The color out of space: learning self-supervised representations for Earth Observation imagery

Stefano Vincenzi*, Angelo Porrello*, Pietro Buzzega*, Marco Cipriano*, Pietro Fronte‡, Roberto Cuccu‡, Carla Ippoliti†, Annamaria Conte†, Simone Calderara*

*AImageLab, University of Modena and Reggio Emilia, Modena, Italy
†Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise 'G.Caporale', Teramo, Italy
‡Progressive Systems Srl, Frascati – Rome, Italy

***Abstract*—The recent growth in the number of satellite images fosters the development of effective deep-learning techniques for Remote Sensing (RS). However, their full potential is untapped due to the lack of large annotated datasets. Such a problem is usually countered by fine-tuning a feature extractor that is previously trained on the ImageNet dataset. Unfortunately, the domain of natural images differs from the RS one, which hinders the final performance. In this work, we propose to learn meaningful representations from satellite imagery, leveraging its high-dimensionality spectral bands to reconstruct the visible colors. We conduct experiments on land cover classification (BigEarthNet) and West Nile Virus detection, showing that colorization is a solid pretext task for training a feature extractor. Furthermore, we qualitatively observe that guesses based on natural images and colorization rely on different parts of the input. This paves the way to an ensemble model that eventually outperforms both the above-mentioned techniques.**

I. Introduction

Over the last decades, Remote Sensing has become an enabling factor for a broad spectrum of applications such as disaster prevention [1], wildfire detection [2], vector-borne disease [3], and climate change [4]. These applications benefit from a higher number of satellite imagery captured at unprecedented rhythms [5], thus making every aspect of the Earth's surface constantly monitored. Machine learning and Computer Vision provide valid tools to exploit these data in an efficient way. Indeed, a synergy between Earth Observation and Deep Learning techniques led to promising results, as highlighted by recent advances in land use and land cover classification [6], image fusion [7], and semantic segmentation [8].

Despite the amount of raw information being significant, the exploitation of these data still raises an open problem. Indeed, the prevailing learning paradigm – the supervised one – frames the presence of labeled data as a crucial factor. However, acquiring a huge amount of ground truth data is expensive and requires expert staff, equipment, and in-field measurements. This often restrains the development of many downstream tasks that are important for paving the way to the above-mentioned applications.

To mitigate such a problem, a common solution [9] exploits models that are pre-trained on the ImageNet [10] dataset. In detail, the learning phase is conducted as follows: firstly, a deep network is trained on ImageNet until it reaches good performance on image categorization; secondly, a fine-tuning step is carried out on a target task (*e.g.* land cover classification). This way, one can achieve acceptable results even in the presence of few labeled examples, as the second step just adapts a set of general-purpose features to the new task. However, this approach is limited only to the tasks involving RGB images as input. Satellite imagery represents a domain that is quite different from the RGB one, thus making the ImageNet pre-training only partially suitable.

These considerations reveal the need for novel approaches that are tailored for satellite imagery. To build transferable representations, two kinds of approaches arise from the literature: annotation-based methods and self-supervised ones. The authors of [11] fulfill the principle of the first branch by investigating in-domain representation learning. They shift the pre-training stage from ImageNet to a labeled dataset specific for remote sensing. As an example, one could leverage BigEarthNet [12], which has been recently released for land-cover classification. On the other hand, Tile2Vec [13] extracts informative features in a self-supervised fashion. The authors rely on the assumption that spatially close tiles share similar information: therefore, their corresponding representations should be placed closer than tiles that are far apart. In doing so, one does not need labeled data for extracting representations, but lacks robustness when close tiles are not similar.

Similarly to [13], we propose a novel representation learning procedure for satellite imagery, which devises a self-supervised algorithm. In more detail, we require the network to recover the RGB information by means of other spectral bands solely. For the rest of the article, we adopt the term "spectral bands" for indicating the subset of the bands not including the RGB. Our approach closely relates to colorization, which turns out to encourage robust and high-level feature representations [14], [15]. We feel this pretext task being particularly useful for satellite imagery, as the connection between colors and semantics appears strong: for instance, sea waters feature the blue color, vegetation regions the green one or arable lands prefer warm tones. We inject such a prior knowledge through an encoder-decoder architecture that – differently from concurrent works – exploits spectral bands (*e.g.* short-wave infrared, near-infrared, etc.) instead of grayscale information to infer color channels. Once the model has reached good capabilities on tile colorization, we use its encoder as a feature extractor

for the later step, namely fine-tuning on a remote sensing task. We found that the representations learnt by colorization leads to remarkable results and semantically diverge from the ones computed on top of RGB channels. Taking advantage of these findings, we set up an ensemble model, which averages the predictions from two distinct branches at inference time (the one fed with spectral bands, the other with RGB information). We show that ensembling features this way leads to better results. To the best of our knowledge, our work is the first investigating colorization as a guide towards suitable features for remote sensing applications.

