ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240739

Context Matters: Leveraging Spatiotemporal Metadata for Semi-Supervised Learning on Remote Sensing Images

Maximilian Bernhard[a,*], **Tanveer Hannan**[a], **Niklas Strauß**[a] **and Matthias Schubert**[a]

[a]LMU Munich, MCML

Abstract. Remote sensing projects typically generate large amounts of imagery that can be used to train powerful deep neural networks. However, the amount of labeled images is often small, as remote sensing applications generally require expert labelers. Thus, semi-supervised learning (SSL), i.e., learning with a small pool of labeled and a larger pool of unlabeled data, is particularly useful in this domain. Current SSL approaches generate pseudo-labels from model predictions for unlabeled samples. As the quality of these pseudo-labels is crucial for performance, utilizing additional information to improve pseudo-label quality yields a promising direction. For remote sensing images, geolocation and recording time are generally available and provide a valuable source of information as semantic concepts, such as land cover, are highly dependent on spatiotemporal context, e.g., due to seasonal effects and vegetation zones. In this paper, we propose to exploit spatiotemporal metainformation in SSL to improve the quality of pseudo-labels and, therefore, the final model performance. We show that directly adding the available metadata to the input of the predictor at test time degenerates the prediction quality for metadata outside the spatiotemporal distribution of the training set. Thus, we propose a teacher-student SSL framework where only the teacher network uses metainformation to improve the quality of pseudo-labels on the training set. Correspondingly, our student network benefits from the improved pseudo-labels but does not receive metadata as input, making it invariant to spatiotemporal shifts at test time. Furthermore, we propose methods for encoding and injecting spatiotemporal information into the model and introduce a novel distillation mechanism to enhance the knowledge transfer between teacher and student. Our framework dubbed *Spatiotemporal SSL* can be easily combined with several state-of-the-art SSL methods, resulting in significant and consistent improvements on the BigEarthNet and EuroSAT benchmarks. Code is available at https://github.com/mxbh/spatiotemporal-ssl.

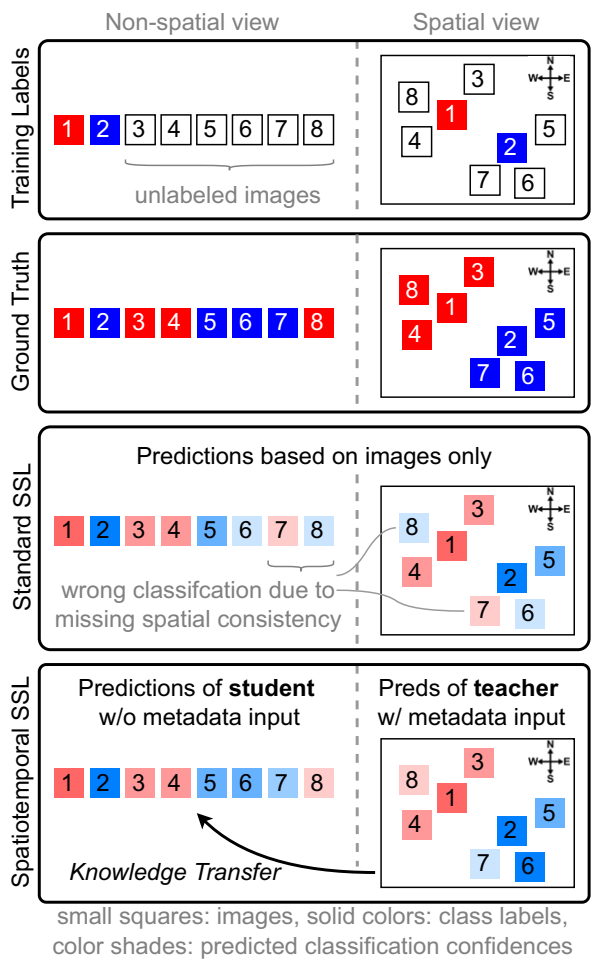


Figure 1: **Intuition for our proposed Spatiotemporal SSL framework.** Additional spatiotemporal input data facilitates learning for the teacher, leading to better predictions and pseudo-labels (7 and 8). The student model in Spatiotemporal SSL learns from these improved pseudo-labels without relying on the additional input, thereby also achieving a better performance. *Temporal dimension omitted for simplicity. Best viewed digitally.*

1 Introduction

Applying deep learning models to analyze and interpret remote sensing imagery is a powerful tool that has been successfully applied in various use cases [12, 41, 4, 21]. However, deep neural networks are known to be data and label-hungry, i.e., they usually require large amounts of labeled samples to reach the desired performance [11]. This requirement poses a problem in many real-world applications as datasets with large quantities of high-quality labels matching the specific use case at hand are often unavailable, and labeling data is costly. In this regard, semi-supervised learning (SSL) [32] constitutes a promising remedy as it aims to narrow the performance gap to fully supervised learning when only a small part of the training dataset is labeled. While SSL is an active field of research with a multitude of methods and approaches [27, 15, 3, 35, 26, 37, 16], its primary focus lies on optimizing the usage of unlabeled samples. At the same time,

* Corresponding Author. Email: bernhard@dbs.ifi.lmu.de

the potential of exploiting additional metadata is mostly neglected.

Remote sensing images are accompanied by metadata such as geolocation and acquisition time as they describe the Earth's surface at a certain place and time. This spatiotemporal metadata can be used during training without additional labeling effort and improve the performance of models when used aptly [29, 20, 18, 6, 39]. Location and recording time yield valuable information as many semantic concepts in remote sensing, such as land cover, are spatiotemporally coherent, and visual features are often highly dependent on the spatiotemporal context. For example, vegetation drastically varies with different climate zones, countries, and seasons. Including this type of information is especially useful for SSL, where labeled data is scarce, as the additional features can alleviate the learning problem and, thus, ultimately reduce the label requirement [11]. However, current SSL methods do not include this valuable source of information.

In this paper, we propose an approach for leveraging this additional metainformation by learning the joint distribution of labels, visual features, and spatiotemporal context. However, models using spatiotemporal metadata as an input are prone to overfitting to the metadata of the training set due to the small number of labeled samples in SSL and spatiotemporal sampling biases introduced during dataset creation. As a result, models relying on spatiotemporal metadata as an input poorly generalize at test time for out-of-distribution metadata (see experiments in Section 4.5). Especially in remote sensing, datasets often cover only a limited area and a few points in time. Therefore, applying a model in spatiotemporal contexts beyond the training data, e.g., a different geographical region or season, is highly desirable in practice.

