications as described above and the additional weighting hyperparameter is set to $\lambda = 0.1$ for our experiments.

After training, the bottleneck representation z must preserve as much visual information about the image as possible in order for the decoder to be able to reconstruct the image. As a result, the trained encoder serves as a reasonable initialization for the subsequent supervised downstream task of land cover class prediction, for which training is otherwise performed as in the supervised setting described above.

3.3 Leveraging Bioclimatic Data

In order to alleviate the issues regarding data shift in the neural networks, the models are not only given the images x, but also the bioclimatic auxiliary vector a. Thereby, the resulting models explicitly approximate $P(Y|X, A)$. We explore two methods for incorporating the bioclimatic context into the neural networks:

In the first case, we apply a fully connected linear layer $\mathcal{M}$ with learnable parameters to a. This embedding of the bioclimatic context is then simply added to the encoder output to yield the intermediate representation $z = \mathcal{E}(x) + \mathcal{M}(a)$.

As the second method for leveraging the bioclimatic context, we replace all batch normalization layers in both the encoder and the decoder with conditional batch normalization layers. Batch normalization in itself is one of the most popular regularization techniques in deep neural networks [18]. Its aim is to reduce the internal covariate shift of a layer's hidden representations h_{ij}, where the index i refers to the batch dimension, and j refers to the feature or channel dimension. First, they are standardized using the features' means μ_j and variances σ_j^2, for which running estimates are stored:

$$h_{ij} \leftarrow \frac{h_{ij} - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}}. \tag{2}$$

Afterward, they are rescaled by learnable scale and offset parameters γ_j and β_j:

$$h_{ij} \leftarrow \gamma_j h_{ij} + \beta_j. \tag{3}$$

As an extension to regular batch normalization, conditional batch normalization [6] makes these layers' parameters dependent on some auxiliary data which in our application is the images' bioclimatic context a_i. Thus, the learnable parameters are not optimized for directly, but as the result of a fully connected linear layer $\mathcal{T}$, which is individually applied to each a_i:

$$\gamma_{ij}, \beta_{ij} = \mathcal{T}(a_i). \tag{4}$$

Note that in conditional batch normalization, the learned scale and bias parameters are different for the samples within a batch, which is not the case in standard batch normalization. The statistics μ_j and σ_j^2, on the other hand, are still computed and applied independently of a_i.

The resulting bioclimate-aware neural networks, abbreviated as ResNet18-Emb and ResNet18-CBN, can be trained in the exact same fashion as the bioclimate-agnostic ResNet18, allowing for a direct comparison between the three different methods in both the supervised and the self-supervised setting.

4 Experiments and Evaluation

4.1 Accuracy

First, we train and test the different neural networks on EuroSAT. We compare the overall accuracy, $\mathrm{Acc}(y, \widehat{y}) = \frac{1}{N} \sum_{i=0}^{N} 1(y_i = \widehat{y}_i)$ with N denoting the number of samples, of the bioclimate-aware and bioclimate-agnostic models in Tab. 1. All models are trained from both random initializations and initializations from self-supervised pre-training, which we simply refer to as the supervised and self-supervised settings, respectively.

Table 1: Averages and standard deviations of test overall accuracy across 10 train runs on EuroSAT.

Model	Acc, Supervised in %	Acc, Self-Supervised in %
ResNet18	95.68 ± 0.41	96.04 ± 0.36
ResNet18-Emb	95.90 ± 0.34	96.13 ± 0.28
ResNet18-CBN	$\mathbf{96.93} \pm 0.61$	$\mathbf{97.10} \pm 0.23$

The results show that ResNet18-CBN significantly outperforms both ResNet18-Emb and the bioclimate-agnostic model. Unexpectedly, a learned embedding did not lead to significant improvements, which indicates that it matters how the bioclimatic context is incorporated into the neural network. Secondly, small improvements can be seen for all models if the encoder is initially trained on the self-supervised reconstruction objective. It is notable, that the randomly initialized ResNet18-CBN still outperforms the pretrained, bioclimate-agnostic model.

4.2 Generalizability

Besides their accuracy, we also quantify the models' generalizability as an additional assessment metric in Tab. 2. In fact, we consider two different notions of generalizability: in a classical machine learning sense and in a geographic sense.

In the classical machine learning sense, generalization describes the difference in accuracy between training and test data. This difference, which we denote by Gen-ML, should be as small as possible, as a large gap indicates that the model overfits to the distribution of the training data and is not sufficiently regularized.