To show the effectiveness of our proposal, we assess it in two different settings. Firstly, we conduct experiments on land-cover classification, comparing our solution with two baselines, namely training from scratch and fine-tuning the ImageNet pre-training. We show that colorization is particularly effective when few annotations are available for the target tasks. This makes our proposal viable for scenarios where gathering many labeled data is not practicable. To demonstrate such a claim, we additionally conduct experiments on the "West Nile Virus" cases collected in the frame of the Surveillance plan put in place by the Ministry of Health, with the aim of predicting presence/absence across the Italian territory.

II. Related Works

A. Land cover - Land use classification

Recently, the categorization of land-covers has attracted wide interest, as it allows for the collection of statistics, activities planning, and climate changes monitoring. To address these challenges, the authors of [16] exploit Convolutional Neural Networks (CNN) to extract representations encoding both spectral and spatial information. To speed up the learning process, they advocate for a prior dimensionality reduction step across the spectra, as they observe a high correlation in this dimension. Among works focusing on how to exploit spectral bands, [17] devises Recurrent Neural Networks (RNNs) to handle the redundancy underlying adjacent spectral channels. Similarly, [18] proposes a 3D-CNN framework, which can naturally joint spatial and spectral information in an end-to-end fashion without requiring any pre-processing step.

While these approaches concern the design of the feature extractor, our work is primarily engaged in the scenarios in which few labeled examples are available. In these contexts, fine-tuning pre-trained models often mitigate the lack of a large annotated dataset, yielding great performance in some cases [19], [20]. Intuitively, the representations learned from ImageNet (1 million images belonging to 1000 classes) encode a prior knowledge on natural images, thus facilitating the transfer to different domains. Instead, [11] proposes in-domain fine-tuning, where the pre-training stage performs on a remote sensing dataset. The authors found in-domain representations to be especially effective with limited data (1000 training examples), surpassing the performance yielded by the ImageNet initialization. As a final remark, one could reduce overfitting through data augmentation [21] (*i.e.* flip, translation, and rotation), which increases both the diversity and volume of training data.

B. Unsupervised Representations Learning

Unsupervised and self-supervised methods were introduced to learn general visual features from unlabeled data [22]. These approaches often rely on *pretext tasks*, which attempt to compensate for the lack of labels through an artificial supervision signal. In so doing, the learned representations hopefully embody meaningful information that is beneficial to downstream tasks.

Reconstructions-based methods. Under this perspective, generative models can be considered as self-supervised methods, where the reconstruction of the input acts as a pretext task. Denoising Autoencoders [23] contribute to this line of research: here, the learner has to recover the original input from a corrupted version. The idea is that good representations are those capturing stable patterns, which should be recovered even in the presence of a partial or noisy observation. In remote sensing, autoencoders are often applied [16], [24], [25] to reduce the dimensionality of the feature space. This yields the twofold advantage of decreasing the correlation lying in spectral bands and reducing the overall computational effort.

Classification-based methods. [26] frames the pretext task as a classification problem, where the learner guesses which rotation (0°, 90°, 180° and 270°) has been applied to its input. The authors observe that recognizing the input transformation behaves as a proxy for object recognition: the higher the accuracy on the upstream task, the higher the accuracy on the downstream one. Considering two random patches from a given image, [27] asks the network to infer the relative position between those. This encourages the learner to recognize the parts that make up the object as well as their relations. Similarly, [28] presents a jigsaw puzzle to the network, which has to place the shuffled patches back to their original locations.