To leverage spatiotemporal metainformation in SSL without compromising generalization at test time, we propose a novel student-teacher framework, which we call *Spatiotemporal SSL* (see Figure 1). The teacher model receives the metadata as additional input and is trained in a semi-supervised way to generate high-quality pseudo-labels. These pseudo-labels are then used to train the student model, which does not rely on the metadata input but solely operates on images. This design has the following advantages: First, the teacher model benefiting from the spatiotemporal information is only employed on the training set, avoiding the generalization issues at inference mentioned above. Second, the student model indirectly benefits from the spatiotemporal information as it receives strong pseudo-labels generated with the help of the additional input. Third, since the student model does not receive the metadata input, it is invariant to shifts in the spatiotemporal distribution of test samples, allowing it to generalize to unseen spatiotemporal contexts (see experiments in Section 4.5).

To jointly model visual features and the spatiotemporal distribution in the teacher, we modify the vision transformer (ViT) [8] by supplementing the visual patch tokens with a specialized metatoken encoding the spatiotemporal information. Furthermore, we optimize the knowledge transfer between teacher and student beyond passing on pseudo-labels by integrating a novel distillation mechanism. Here, a dedicated distillation token in the student model is supervised to align with the metatoken embedding of the teacher model, allowing the student model to access the spatiotemporal reasoning of the teacher without actually receiving spatiotemporal inputs.

Notably, our Spatiotemporal SSL framework does not rely on restrictive assumptions about the underlying SSL algorithm, making it versatile and compatible with recent developments in SSL. In our experiments on the popular BigEarthNet and EuroSAT benchmarks, we demonstrate that combining Spatiotemporal SSL with several state-of-the-art SSL methods such as FixMatch [27] and DeFixMatch [26] leads to substantial and consistent improvements. We also perform detailed experiments and ablation studies to identify and analyze relevant factors in our approach.

2 Related Work

Semi-Supervised Learning Semi-supervised learning (SSL) is an active field of research with a large variety of methods [32, 40, 22]. A major cornerstone in this field is FixMatch [27]. FixMatch represents a framework where hard pseudo-labels are generated from confident predictions on weakly augmented, unlabeled samples in order to supervise the predictions for strongly augmented versions of these samples. Many works extend FixMatch, e.g., by applying specialized consistency losses [10, 17, 43, 44, 26] or replacing the hard, threshold-based pseudo-labeling with more sophisticated techniques, oftentimes aiming to align the distribution of pseudo-labels with the observed distribution of labels [42, 38, 2, 35, 3]. More specifically, FreeMatch [35] adapts FixMatch's fixed confidence threshold for pseudo-labeling for each class separately and combats pseudo-label class imbalance with an additional loss. SoftMatch [3] applies a soft pseudo-label weighting and, similarly to FreeMatch, aligns the pseudo-labels with a uniform distribution. DeFixMatch [26] follows a rather different approach, which is debiasing the learner by applying the negative unsupervised loss on the labeled data.

Many works in SSL focus on problems that are often met when working with remote sensing data. For example, [36, 23, 15] consider SSL with imbalanced and long-tailed data. In particular, UDAL [15] presents a way to combat label imbalance by integrating the distribution alignment into the cross-entropy computation via modulating predicted scores. CAP [37] addresses the problem of semi-supervised multi-label classification (instead of multi-class, single-label classification as most SSL benchmarks). Their approach is centered around finding suitable pseudo-labeling thresholds for the individual classes.

Our contribution can be considered orthogonal to previous methods since our framework can be combined with several state-of-the-art SSL algorithms as shown in our experiments. Hence, even future SSL techniques might benefit from our approach.

Image Classification with Additional Metadata Several works explored feeding image metadata as additional input to the model in fully supervised settings. In [29], it has been shown that geolocation as additional network input can improve supervised image classification on YFCC100M [30]. Similarly, [20] uses metadata, including geolocation, as additional input in supervised classification on FMOW [5]. However, their main focus lies on network ensembling technique instead of the inclusion of additional input data. In [25], metadata are used via context networks to learn a dynamic map of visual appearance, which has been shown to be beneficial for image localization, image retrieval, and metadata verification. [18] exploits metadata to learn a spatiotemporal prior in order to refine predictions in supervised fine-grained classification on YFCC100M [30], BirdSnap [1], and iNaturalist [33]. Especially for fine-grained image classification, the usage of spatiotemporal metadata has proven useful as the subsequent works of [6] and [39] demonstrate. Specifically, [6] investigates different ways to include metadata, i.e., via geographical priors, prediction postprocessing, or feature modulation, whereas [39] proposes DynamicMLP, a novel building block for fusing multimodal features effectively.

Though existing literature shows the value of spatiotemporal metadata for image classification, this line of research only examines fully supervised settings. Furthermore, these works do not address spatiotemporal generalization at test time. In contrast, we examine spa-

tiotemporal metadata in SSL where directly using the metadata as input might lead to generalization issues.

3 Spatiotemporal SSL

In the following, we introduce the necessary notation before we describe the training and architecture of the spatiotemporal teacher model, the student's training, and the novel distillation scheme. Figure 2 provides an overview of our approach.

3.1 Notation

Let I and y denote images and their ground-truth labels. In SSL, we further distinguish between labeled and unlabeled images I_L and I_U. We denote the number of labeled images and unlabeled images in a batch as n_L and n_U, respectively. During training, we only have access to the ground truth of the labeled images y_L. We consider the geolocation G and time T of image acquisition as additional metadata, consolidated to $M = (G, T)$ (M_L and M_U for labeled and unlabeled samples, respectively). Our ultimate goal is to estimate the label y for an image I with a neural network f_θ (student) as $f_\theta(I)$. The metadata M will be consumed by another neural network f_ϑ (teacher), providing the pseudo-label $f_\vartheta(I, M)$. We assume f_θ and f_ϑ to be two separate neural networks as their input differs.