In the geographic sense, we define generalizability as a model's ability to perform equally well across different biogeographical regions b. For quantification, we compute the accuracies over each regional subset of the test data and calculate their standard deviation with respect to the overall accuracy, as described in Sec. 4.1:

$$\text{Gen-Geo}(y, \widehat{y}) = \sqrt{\frac{1}{B} \sum_{b} (\text{Acc}(y_b, \widehat{y}_b) - \text{Acc}(y, \widehat{y}))^2}, \tag{5}$$

where B is the number of biogeographical regions considered. We acknowledge that by equally dividing the samples from all biogeographic regions in our train-test-split, we only consider the geographic generalizability within the domain of the training data. This is different from domain adaptation-related works which evaluate and improve models regarding their geographic generalizability outside the domain of the training data, e.g. different continents or cities [42].

Table 2: Averages and standard deviations of test generalizability metrics Gen-ML and Gen-Geo across 10 training runs on EuroSAT.

Model	Gen-ML in %		Gen-Geo in %	
	Supervised	Self-Supervised	Supervised	Self-Supervised
ResNet18	3.36 ± 0.55	2.64 ± 0.40	1.21 ± 0.34	$\mathbf{1.01} \pm 0.19$
ResNet18-Emb	3.32 ± 0.53	2.60 ± 0.60	1.21 ± 0.27	1.18 ± 0.18
ResNet18-CBN	$\mathbf{1.85} \pm 1.09$	$\mathbf{1.80} \pm 0.26$	$\mathbf{1.02} \pm 0.23$	$\mathbf{1.01} \pm 0.16$

Both metrics of generalizability are best for ResNet18-CBN, while ResNet18-Emb again only slightly outperforms the bioclimate-agnostic baseline. Improvements with self-supervised pretraining are most significant for the bioclimate-agnostic model, but also improve both metrics of the bioclimate-aware models. Based on the results on Gen-ML, we conclude that leveraging bioclimatic context through conditional batch normalization effectively regularizes neural network training for land cover classification. As for Gen-Geo, we can also see that ResNet18-CBN models generally achieve more geographically consistent classification results, which implies that bioclimatic data shift effects are successfully counteracted, especially in the supervised setting.

4.3 Training Efficiency

Here, we report the number of training epochs the model needs to pass certain accuracy thresholds p on the validation data, which we denote by Eff-p.

To account for variability regarding floating point operations, and thereby the time it takes to train a single epoch, we also need to determine relative walltime multiplicators α for each model. To this end, 10 epochs of training are run independently of training on the same physical GPU. The walltimes of each epoch are recorded and averaged. We define the supervised, bioclimate-agnostic ResNet18 as our baseline and set its $\alpha = 1$, accordingly. We track the validation accuracy throughout training with respect to the relative number of epochs, i.e. the number of epochs multiplied by α. The curves are averaged across the ten different runs.

Table 3: Efficiency characteristics and metrics Eff-0.95 of the implemented models as derived from the averaged training curves across 10 training runs on EuroSAT.

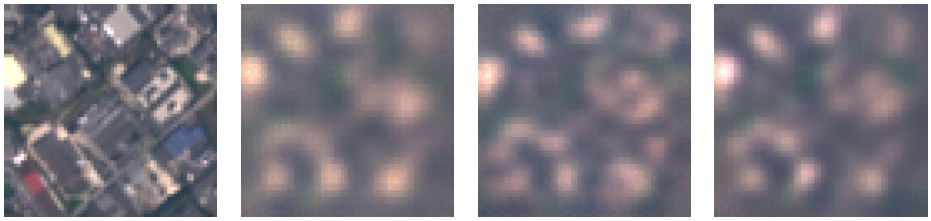
Model	α	#Param.s	Eff-0.95, Supervised	Eff-0.95, Self-Supervised
ResNet18	1.000	$12.556M$	13.208	12.000
ResNet18-Emb	1.080	$12.567M$	13.651	15.379
ResNet18-CBN	1.242	$12.749M$	**12.993**	**9.954**

The results largely confirm the findings from Secs. 4.1 and 4.2. Training times of ResNet18-CBN are reduced compared to the bioclimate-agnostic baseline in both the supervised setting and the self-supervised setting, although the difference is more significant with respect to the latter, where training is about 23% faster on average. Meanwhile, no improvements can be detected for ResNet18-Emb, where we even observe a surprising decline in efficiency when pre-training the model in a self-supervised manner.