Colorization-based methods. Given a grey-scale image as input, colorization is the process of predicting realistic colors as output. A qualitative analysis conducted in [29] shows that colorization-driven representations capture semantic information, grouping together high-level objects that display low-level variations (*e.g.* color or pose). [30] concerns the ambiguity and ill-posedness of colorization, arguing that several solutions may be assessed for a given grey-scale image. On this basis, the authors exploit Conditional Variational Autoencoder (CVAE) to produce diverse colorizations, thus naturally complying with the multi-modal nature of the problem. Instead, [31] focuses on the design of the inference pipeline and proposes a two-stage procedure: *i)* a pixel-wise descriptor is built by VGG-16 feature maps taken at different resolutions; *ii)* the descriptors are then fed into a fully connected layer, which outputs hue and chroma distributions. Split-Brain Autoencoders [15] relies on a network composed of two disjoint modules, each of which predicts a subset of color channels from another. The authors argue that this schema induces transferable representations, the latter taking into account all input dimensions (instead of gray-scale solely).

III. MODEL

Overview. Our main goal consists in finding a good initialization for the classifier, in such a way that it can later capture meaningful and robust patters even in presence of few labeled data. To this purpose, we devise a two-stage procedure tailored for satellite imagery tasks, which prepends a colorization step (Sec. III-A) to a fine-tuning one (Sec. III-B).

As depicted in Fig. 1 (a), our proposal leverages an encoder-decoder architecture for feature learning. In doing so, we do not require the model to reconstruct its input: differently, we set up an asymmetry between input (spectral bands) and output (color channels). This way, we expect the encoder to capture meaningful information about soil and environmental characteristics. Afterward, we exploit the encoder and its representation capabilities to tackle a downstream task (*e.g.* land cover classification, see Fig. 1 (b)). Eventually, an ensemble model (see Sec. III-C for additional details) further refines the final prediction combining the outputs from the two input modalities (RGB and spectral bands).

A. Colorization

In formal terms, the encoder network $\mathcal{F}$ takes $\mathbf{S} \in \mathbb{R}^{H \times W \times C}$ as input, where C equals the number of spectral bands available to the model and H and W the input resolution (height and width respectively). The decoder network produces a tensor $\widehat{\mathbf{X}} \in \mathbb{R}^{H \times W \times 2}$, which yields the pixel-wise predictions in terms of a and b coordinates in the CIE *Lab* color space. On this latter point, a naive strategy would simply define the expected output in terms of RGB: nevertheless, as pointed out in [31], modeling colors as RGB values may not yield an effective training signal. Differently, we adhere to the guideline described in [32] and frame the problem in the CIE *Lab* space. Here, a color is defined with a lightness component L and $a*b$ values carrying the chromatic content. The effectiveness of this space comes from the fact that colors are encoded accordingly to human perception: namely, the distance between two points reflects the amount of visually perceived change between the corresponding colors.

Encoder. We opt for ResNet18 [33] as backbone network for the encoder, which hence consists of four blocks with two residual units each. As pointed out in [34], thanks to their residual units and skip connections, ResNet-based networks are more suitable for self-supervised representation learning. Indeed, when compared to other popular architectures (*e.g.* AlexNet), residual networks favorably preserves representations from degrading towards the end of the network and therefore results in better performance.

Decoder. In designing the decoder network, we mirror the architecture of the encoder, replacing the first convolutional layer of each residual block with its transposed counterpart. Moreover, we add an upsampling operation to the top of the decoder, followed by a batch normalization layer, a ReLU activation, and a transposed convolution. The latter reduces the number of features maps to 2: this way, the output dimensionality matches the ground truth one.

Colorization Loss. Recent works [15], [29], [32] investigate various loss functions, questioning their contributions to colorization results (intended as performance on either the target task or the pretext one). Despite a regression objective (*e.g.* the mean squared error) being a valid baseline, these works show that treating the problem as a multinomial classification leads to better results. However, the overall training time increases considerably because of the additional information taken into account. In our case, this would add up to the burdensome computations required by hyperspectral images, thus resulting even more expensive. For this reason, we limit our experiments to the mean absolute error $\mathcal{L}_1(\cdot, \cdot)$, as follows:

$$\mathcal{L}_1(\widehat{\mathbf{X}}, \mathbf{X}) = \lambda \sum_{h,w} \left| \widehat{x}_{h,w}^{(a)} - x_{h,w}^{(a)} \right| + \left| \widehat{x}_{h,w}^{(b)} - x_{h,w}^{(b)} \right|, \quad (1)$$

where $\mathbf{X}$ represents the $a*b$ ground truth colorization and $\lambda = 100$ is a weighting term that prevents numerical instabilities.

B. Fine-tuning

Once the encoder-decoder has been trained, we turn our attention to the downstream task and exploit the encoder $\mathcal{F}(\cdot)$ as a pre-trained feature extractor. To achieve this, we need a single amendment to the network: a final linear transformation that maps bottleneck features $\mathbf{H} = \mathcal{F}(\mathbf{S})$ to the classification output space $\widehat{\mathbf{y}} = \mathbf{W}^T\mathbf{H} + \mathbf{b}$.