3.2 Training the Spatiotemporal Teacher Model

To train the spatiotemporal teacher model, we assume a generic SSL framework. In the following, we adopt the basic structure of FixMatch [27], which is the basis for most recent SSL methods. Here, the teacher model f_ϑ is trained with separate losses for labeled and unlabeled data. The loss for the labeled data $\mathcal{L}_L$ can be any supervised loss function, i.e., we write

$$\mathcal{L}_L^\vartheta = \frac{1}{n_L}\sum_{i=0}^{n_L} \ell_L(y_i, f_\vartheta(I_i, M_i)) \tag{1}$$

where ℓ_L denotes the loss function for a single sample and is chosen according to the underlying SSL algorithm (e.g., cross-entropy in FixMatch). On the other hand, the unsupervised loss $\mathcal{L}_U^\vartheta$ is applied to ensure consistency on unlabeled data. This loss employs pseudo-labels generated from weakly augmented image versions $I_U^{(w)}$ to supervise the predictions for strongly augmented image versions $I_U^{(s)}$. To mitigate the effect of noisy and unreliable pseudo-labels, a weighting function α determines the loss contribution of each individual sample (e.g., confidence-based hard thresholding as in FixMatch [27]). Formally, the unlabeled loss $\mathcal{L}_U$ is defined as

$$\mathcal{L}_U^\vartheta = \frac{1}{n_U}\sum_{i=1}^{n_U} \alpha\Big(f_\vartheta(I_i^{(w)}, M_i)\Big) \cdot \ell_U\Big(f_\vartheta(I_i^{(w)}, M_i), f_\vartheta(I_i^{(s)}, M_i)\Big). \tag{2}$$

For example, we can use cross-entropy as criterion ℓ_U after generating hard pseudo-labels from $f_\vartheta(I_i^{(w)}, M_i)$. Let us note that more sophisticated techniques, such as modulating prediction scores with distribution alignment [35, 3], can also be used. The overall loss for the teacher model f_ϑ is

$$\mathcal{L}^\vartheta = \mathcal{L}_L^\vartheta + \lambda_U \cdot \mathcal{L}_U^\vartheta, \tag{3}$$

where λ_U is a parameter for balancing the two described loss terms.

3.3 Spatiotemporal Teacher Architecture

A straightforward way to exploit spatiotemporal metadata would be to estimate class priors from the spatiotemporal occurrences and use these to post-process the predicted classification scores, e.g., with a Bayesian approach [6, 18]. This approach assumes that the visual appearance of a class is generally the same for different geolocations and times. More formally, the assumption is that the visual appearance $I|y$ given the class is conditionally independent of the spatiotemporal context M [18]. However, in our setting, we argue that the spatiotemporal context has a considerable effect on the visual appearance as, for instance, the same type of land cover may look very different across varying countries, climate zones, and seasons.

Therefore, it is sensible to process and modulate image features conditional to the spatiotemporal information as it allows the model to capture such variation in the visual appearance. To put this into practice, we opt for an early-fusion architecture where spatiotemporal and image information are modeled jointly. In other words, with an early interaction of visual and spatiotemporal information, the model can not only learn which classes are likely for a certain location and time but also how they visually depend on the location and time.

On a technical level, we concatenate latitude, longitude, and time, represented as the relative day of the year, and feed the resulting vector into a two-layer MLP to generate a single *metatoken*. Afterward, we pass the metatoken along with the visual tokens encoding image patches on to the Vision Transformer (ViT) [8] architecture (see the teacher in Figure 2 (b)). This design allows maximal interaction between the visual and spatiotemporal information while introducing minimal methodological overhead. In addition, the metatoken can be easily injected into other transformer-based vision backbones such as [9, 7, 14, 24]. An empirical analysis of the proposed early-fusion approach can be found in Section 4.4.

3.4 Training the Student Model

Building upon the training of the teacher model f_ϑ, we now describe the training of the student f_θ. Our goal is to transfer the teacher's learned knowledge to the student not receiving metadata as input. To this end, we train f_θ with its own loss for labeled data $\mathcal{L}_L^\theta$, which can be obtained from Equation 1 by substituting $f_\theta(I_i)$ for $f_\vartheta(I_i, M_i)$. However, we want the student to benefit from the spatiotemporal teacher by providing it with the teacher's pseudo-labels for unlabeled data. More precisely, we modify $\mathcal{L}_U$ from Equation 2 as follows:

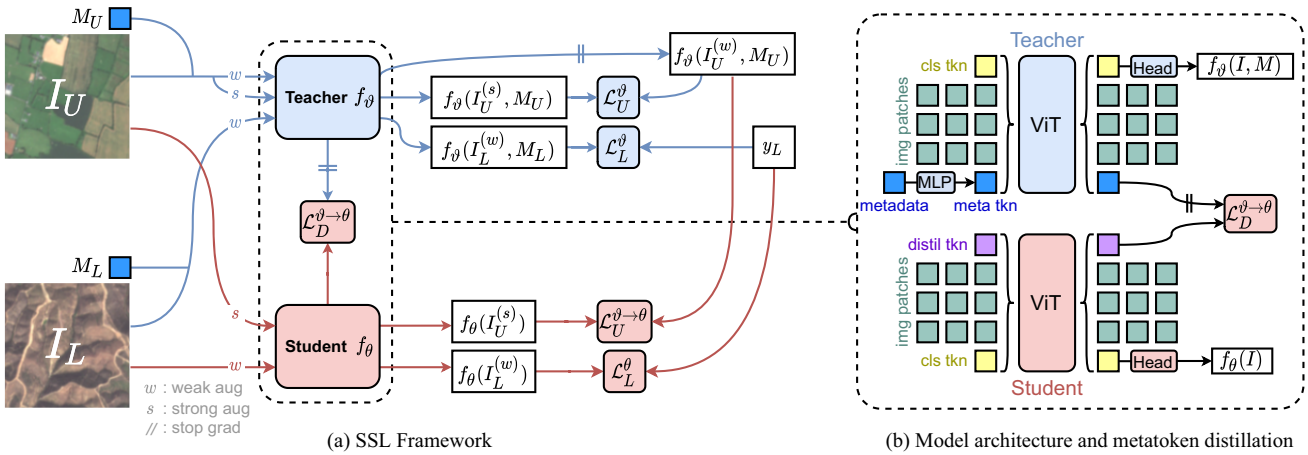
$$\mathcal{L}_U^{\vartheta\to\theta} = \frac{1}{n_U}\sum_{i=1}^{n_U} \alpha\Big(f_\vartheta(I_i^{(w)}, M_i)\Big) \cdot \ell_U\Big(f_\vartheta(I_i^{(w)}, M_i), f_\theta(I_i^{(s)})\Big). \tag{4}$$

That is, instead of bootstrapping the pseudo-labels for f_θ from f_θ's predictions itself, we employ the pseudo-labels generated from the teacher f_ϑ. As f_ϑ has access to the additional input, these pseudo-labels are of higher quality and, therefore, improve the student's training. Putting everything together, we obtain the training objective for f_θ as

$$\mathcal{L}^\theta = \mathcal{L}_L^\theta + \lambda_U \cdot \mathcal{L}_U^{\vartheta\to\theta}. \tag{5}$$

The teacher and the student are trained simultaneously on the same images, resulting in convenient single-stage training.