4.4 Autoencoder Reconstructions and Representations

Below, we show an example image, as well as the corresponding reconstructions by the different autoencoders during the pre-training step in the self-supervised setting. The results for l_1 and l_{SSIM} on the test set are reported in Tab. 4.

Once again, ResNet18-CBN quantitatively outperforms the other two approaches in terms of reconstruction loss. The poor reconstruction quality is to be expected due to the very large drop in dimensionality from the original image to the latent representations: The images originally contain $64 \times 64 \times 13 = 53248$ values and are compressed into just 512. Despite this subjectively poor performance on the pretext task, the quantitative results from the previous sections suggest that the self-supervised pretraining benefits subsequent land cover classification and thus, that the encoder has learned useful land cover-relevant features.



(a) Original image (b) ResNet18 (c) ResNet18-Emb (d) ResNet18-CBN

Fig. 4: RGB visualization of the original EuroSAT image and autoencoder reconstructions of the different models.

To quantitatively evaluate the quality of the representations, we fit a simple Support Vector Machine (SVM) classifier with a Gaussian radial basis function kernel [5] to predict y from z with frozen encoder weights. The resulting overall accuracies are reported in Tab. 4. Again, ResNet18-CBN outperforms both the bioclimate-agnostic baseline and ResNet18-Emb. We also investigate the effect of applying CBN to only $\mathcal{E}$ and only $\mathcal{D}$. We notice that reconstructing quality suffers when only applying CBN to $\mathcal{E}$, whereas performance on the downstream task is similar. On the other hand, accuracy of the SVM declines significantly when applying CBN only to $\mathcal{D}$, while reconstruction quality even slightly improves with respect to both l_1 and l_{SSIM}.

Table 4: Averages and standard deviations of evaluation metrics of autoencoder reconstructions and representations across 10 training runs.

Model	l_1	l_{SSIM}	SVM-Acc in %
ResNet18	0.0110 ± 0.0005	0.9232 ± 0.0052	76.45 ± 5.34
ResNet18-Emb	0.0104 ± 0.0004	0.9300 ± 0.0035	79.52 ± 2.30
ResNet18-CBN	0.0095 ± 0.0002	0.9344 ± 0.0024	$\mathbf{83.73} \pm 1.01$
ResNet18-CBN ($\mathcal{E}$ only)	0.0114 ± 0.0006	0.9247 ± 0.0038	83.05 ± 2.35
ResNet18-CBN ($\mathcal{D}$ only)	$\mathbf{0.0091} \pm 0.0002$	$\mathbf{0.9384} \pm 0.0021$	75.25 ± 3.05

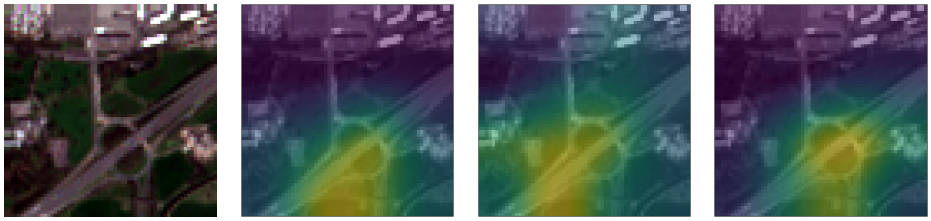
4.5 Sanity Checks for Conditional Batch Normalization

To confirm the integrity of ResNet18-CBN, we perform two basic sanity checks: First, we study if our observed improvements are actually due to the incorporation of bioclimatic context, and not due to differences between standard batch normalization and conditional batch normalization with respect to the overall optimization scheme as described in Sec. 3.3. To this end, we shuffle the auxiliary vectors a within each data split so that images are associated with 'fake' bioclimatic data which were originally derived for a different image. If the bioclimatic context in itself were irrelevant, this would have only little effect on the results.

However, we find that models perform worse under this manipulation with an average overall accuracy of 0.9496 across 10 runs, which is worse than ResNet18-CBN with unshuffled bioclimatic auxiliary vectors and the bioclimate-agnostic baseline, as reported in Sec. 4.1. This strengthens our hypothesis that bioclimatic context is useful auxiliary information in the context of land cover classification by ruling out the possibility that improvements are merely to changes in the optimization scheme.

Another recently raised concern about multi-modal learning using conditional batch normalization is that depending on the dataset, particularly the type and quality of the auxiliary data, shortcut learning may be encouraged [34]. Shortcut learning describes the phenomenon where models counterintuitively focus too strongly on the auxiliary information and thus do not learn meaningful features from the primary data source, i.e. the images.