Classification Loss We make use of two different losses in our experiments: when dealing with a multi-label task as the land cover classification one (*i.e.* each example can be categorized into multiple classes), the objective function resembles a binary cross-entropy term averaged over C classes:

$$\mathcal{L}(\widehat{\mathbf{y}}, \mathbf{y}) = -\frac{1}{C} \sum_i \mathbf{y}_i \log \sigma\left(\widehat{\mathbf{y}}_i\right) + \left(1 - \mathbf{y}_i\right) \log\left(1 - \sigma\left(\widehat{\mathbf{y}}_i\right)\right),$$

where $\mathbf{y}$ indicates the ground-truth multi-hot encoding vector and σ the sigmoid function. Differently, we use the binary cross-entropy loss to treat the West Nile Disease case study.

C. Model ensemble

As pointed out in [15], a network trained on colorization specializes just on a subset of the available data (in our case, spectral bands) and cannot exploit the information coming from its ground truth (the RGB color images). To further take advantage of color information, we set up an ensemble model at inference time (so, no additional training steps required). As shown in Fig. 1 (c), the ensemble is formed by two independent branches taking the RGB channels and the spectral bands as input respectively. The first one is pre-trained on classification (ImageNet) and the second one on colorization; both are fine-tuned separately on the given classification task. The ensemble-level predictions are simply computed by averaging the responses from the two branches:

$$\widehat{\mathbf{y}}_{\text{ENS}} = \frac{\sigma(\widehat{\mathbf{y}}_{\text{RGB}}) + \sigma(\widehat{\mathbf{y}}_{\text{SPECTRAL}})}{2}. \quad (2)$$