Let us note that Spatiotemporal SSL does not make restrictive assumptions about the underlying SSL algorithm, i.e., α, ℓ_L, ℓ_U as well



(a) SSL Framework (b) Model architecture and metatoken distillation

Figure 2: **Overview of the proposed Spatiotemporal Semi-supervised Learning.** The spatiotemporal teacher f_ϑ utilizes both images and metadata and generates strong pseudo-labels that are used as supervision for the student f_θ, which does not access the metadata.

as additional loss functions such as the debiasing loss of [26] can be freely chosen. Hence, our approach is versatile and orthogonal to recent developments in SSL, allowing to combine it with several state-of-the-art SSL algorithms (see Section 4.3).

3.5 Metatoken Distillation

To enhance the knowledge transfer from the teacher to the student beyond the exchange of pseudo-labels, we integrate a specialized knowledge distillation mechanism into our framework (see Figure 2 (b)). Inspired by [31], we introduce a dedicated distillation token to the student network, which interacts with all the other tokens through self-attention. In contrast to [31], we do not use this token for distilling knowledge from another model with a fundamentally different architecture but from a model with different inputs. That is, the output embedding of the distillation token is supervised to be similar to the output embedding of the teacher's spatiotemporally informed metatoken. Our proposed metatoken distillation loss $\mathcal{L}_D^{\vartheta\to\theta}$ is defined as the mean squared error of the teacher's metatoken embeddings and the student's distillation token embeddings over all samples (labeled and unlabeled). We use this loss to update only the student model f_θ but not the teacher f_ϑ to prevent the teacher from adapting to the student. Since the student does not access the additional metainformation, it cannot provide meaningful guidance for the teacher (see Section 4.4 for an empirical justification). Hence, the overall loss for the student f_θ is defined as

$$\mathcal{L}^\theta = \mathcal{L}_L^\theta + \lambda_U \cdot \mathcal{L}_U^{\vartheta\to\theta} + \lambda_D \cdot \mathcal{L}_D^{\vartheta\to\theta}, \tag{6}$$

where λ_D is a weighting hyperparameter for the distillation loss. A detailed analysis of design choices and ablation studies for this additional loss can be found in Section 4.4.

4 Experiments

4.1 Datasets

BigEarthNet [28] is a large-scale land cover dataset. It contains 590,326 Sentinel-2 images of 120×120 pixels and 43 classes taken from the CORINE Land Cover database (CLC18). BigEarthNet is an imbalanced multi-label classification dataset, where the relative frequencies of labels ranges from 37% to 0.05%. We adopt the train-val-test split of [19] and further split the training set into labeled and unlabeled samples in a stratified way. Apart from the RGB-bands of the images, we consider the geolocation and the image acquisition time, represented as the relative day of the year, as the input to the network. The BigEarthnet images were acquired between June 2017 and May 2018 in ten European countries (Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, Switzerland).

EuroSAT [13] is another, highly popular land use and land cover dataset containing 27,000 Sentinel-2 patches of 64×64 pixels. Each image patch belongs to one out of 10 classes. We adopt the dataset split provided by the Unified Semi-supervised Learning Benchmark (USB) [34], i.e., we use a fixed train-test split and the training images are further divided into labeled and unlabeled images such that exactly the same number of labels are available for every class. Similar to BigEarthNet, we use the RGB-bands of the images and the geolocation as metadata but omit the image acquisition time, which is not available for EuroSAT. The image locations of EuroSAT are scattered over 34 European countries (see supplementary material[1]).

4.2 Implementation Details

We implement our method within the USB codebase [34]. We use ViT-S [8] as our base architecture. On BigEarthNet, we resize images to 128×128, use a ViT patch size of 16 pixels, and train with a cosine learning rate schedule for 64k steps with an initial learning rate of 1e-4 and a batch size of 512 (64 labeled, 448 unlabeled). On EuroSAT, we adopt the configuration of USB [34], i.e., we resize images to 32×32 pixels, use a ViT patch size of 2 pixels, and train with a cosine learning rate schedule for 204,800 steps with an initial learning rate of 5e-5 and a batch size of 16 (8 labeled, 8 unlabeled). As BigEarthNet is a multi-label classification dataset, we extend SSL methods that were originally proposed for single-label problems (FixMatch [27], SoftMatch [3], FreeMatch [35], DeFixMatch [26], UDAL [15]) by applying the pseudo-labeling scheme to every class separately. For the weight parameter λ_U, we adopt the proposed values of the respective underlying SSL algorithm. For λ_D, we choose 1.0 on BigEarthNet and 0.01 on EuroSAT (see Section 4.4). Following the common practice [27], we use the exponential moving average of models for evaluation. For more details, see supplementary material.

[1] supplementary material can be found in our code repository

Table 1: **Results on BigEarthNet with 1% training labels** Our ST-SSL leads to an improvement on every standard SSL method, even surpassing the best standard algorithm (DeFixMatch) with the worst ST-SSL combination (CDMAD + ST-SSL).