We want to test if this effect is prevalent for our application and thus compute attribution maps, which indicate the relevant locations of an image with respect to specific class outputs. In particular, we apply Grad-CAM [33] to the last layer of the trained ResNet18 models from the supervised setting. Because the attribution maps are originally given in the respective layer's spatial dimension, they are upsampled to the original image size using bicubic interpolation.



(a) Original image (b) ResNet18 (c) ResNet18-Emb (d) ResNet18-CBN

Fig. 5: RGB visualization of the original image and overlayed Grad-CAM attribution maps from the different models for a data sample of the class *Highway*.

We visually compare the attribution maps from the models, which are once again averaged over the ten different runs, for a test data sample in Fig. 5. The classes *Highway* and *River* are particularly useful in this context, as they contain localizable objects in the image, which the Grad-CAM attributions should ideally highlight. We find that the attribution maps derived from both bioclimate-aware and -agnostic models are well aligned in this context, indicating that sensible visual features are learned despite the use of auxiliary information. There is thus no indication that the bioclimate-aware models suffer from the pitfall of shortcut learning. The most likely explanation is that bioclimatic context is not in itself informative enough to solve the prediction task, but still represents useful information to regularize neural network training.

4.6 Towards Large-Scale Application

While the simplicity of the EuroSAT dataset makes it suitable for the kinds of extensive experiments described above, it also causes the baseline ResNet18 to perform rather well, leaving only little room for improvement. We therefore provide additional accuracy metrics for supervised training on BigEarthNet, which is more challenging due to its about 200 times larger size and the presence of multiple labels for each image, as described in Sec. 3.1. In addition, unlike in EuroSAT, the label distribution is highly unbalanced, as shown in Fig. 2, which is why we report the micro- and macro-averaged F1-scores besides overall accuracy in Tab. 5.

Table 5: Test accuracy metrics on BigEarthNet.

Model	Acc in %	F1(micro) in %	F1(macro) in %
ResNet18	93.79	78.68	72.95
ResNet18-Emb	93.83	78.73	72.47
ResNet18-CBN	**94.30**	**79.57**	**75.68**

The results confirm our conclusions from the experiments on EuroSAT as ResNet18-CBN stands out as the best performer across all metrics and ResNet18-Emb does not offer significant improvements over the baseline. This is an indication that the approach is generally transferable to more data-intensive and complex scenarios. We are therefore optimistic, that it will also be a suitable building block within other types of neural networks, e.g. for semantic segmentation, which we plan to investigate in future work.

5 Conclusion

We showed that leveraging bioclimatic auxiliary data in a multi-modal setup benefits the training of neural networks for both supervised and self-supervised land cover classification regarding all considered quantitative and qualitative aspects. The incorporation by means of conditional batch normalization lead to particularly good results, whereas improvements were surprisingly marginal when using added embeddings. Because of the universal applicability to a wide range of other datasets and architectures, these insights can have a great impact on the growing interdisciplinary field of machine learning for Earth observation.

References

1. Alemohammad, H., Booth, K.: LandCoverNet: A global benchmark land cover classification training dataset. In: AI for Earth Sciences Workshop at NeurIPS (2020)

2. Audebert, N., Le Saux, B., Lefèvre, S.: Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing **140**, 20–32 (2018)
3. Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S.: Geography-aware self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10181–10190 (2021)
4. Cao, Y., Steffey, S., He, J., Xiao, D., Tao, C., Chen, P., Müller, H.: Medical image retrieval: a multimodal approach. Cancer Informatics **13**, CIN–S14053 (2014)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**, 273–297 (1995)
6. De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., Courville, A.C.: Modulating early visual processing by language. Advances in Neural Information Processing Systems **30** (2017)
7. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R.: DeepGlobe 2018: A challenge to parse the earth through satellite images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 172–181 (2018)
8. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. In: International Conference on Learning Representations (2016)
9. Eddin, M.H.S., Roscher, R., Gall, J.: Location-aware adaptive normalization: A deep learning approach for wildfire danger forecasting. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–18 (2023)
10. EEA: Biogeographical regions Europe 2016 (2016), https://www.eea.europa.eu/ds_resolveuid/9b7911cc33ad4a9c940847a7ff653a40
11. Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., et al.: The shuttle radar topography mission. Reviews of Geophysics **45**(2) (2007)
12. Fick, S.E., Hijmans, R.J.: Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology **37**(12), 4302–4315 (2017), https://worldclim.org/data/worldclim21.html
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. Heidler, K., Mou, L., Hu, D., Jin, P., Li, G., Gan, C., Wen, J.R., Zhu, X.X.: Self-supervised audiovisual representation learning for remote sensing data. International Journal of Applied Earth Observation and Geoinformation **116**, 103130 (2023)
15. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)
16. Hu, L., Robinson, C., Dilkina, B.: Model generalization in deep learning applications for land cover mapping. arXiv preprint arXiv:2008.10351 (2020)
17. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1501–1510 (2017)
18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456 (2015)