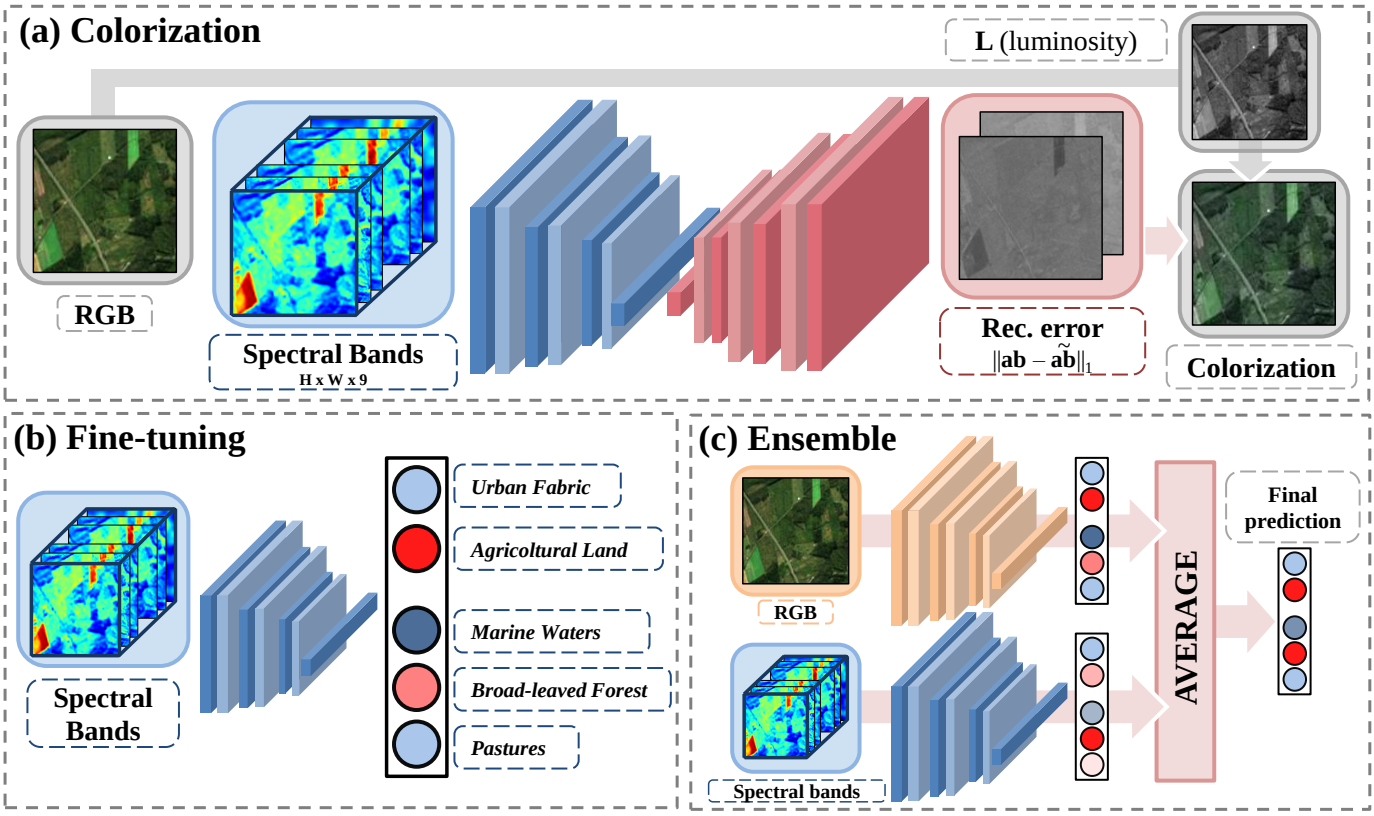

Fig. 1. An overview of the proposed pipeline for feature learning on satellite imagery.

IV. Datasets

The two datasets we rely on data acquired through the Sentinel-2A and 2B satellites developed by the European Space Agency (ESA). These satellites provide a multi-spectral imagery over the earth with 12 spectral bands (covering the visible, near and short wave infrared part of the electromagnetic spectrum) at three different spatial resolutions (10, 20 and 60 meters per pixel).

A. Land-cover classification - BigEarthNet

In Remote Sensing, the main bottleneck in the adoption of deep networks was the lack of a large training set. Indeed, existing datasets (as Eurosat [35], PatterNet [36], UC Merced Land Use Dataset [37]) include a small number of annotated images, hence resulting inadequate for training very deep networks. To overcome this problem, [12] introduces BigEarthNet, a novel large scale dataset collecting 590 326 tiles. Each example comprises of 12 bands (RGB included) and multiple land-cover classes (provided by the CORINE Land Cover (CLC) database [38]) as ground truth.

Originally, the number of classes amounted to 43: but, the authors of [39] argue that some CORINE classes cannot be easily inferred by looking at Sentinel-2 images solely. Indeed, some labels may not be recognizable at such low resolution (the highest one is 120×120 pixels for $10m$ bands) and other ones would require temporal information for being correctly discriminated (*e.g.* non-irrigated arable land *vs.* permanently irrigated land). For these reasons, in our experiments we adopt the class-nomenclature proposed in [39], which reduces the number of classes to 19. Moreover, we discard the 70 987 patches displaying lands that are fully covered by clouds, cloud shadows, and seasonal snow.

B. West Nile Disease Dataset

In the last decade, numerous studies have examined the complex interactions among vectors, hosts, and pathogens [3], [40]. In particular, one of the major threat worldwide studied is represented by West Nile Disease (WND), a mosquito-borne disease caused by West Nile virus (WNV). Mosquitoes presence and abundance have been extensively proved to be associated with climatic and environmental factors such as temperatures, vegetation, rainfall [40]–[42], and remote sensing has been an important key source for data collection. Our capacity to collect and store data continues to expand rapidly and this requires the incorporation of new analytical techniques able to process Earth Observation (EO) data establishing pipelines to turn near real-time "big data" into "smart data" [43]. In this context, Deep techniques could provide useful tools to process data and automatically identify patterns able to make accurate predictions of the spatio-temporal re-emergence and spread of the West Nile Disease in Italy. With

this aim, we collected data from the Copernicus program and paired Sentinel 2 (S2) EO data with ground truth WND data.

Disease sites are collected through the National Disease Notification System of the Ministry of Health (SIMAN www.vetinfo.sanita.it) [44]. We start with the analysis of the 2018 epidemic, one of the most spread on the Italian territory. We frame the problem as a binary classification task with the final purpose of predicting positive and negative WND sites analyzing multi-spectral bands. Positive cases are geographically located mainly in Po valley, in Sardinia and some spots in the rest of Italy [45]: the location of each case of birds, mosquitoes and horses, was visually inspected for the accuracy needs in the analysis. Negative sites, being not always available in the national database due to the surveillance plan strategy, were derived as pseudo-absence ground truth data, either in the space (points located in areas where the disease was never reported in the past) and in the time (a random date in months previous the reported positivity in mosquitoes collections).