Method	*(Venue)*	mAP	*(±Δ)*
Supervised only baseline		40.47	–
FixMatch [27]	*(NeurIPS'20)*	42.56	–
UDAL [15]	*(WACV'23)*	42.32	–
SoftMatch [3]	*(ICLR'23)*	41.95	–
FreeMatch [35]	*(ICLR'23)*	42.46	–
DeFixMatch [26]	*(ICLR'23)*	43.09	–
CAP [37]	*(NeurIPS'23)*	41.81	–
CDMAD [16]	*(CVPR'2024)*	41.48	–
FixMatch + ST-SSL		46.12	*(+3.56)*
UDAL + ST-SSL		45.84	*(+3.52)*
Softmatch + ST-SSL		45.34	*(+3.39)*
FreeMatch + ST-SSL		45.20	*(+2.74)*
DeFixMatch + ST-SSL		**46.65**	*(+3.56)*
CAP + ST-SSL		45.53	*(+3.72)*
CDMAD + ST-SSL		43.74	*(+2.26)*

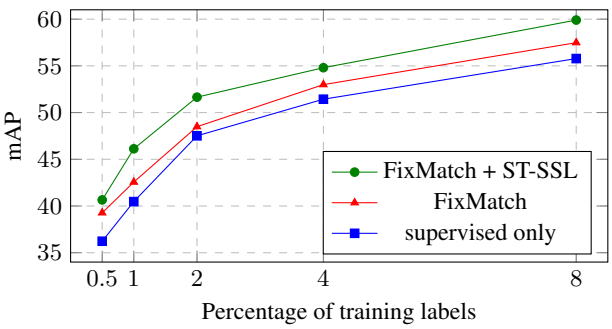


Figure 3: **Results for different numbers of labeled training samples on BigEarthNet.** Our method consistently outperforms FixMatch and the supervised baseline.

4.3 *Combination and Comparison with State-of-the-Art SSL Methods on BigEarthNet*

In Table 1, we provide results for several SSL methods on BigEarthNet when 1% of the training labels are available. The existing SSL methods reach mAP scores in the range of 41.48% (CDMAD) to 43.09% (DeFixMatch), thereby surpassing the "supervised only baseline" (40.47%), which is trained in a fully supervised way on only the labeled samples. When combining these SSL methods with the Spatiotemporal SSL framework, we can observe consistent performance improvements leading to mAP scores in the range of 43.74% (CDMAD + ST-SSL) and 46.65% (DeFiXMatch + ST-SSL). That is, even the least performing ST-SSL variant outperforms the best SSL method without ST-SSL. At the same time, the top-performing conventional SSL method DeFixMatch improves by 3.56% mAP with ST-SSL. This observation is even more remarkable since we did not tune hyperparameters for the single ST-SSL combinations, but took the conventional SSL configuration and transferred it to ST-SSL without modification.

Furthermore, we investigate the generalization of our approach to other settings with different proportions of labeled data by comparing FixMatch and FixMatch + ST-SSL in Figure 3. Once again, ST-SSL consistently improves the baseline across all settings.

4.4 *Ablations*

To identify contributing factors within our method, we perform a variety of ablation studies presented in Table 2. We base the ablations on FixMatch [27] as it is a strong and highly popular SSL method that is conceptually simple at the same time.

Table 2: **Ablations and detailed experiments** on BigEarthNet with 1% training labels.

	Method	mAP	*(±Δ)*
(a)	FixMatch + ST-SSL (MSE distil, $\lambda_D = 1$)	46.12	–
(b)	without acquisition time T	45.59	*(-0.53)*
(c)	without geolocation G	44.96	*(-1.16)*
(d)	late fusion	43.59	*(-2.53)*
(e)	single model	44.61	*(-1.51)*
(f)	without distillation	44.92	*(-1.20)*
(g)	MAE distillation	45.86	*(-0.36)*
(h)	cosine sim. distillation	45.17	*(-0.95)*
(i)	classification token distillation	42.89	*(-3.23)*
(j)	without stopping gradients	45.86	*(-0.26)*
(k)	$\lambda_D = 0.1$	45.27	*(-0.85)*
(l)	$\lambda_D = 0.5$	45.76	*(-0.36)*
(m)	$\lambda_D = 2.0$	45.78	*(-0.34)*

Table 2 is divided into five sections: The first section (a) represents our full method as already seen in Table 1. Next, experiments (b,c), analyze the effect of geolocation G and acquisition time T. If we omit the acquisition time T and only consider the geolocation G as additional input (b), we observe a drop in mAP of about 1%. Conversely, if we only use the image acquisition time and omit the geolocation (c), we observe a performance drop of about 2%. That is, both the geolocation and acquisition time provide useful information for the model. However, the geolocation seems to have a stronger effect which is in line with the reasoning that land cover is geographically coherent. On the other hand, the time information is helpful as it allows the network to model differences in visual appearance of land cover throughout seasons and potentially also because of sampling bias in the training data. Note that exploiting sampling bias in the training data is not problematic in our framework as the final model has no direct access to the metadata, preventing it from transferring the bias to test time.

The third group (d,e) in Table 2 is concerned with the modeling of spatiotemporal information within the teacher network. First, we compare our early fusion via the metatoken with a late-fusion approach (d), where we add the encoded metadata to the encoded image representation before passing it on to the classification head. The resulting performance degradation of about 2.5% mAP indicates that the early interaction of image features and metadata information is, in fact, beneficial as it allows the modulation of visual features based on the spatiotemporal context. In the late-fusion architecture, this kind of interaction is not possible, restricting the model to primarily learn a spatiotemporal prior for the occurrence of classes. Furthermore, we investigate the case where the teacher and the student are represented by the same network (e). Here, we observe a decrease in performance of 1.51%, indicating that specialization with separate models for spatiotemporal and purely visual reasoning is beneficial.

In the next section (f-j), we investigate several design choices regarding our metatoken distillation. First of all, not using the distillation mechanism at all (f), leads to an inferior performance of 44.92% mAP (vs. 46.12%). Replacing the mean squared error (MSE) in the distillation loss $\mathcal{L}_D^{\vartheta \to \theta}$ with the mean absolute error (MAE) (g) and the cosine similarity (h) results in mAP drops of 0.36% and 0.95%, respectively. Furthermore, using the teacher's classification token instead of the metatoken for distillation (i) leads to a drop of 3.23% mAP. Moreover, omitting the gradient stopping and letting the gradients of the distillation loss flow into the teacher network (j) decreases

Table 3: **Generalization to out-of-distribution metadata** on BigEarthNet with 1% training labels. †: Different dataset split based on geography, i.e., (e,f) are *not* directly comparable to (a-d).