19. Jain, A., Singh, A., Koppula, H.S., Soh, S., Saxena, A.: Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In: IEEE International Conference on Robotics and Automation. pp. 3118–3125 (2016)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference for Learning Representations (2014)
21. Kramer, M.A.: Nonlinear principal component analysis using autoassociative neural networks. AIChE journal **37**(2), 233–243 (1991)
22. Lenz, I., Lee, H., Saxena, A.: Deep learning for detecting robotic grasps. The International Journal of Robotics Research **34**(4-5), 705–724 (2015)
23. Li, W., Chen, K., Chen, H., Shi, Z.: Geographical knowledge-driven representation learning for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–16 (2021)
24. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive batch normalization for practical domain adaptation. Pattern Recognition **80**, 109–117 (2018)
25. Liang, M., Li, Z., Chen, T., Zeng, J.: Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics **12**(4), 928–937 (2014)
26. Lu, X., Gong, T., Zheng, X.: Multisource compensation network for remote sensing cross-domain scene classification. IEEE Transactions on Geoscience and Remote Sensing **58**(4), 2504–2515 (2019)
27. Marín, J., Escalera, S.: SSSGAN: Satellite style and structure generative adversarial networks. Remote Sensing **13**(19), 3984 (2021)
28. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning. pp. 689–696 (2011)
29. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2337–2346 (2019)
30. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
31. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine **34**(6), 96–108 (2017)
32. Scheibenreif, L., Hanna, J., Mommert, M., Borth, D.: Self-supervised vision transformers for land-cover segmentation and classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1422–1431 (2022)
33. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
34. Sheth, I., Rahman, A.A., Havaei, M., Kahou, S.E.: Pitfalls of conditional batch normalization for contextual multi-modal learning. In: I Can't Believe It's Not Better Workshop at NeurIPS (2022)
35. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. Advances in Neural Information Processing Systems **25** (2012)
36. Suel, E., Bhatt, S., Brauer, M., Flaxman, S., Ezzati, M.: Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. Remote Sensing of Environment **257**, 112339 (2021)

37. Sumbul, G., Charfuelan, M., Demir, B., Markl, V.: BigEarthNet: A large-scale benchmark archive for remote sensing image understanding. In: IEEE International Geoscience and Remote Sensing Symposium. pp. 5901–5904. IEEE (2019)
38. Sumbul, G., De Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., Markl, V.: BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. IEEE Geoscience and Remote Sensing Magazine **9**(3), 174–180 (2021)
39. Sun, Y., Hua, Y., Mou, L., Zhu, X.X.: CG-Net: Conditional GIS-aware network for individual building segmentation in VHR SAR images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2021)
40. Tang, K., Paluri, M., Fei-Fei, L., Fergus, R., Bourdev, L.: Improving image classification with location context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1008–1016 (2015)
41. Townshend, J.G.: Land cover. International Journal of Remote Sensing **13**(6-7), 1319–1328 (1992)
42. Tuia, D., Persello, C., Bruzzone, L.: Domain adaptation for the classification of remote sensing data: An overview of recent advances. IEEE Geoscience and Remote Sensing Magazine **4**(2), 41–57 (2016)
43. Turner, M.G., Gardner, R.H.: Landscape ecology in theory and practice. Springer (2015)
44. Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In: NeurIPS Datasets and Benchmarks Track (2021)
45. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 606–615 (2018)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
47. Wickramasinghe, C.S., Marino, D.L., Manic, M.: ResNet autoencoders for unsupervised feature learning from high-dimensional data: Deep models resistant to performance degradation. IEEE Access **9**, 40511–40520 (2021)
48. Wulder, M.A., Coops, N.C., Roy, D.P., White, J.C., Hermosilla, T.: Land cover 2.0. International Journal of Remote Sensing **39**(12), 4254–4284 (2018)
49. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging **3**(1), 47–57 (2016)
50. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: SEAN: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5104–5113 (2020)