WND dataset comprises of $1\,488$ distinct cases, divided into 962 negatives and 526 positives. Each case comes with a variable number of Sentinel-2 patches (corresponding to various acquisitions over time), thus leading to $18\,684$ spectral images in total.

V. Experiments

In this section, we test our proposal as a pre-training strategy for the later fine-tuning step. We compare the results yielded by colorization to those achieved by two baselines: training from scratch [46] and the common ImageNet pre-training. In doing so, we mimic scenarios with few labeled data by reducing the amount of examples available at training time (*e.g.* $1\,000$, $5\,000$, etc...).

A. Evaluation Protocols

Land-Cover Classification. We strictly follow the guidelines provided by [11] when assessing the performance on the BigEarthNet benchmark. Namely, we form the training set by sampling 60% of the total examples considered, retaining 20% for the validation set and 20% for the test set. We measure the results in terms of Mean-Average-Precision (mAP), which also considers the order in which predictions are given to the user. We check the performance every 10 epochs and retain the weights that yield the higher mAP score on the validation set.

West Nile Disease. Here, we adopt the stratified holdout strategy, which ensures the class probabilities of training and test being close to each other. The metrics of interest are precision, recall and F1 score, the latter accounting for the slight imbalance that occurs at class level (indeed, negatives cases appear more frequently than positives ones).

B. Implementation details

BigEarthNet. We exploit the normalization technique described in [43], [47] computing the 2^{nd} and 98^{th} percentile values to normalize each band. This method is more robust than the common min-max normalization, as it is less sensitive to outliers. Before feeding the spectral bands into the model – as they come with different spatial resolutions – we apply a cubic interpolation to get a dimension of 128×128.

Colorization. To broaden the diversity of available data, we apply data augmentation (*i.e.* rotation, horizontal and, vertical flip). We initialize the network according to [46] and train for 50 epochs on the full BigEarthNet, setting the batch size equal to 16 and leveraging Stochastic Gradient Descent (SGD) as optimizer (with a learning rate fixed at 0.01).

Land-Cover Classification. We train the model for 30 epochs whether the full dataset is available; otherwise we increase the epochs to 50. The learning rate is set to 0.1 and divided by 10 at the 10^{th} and 40^{th} epoch. The batch size equals 64.

West Nile Disease Differently from the previous cases, we apply neither upscaling nor pixel-normalization, as all channels are provided at the same resolution (224×224) and their values lie within the range $[0, 1]$. We leverage the network trained for colorization on BigEarthNet. Since we rely on a subset of the spectral bands (B_1, B_{8A}, B_{11} and B_{12}), we fix the first convolutional layer so that it takes 4 channels as input. We optimize the model for 30 epochs, with a batch size of 32 and an initial learning rate of 0.001, multiplied by 0.1 after 25 epochs.

TABLE I
Performance (mAP) on BigEarthNet for different strategies to vary the number of training examples.

Input	pre-training	1k	5k	10k	50k	Full
RGB	from scratch	.486	.608	.645	.744	.851
RGB	ImageNet	.620	.695	.726	.786	**.879**
Spectral	from scratch	.555	.667	.711	.767	.866
Spectral	ImageNet	.578	.627	.681	.773	**.879**
Spectral	Color. (our)	**.622**	**.730**	**.760**	**.793**	.860

TABLE II
Ensemble model – results (mAP) on BigEarthNet.

Input	pre-training	1k	5k	10k	50k	Full
RGB	ImageNet	.620	.695	.726	.786	.879
Spectral	Colorization	.622	.730	.760	.793	.860
Ensemble	ImagNet+ImageNet	.649	.707	.749	.815	**.904**
Ensemble	Color.+ImageNet	**.656**	**.751**	**.778**	**.823**	.896

C. Results of Colorization pre-training

Based on the final performance reported in Tab. I, one can observe the initialization offered by colorization surpassing the other alternatives. Such a claim especially holds in presence of scarce data, thus complying with the goals we have striven for in this work. This does not apply when the learner faces up to the entire training set (519k examples): such evidence – already encountered in [11] – deserves more investigations that we will conduct in future works.

TABLE III
PERFORMANCE (ACC. ACCURACY, PR. PRECISION, RC. RECALL) ON THE WEST NILE DISEASE CASE STUDY, FOR DIFFERENT METHODS AND PRE-TRAINING STRATEGIES.