	Model	Metadata	OOD component	mAP
(a)	Student	✗	–	46.12
(b)	Teacher	✓	–	44.93
(c)	Teacher	✓	$p_{train}(G) \neq p_{test}(G)$	30.15 ± 1.38
(d)	Teacher	✓	$p_{train}(T) \neq p_{test}(T)$	42.75 ± 1.30
(e)	Student†	✗	$p_{train}(y, I) \neq p_{test}(y, I)$	16.65
(f)	Teacher†	✓	$p_{train}(y, I, G) \neq p_{test}(y, I, G)$	15.43

Table 4: **Results (accuracy) on EuroSAT.** With few labels, FixMatch + ST-SSL clearly outperforms the baselines, while we see diminishing value of the metadata and ST-SSL for larger numbers of labels.

Method	Number of labels			
	10	20	40	80
Supervised only	57.56	71.70	85.13	88.50
FixMatch	64.19	89.13	**94.20**	96.39
FixMatch + ST-SSL	**77.85**	**90.56**	93.96	**96.52**

the performance by a slight margin of 0.26%.

Finally, we study the influence of the weight of the distillation loss λ_D in the last section of the table (k-m). The overall performance is relatively robust to the choice of λ_D as setting λ_D suboptimally still gives us better mAP scores than not using the metatoken distillation at all (f). For an analogous experiment on EuroSAT, please refer to the supplementary material.

4.5 Generalization to Out-of-Distribution Metadata

A central aspect of this work is the assumption that it is not advisable to use metadata inputs at test time as it would have a detrimental effect on the generalization of the model. To investigate this, we analyze different scenarios in Table 3. For this experiment, we train FixMatch + ST-SSL on BigEarthNet with 1% of training labels and evaluate the performances of the student (not relying on metadata inputs) and the teacher (relying on metadata inputs).

The first row (a) of the table is our final student model as presented in Table 1. The second model (b) corresponds to the teacher model, which relies on the metadata as input at test time. With this model, we observe a slight tendency to overfit on the training samples as the test mAP is 1.19% lower than the student's. In contrast, the quality of predictions, i.e., pseudo-labels, on the training set is consistently higher than for the non-spatiotemporal counterpart, while the pseudo-label quantity is virtually identical (see Figure 6).

In experiments (c) and (d), we simulated test samples where only the test locations (c) or only the test image acquisition times (d) are outside the training distribution. To this end, we select five locations and acquisition times outside the training distribution and replace the true test locations or times with these fixed values (see supplementary material for details). We report the mean and standard deviation of mAP for the five selected values. When the test locations are outside the training support (c), we observe a huge performance drop of about 15% for the teacher model. For the acquisition times (d), we observe a significant drop of about 2%. However, the drop is substantially smaller as we encode the acquisition times using the relative day of the year, leaving little space for far out-of-distribution values. Note that the proposed model (a) is completely unaffected by such distribution shifts in the metadata.

In a practical scenario, however, we may also observe a change in the label and image distribution when the spatiotemporal context changes drastically. To investigate this, we create a geographical data split on BigEarthNet, using all samples from Portugal and Ireland only for testing the student (e) and teacher (f). In line with our previous results, the student is more robust to the distribution shift than the teacher model. Nonetheless, comparing the order of magnitude of the mAP values for (e,f) with (a-d) indicates that the distributional shifts in the labels and the images have a bigger impact than the distributional shifts in the metadata alone. We leave a deeper investigation to future work as our main concern is SSL and not domain adaption or transfer learning.

Altogether, the takeaway of this set of experiments is that models relying on metadata suffer from inferior generalization to unseen spatiotemporal settings. Therefore, it is desirable to develop models not relying on spatiotemporal metadata, even though the teacher models using the additional input perform better on the training samples and allow to enhance the student's training and performance.

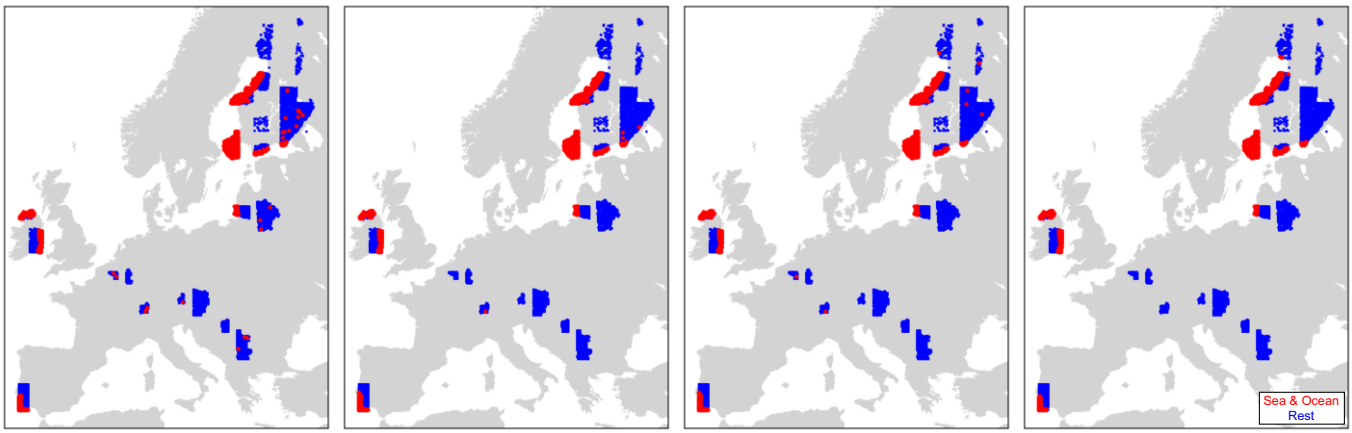
4.6 EuroSAT

We also evaluate our approach on EuroSAT, a highly popular benchmark dataset. The results in Table 4 demonstrate that ST-SSL is able to substantially improve the performance in settings with few labeled samples, e.g., by about 13% accuracy for ten labels. For 40 and 80 labels, the performances of FixMatch and FixMatch + ST-SSL become similar. We reason that this is because the overall performance of the model is already at a very high level in these settings. Therefore, the additional metadata does not add much value to the model, which is able to classify the vast majority of images correctly anyway.