Input	pre-training	acc.	pr.	rc.	F1
Random classifier	-	.503	.391	.395	.393
RGB	from scratch	.652	.542	.941	.688
RGB	ImageNet	.865	.819	.857	.838
$B_{1,8A,11,12}$	from scratch	.756	.662	.817	.732
$B_{1,8A,11,12}$	Colorization	.852	.823	.811	.817
Ensemble	Color.+ImageNet	**.880**	**.855**	**.850**	**.852**

TABLE IV
COMPARISON BETWEEN SEVERAL BASELINES AND OUR ENSEMBLE METHOD ON BIGEARTHNET.

Method	pr.	rc.	F1
K-Branch CNN [12]	.716	**.789**	.727
VGG19 [12]	.798	.767	.759
ResNet-50 [12]	.813	.774	.771
ResNet-101 [12]	.801	.774	.764
ResNet-152 [12]	.817	.762	.765
Ensemble (our)	**.843**	.781	**.811**

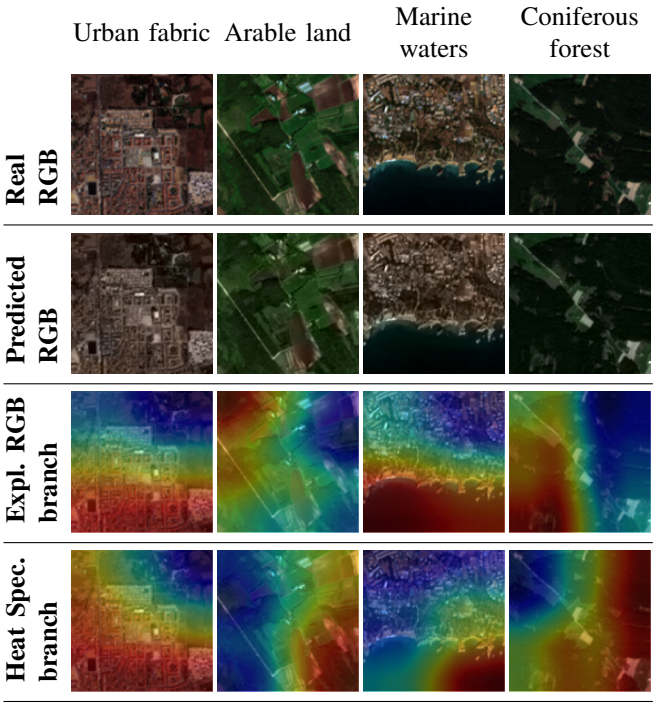

Fig. 2. Some examples of the BigEarthNet dataset, coupled with the predicted colorization and visual explanations provided by the ensemble method for RGB and spectral inputs.

Results shown by Tab. I let us draw additional remarks: *i)* as one would expect, the ImageNet pre-training performs good for an RGB input; however, when dealing with the spectral domain, even a random initialization outperforms it; *ii)* colorization is the sole that rewards the exploitation of spectral bands and justifies their usage in place of RGB.

D. Results of the Model ensemble

Here, we primarily assess the effectiveness of the ensemble discussed in Sec. III-C on BigEarthNet. In this regard, Tab. II compares the performance that can be reached when leveraging a twofold source of information (RGB and spectral bands): firstly, the ensemble model largely outperforms those that consider a single input modality; secondly, colorization presents an improvement over the ImageNet pre-training.

Tab. III reports the results achieved on the West Nile Disease case study discussed in Sec IV-B. To provide a better understanding, we additionally furnish a simple baseline (*i.e.* "random classifier") that computes predictions by randomly guessing from the class-prior distribution of the training set. As a first remark, all the networks we trained exceed random guessing, hence suggesting they effectively learned meaningful and suitable features for the problem at hand. Secondly, the ensemble model plays an important role even in this case, surpassing networks based on a single modality by a consistent margin.

E. Comparison with the state of the art

To further highlight the contributions of our proposal, we compare it with the networks discussed in [12]. Results reported in Tab. IV confirm the above intuitions: the ensemble we build upon ResNet-18 outperforms heavier and overparametrized networks like ResNet-101 or ResNet-152. Notably, we found a large improvement in precision, suggesting that our proposal is capable of returning only the categories that are relevant to the semantics of the input tile.