4.7 Qualitative Results

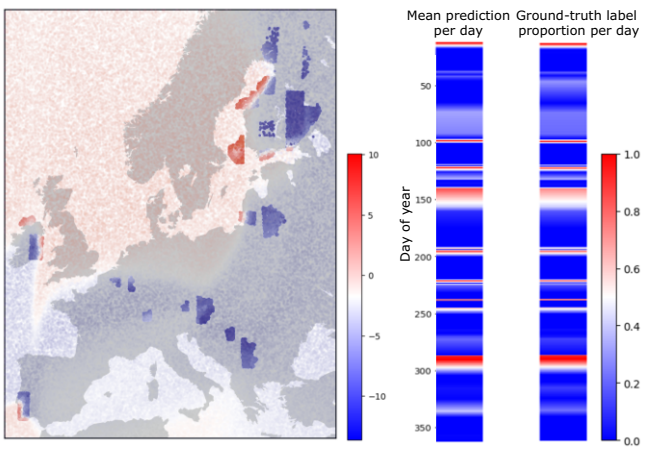
In this section, we qualitatively analyze the effects of the metadata when used as an additional input in our spatiotemporal SSL framework. First, in Figure 4, we illustrate how the metadata improve the predictions on BigEarthNet using the class "sea & ocean" as an example. The ST-SSL teacher (Subfigure (b)) is able to learn the spatial occurrence of the class, which is geographically contiguous in the real world. In concrete terms, this means that the teacher model learns the concept of maritime and continental regions and, thus, mostly avoids false positive predictions in the latter. The learned knowledge is successfully transferred to the student model (Subfigure (c)), which also mostly avoids this type of error. In contrast, the plain FixMatch model (Subfigure (a)) produces many false positives in mainland regions, e.g., Eastern Finland.

To shed more light on this, we conduct an additional experiment where we examine the predicted confidences when various metadata instances are paired with a monochrome, gray image. This allows us to assess the spatiotemporal prior learned by the teacher model. In Figure 5 (a), we can observe high confidences for the class "sea & ocean" in coastal and maritime regions and low confidences in inland regions on the training dataset (solid regions). On the other hand, the prior is clearly useless on out-of-distribution locations. Similarly, we visualize the predicted confidences for "sea & ocean" depending on the time of image acquisition. We observe a considerable sampling bias toward certain times, which the teacher model can learn. Even though the temporal distribution has no real-world semantic meaning in this case (in contrast to the geospatial distribution of the class), the teacher can exploit it to provide strong pseudo-labels on the training data. The student model benefits from the pseudo-labels, but it cannot take up such a bias as it does not access the metadata.



(a) FixMatch (b) FixMatch + ST-SSL Teacher (c) FixMatch + ST-SSL Student (d) Ground Truth

Figure 4: **Qualitative comparison of FixMatch and our method** on BigEarthNet. As FixMatch (a) does not exploit the geospatial context of images, it produces numerous false positive predictions for the class "sea & ocean" in mainland regions (e.g., Eastern Finland). In contrast, our spatiotemporal teacher model (b) is able to largely mitigate this type of error by considering the image geolocation, leading to more accurate pseudo-labels, from which the student model (c) benefits as well. *Best viewed digitally.*



(a) Geolocation vs. predicted confidence (logit); solid regions correspond to dataset support (b) Day of acquisition vs. predicted confidence (prob)

Figure 5: **Visualizing the spatiotemporal prior learned by the teacher model**. We feed a monochrome, gray image together with all metadata instances of BigEarthNet into the model and visualize the confidence for the class "sea & ocean" depending on geolocation (a) and time of image acquisition (b). We can see that the model has learned the dataset's spatiotemporal distribution well (see Figure 4 (d) for geolocation vs. ground-truth label) but does not generalize to OOD locations. *Best viewed digitally.*

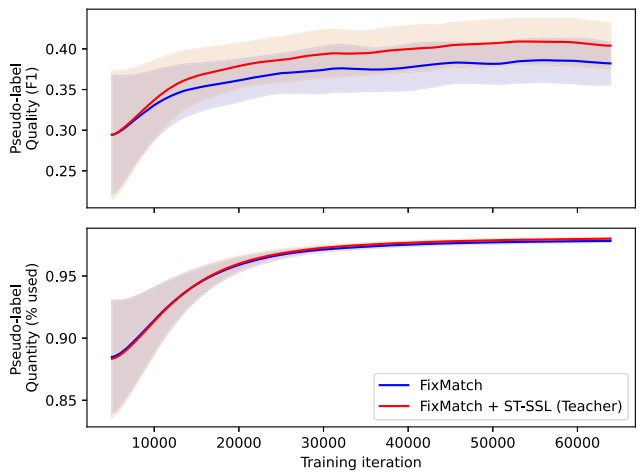


Figure 6: **Comparing pseudo-label quality (top) and quantity (bottom) of FixMatch and FixMatch + ST-SSL**. The additional metadata in ST-SSL allows generate pseudo-labels of higher quality while the quantity, i.e., the fraction of samples passing FixMatch's confidence threshold, is almost identical. *Curves smoothed for visualization. Best viewed digitally.*

5 Conclusion

In this work, we propose a new SSL framework called *Spatiotemporal SSL*. In this framework, a teacher leverages spatiotemporal metadata to generate high-quality pseudo-labels for a student not receiving the additional input. That way, the student can generalize to unseen spatiotemporal contexts while still benefiting from the spatiotemporal information during training. Moreover, we propose a method for joint visual and spatiotemporal modeling and introduce a novel distillation mechanism to enhance the knowledge transfer between teacher and student. We combine Spatiotemporal SSL with several state-of-the-art SSL algorithms and observe consistent and substantial improvements.

Limitations Even though we have seen consistent improvements with Spatiotemporal SSL, there is a dependency on the information contained in the metadata itself. For instance, in a scenario where the distributions of labels and images are independent of the spatiotemporal context, we cannot expect performance gains with our approach as the metadata then does not convey any useful information for the model.

Generalizations and Future Work The proposed paradigm of exploiting additional low-cost data from a second modality to improve SSL is not only applicable to the classification of spatiotemporal images. In fact, it may be beneficial in any other situation where acquiring additional and informative features is substantially easier than annotating numerous samples. For example, within the remote sensing domain, imagery of higher spatial, temporal, or spectral resolution may be collected for a fixed area without large efforts, whereas the high-resolution data may not be available or practicable for use at inference time. Furthermore, the transfer of our approach to other tasks such as object detection and segmentation, or even other domains is conceivable, opening up many directions for future research.

References

[1] T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, pages 2011–2018, 2014.

[2] D. Berthelot, R. Roelofs, K. Sohn, N. Carlini, and A. Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732*, 2021.

[3] H. Chen, R. Tao, Y. Fan, Y. Wang, J. Wang, B. Schiele, X. Xie, B. Raj, and M. Savvides. Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In *ICLR*, 2023. URL https://openreview.net/forum?id=ymt1zQXBDiF.

[4] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3735–3756, 2020.

[5] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. In *CVPR*, pages 6172–6180, 2018.

[6] G. Chu, B. Potetz, W. Wang, A. Howard, Y. Song, F. Brucher, T. Leung, and H. Adam. Geo-aware networks for fine-grained recognition. In *ICCVW*, pages 0–0, 2019.

[7] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *NeurIPS*, 35:197–211, 2022.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[9] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6824–6835, 2021.

[10] Y. Fan, A. Kukleva, D. Dai, and B. Schiele. Revisiting consistency regularization for semi-supervised learning. *IJCV*, 131(3):626–643, 2023.

[11] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[12] W. Han, X. Zhang, Y. Wang, L. Wang, X. Huang, J. Li, S. Wang, W. Chen, X. Li, R. Feng, et al. A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:87–113, 2023.

[13] P. Helber, B. Bischke, A. Dengel, and D. Borth. 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.

[14] J. Jakubik, S. Roy, C. E. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, D. Kimura, N. Simumba, L. Chu, S. K. Mukkavilli, D. Lambhate, K. Das, R. Bangalore, D. Oliveira, M. Muszynski, K. Ankur, M. Ramasubramanian, I. Gurung, S. Khallaghi, H. S. Li, M. Cecil, M. Ahmadi, F. Kordi, H. Alemohammad, M. Maskey, R. Ganti, K. Weldemariam, and R. Ramachandran. Foundation Models for Generalist Geospatial Artificial Intelligence. *Preprint Available on arxiv:2310.18660*, Oct. 2023.

[15] J. Lazarow, K. Sohn, C.-Y. Lee, C.-L. Li, Z. Zhang, and T. Pfister. Unifying distribution alignment as a loss for imbalanced semi-supervised learning. In *WACV*, pages 5644–5653, 2023.

[16] H. Lee and H. Kim. Cdmad: Class-distribution-mismatch-aware debiasing for class-imbalanced semi-supervised learning. *CVPR*, 2024.

[17] J. Li, C. Xiong, and S. C. Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *ICCV*, pages 9475–9484, 2021.

[18] O. Mac Aodha, E. Cole, and P. Perona. Presence-only geographical priors for fine-grained image classification. In *ICCV*, pages 9596–9606, 2019.

[19] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *ICCV*, pages 9414–9423, 2021.

[20] R. Minetto, M. P. Segundo, and S. Sarkar. Hydra: An ensemble of convolutional neural networks for geospatial land classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6530–6541, 2019.

[21] L. P. Osco, J. M. Junior, A. P. M. Ramos, L. A. de Castro Jorge, S. N. Fatholahi, J. de Andrade Silva, E. T. Matsubara, H. Pistori, W. N. Gonçalves, and J. Li. A review on deep learning in uav remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 102:102456, 2021.

[22] Y. Ouali, C. Hudelot, and M. Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

[23] H. Peng, W. Pian, M. Sun, and P. Li. Dynamic re-weighting for long-tailed semi-supervised learning. In *WACV*, pages 6464–6474, 2023.

[24] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *ICCV*, pages 4088–4099, 2023.

[25] T. Salem, S. Workman, and N. Jacobs. Learning a dynamic map of visual appearance. In *CVPR*, pages 12435–12444, 2020.

[26] H. Schmutz, O. Humbert, and P.-A. Mattei. Don't fear the unlabelled: safe semi-supervised learning via debiasing. In *ICLR*, 2023. URL https://openreview.net/forum?id=TN9gQ4x0Ep3.

[27] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 33:596–608, 2020.

[28] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019.

[29] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, and L. Bourdev. Improving image classification with location context. In *ICCV*, pages 1008–1016, 2015.

[30] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[31] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357. PMLR, 2021.

[32] J. E. Van Engelen and H. H. Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

[33] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018.

[34] Y. Wang, H. Chen, Y. Fan, W. Sun, R. Tao, W. Hou, R. Wang, L. Yang, Z. Zhou, L.-Z. Guo, H. Qi, Z. Wu, Y.-F. Li, S. Nakamura, W. Ye, M. Savvides, B. Raj, T. Shinozaki, B. Schiele, J. Wang, X. Xie, and Y. Zhang. Usb: A unified semi-supervised learning benchmark for classification. In *NeurIPS Datasets and Benchmarks Track*, 2022. doi: 10.48550/ARXIV.2208.07204. URL https://arxiv.org/abs/2208.07204.

[35] Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Fan, Z. Wu, J. Wang, M. Savvides, T. Shinozaki, B. Raj, B. Schiele, and X. Xie. Freematch: Self-adaptive thresholding for semi-supervised learning. In *ICLR*, 2023. URL https://openreview.net/forum?id=PDrUPTXJI_A.

[36] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *CVPR*, pages 10857–10866, 2021.

[37] M.-K. Xie, J. Xiao, H.-Z. Liu, G. Niu, M. Sugiyama, and S.-J. Huang. Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. *NeurIPS*, 36, 2024.

[38] Y. Xu, L. Shang, J. Ye, Q. Qian, Y.-F. Li, B. Sun, H. Li, and R. Jin. Dash: Semi-supervised learning with dynamic thresholding. In *ICML*, pages 11525–11536. PMLR, 2021.

[39] L. Yang, X. Li, R. Song, B. Zhao, J. Tao, S. Zhou, J. Liang, and J. Yang. Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information. In *CVPR*, pages 10945–10954, 2022.

[40] X. Yang, Z. Song, I. King, and Z. Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[41] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241:111716, 2020.

[42] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *NeurIPS*, 34:18408–18419, 2021.

[43] M. Zheng, S. You, L. Huang, F. Wang, C. Qian, and C. Xu. Simmatch: Semi-supervised learning with similarity matching. In *CVPR*, pages 14471–14481, 2022.

[44] M. Zheng, S. You, L. Huang, C. Luo, F. Wang, C. Qian, and C. Xu. Simmatchv2: Semi-supervised learning with graph consistency. In *ICCV*, pages 16432–16442, 2023.