It is noted the fairness of the comparisons above, as both our ensemble and the baselines leverage the same amount of information in input (namely, spectral bands and color channels). Nevertheless, an important difference subsists in the way information is consumed: while [12] stacks both the input modalities to form a single input tensor, we distinguish two independent paths that eventually cross in the output space. This way, we can benefit from two different pre-training, each one being devoted to its modality: the one offered by colorization – which works well for spectral bands – and the ImageNet one – which instead represents a natural and reasonable choice for dealing with RGB images.

F. Model Explanation - Towards diverse feature sets

We believe the strength of our ensemble approach being a result of the diversity among the individual learners. We investigate the truthfulness of such a claim from a *model explanation* perspective, questioning which information in the input makes our models arrive at their decisions [48]. In particular, we take advantage of GradCam [49] to assess whether the two branches look for different properties within their inputs. The third and fourth rows of Fig. 2 highlight the input regions that have been considered important for predicting the target category (we limit the analysis to the class

denoting the highest confidence score). As one can see, the explanations provided by the two branches visually diverge, thus qualitatively confirming the weak correlation between their representations.

VI. Conclusion

In this work, we propose a self-supervised learning approach for satellite imagery, which moves towards a proper initialization for deep networks facing up to Remote Sensing tasks. Our proposal builds upon two steps: firstly, we ask an encoder-decoder architecture to predict color channels from those capturing spectral information (colorization); secondly, we exploit its encoder as a pre-trained feature extractor for a classification task (*i.e.* land-cover categorization and the West Nile Disease case study). We observe that the initialization we devised leads to remarkable results, exceeding the baselines especially in presence of scarce labeled data. Moreover, we qualitatively observe that representations learned through colorization are different from the ones driven by the RGB channels. Based on this finding, we set up an ensemble model that achieves the highest results in all the scenarios under consideration.

Acknowledgment

The research described in this paper has been conducted within the project 'AIDEO' (AI and EO as Innovative Methods for Monitoring West Nile Virus Spread). The project is being developed within the scope of the ESA EO Science for Society Permanently Open Call for Proposals EOEP-5 BLOCK 4 (ESA AO/1-9101/17/I-NB).

References

[1] G. J. Schumann, G. R. Brakenridge, A. J. Kettner, R. Kashif, and E. Niebuhr, "Assisting flood disaster response with earth observation data and products: a critical assessment," *Remote Sensing*, vol. 10, no. 8, p. 1230, 2018.

[2] F. Filipponi, "Exploitation of sentinel-2 time series to map burned areas at the national level: A case study on the 2017 italy wildfires," *Remote Sensing*, vol. 11, no. 6, p. 622, 2019.

[3] C. Ippoliti, L. Candeloro, M. Gilbert, M. Goffredo, G. Mancini, G. Curci, S. Falasca, S. Tora, A. Di Lorenzo, M. Quaglia *et al.*, "Defining ecological regions in italy based on a multivariate clustering approach: A first step towards a targeted vector borne disease surveillance," *PloS one*, vol. 14, no. 7, 2019.

[4] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown *et al.*, "Tackling climate change with machine learning," *arXiv preprint arXiv:1906.05433*, 2019.

[5] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort *et al.*, "Sentinel-2: Esa's optical high-resolution mission for gmes operational services," *Remote sensing of Environment*, vol. 120, pp. 25–36, 2012.

[6] S. Ji, Z. Chi, A. Xu, Y. Shi, and Y. Duan, "3d convolutional neural networks for crop classification with multi-temporal remote sensing images," *Remote Sensing*, vol. 10, p. 75, 01 2018.

[7] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, no. 7, p. 594, 2016.

[8] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *arXiv preprint arXiv:1606.02585*, 2016.

[9] M. Huh, P. Agrawal, and A. A. Efros, "What makes imagenet good for transfer learning?" *arXiv preprint arXiv:1608.08614*, 2016.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[11] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby, "In-domain representation learning for remote sensing," *arXiv preprint arXiv:1911.06721*, 2019.

[12] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," *arXiv preprint arXiv:1902.06148*, 2019.

[13] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3967–3974.

[14] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6874–6883.

[15] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1058–1067.

[16] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2015, pp. 4959–4962.

[17] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5384–5394, 2019.

[18] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network," *Remote Sensing*, vol. 9, no. 1, p. 67, 2017.

[19] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105–109, 2015.

[20] K. Nogueira, O. A. Penatti, and J. A. Dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[21] X. Yu, X. Wu, C. Luo, and P. Ren, "Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework," *GIScience & Remote Sensing*, vol. 54, no. 5, pp. 741–758, 2017.

[22] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *arXiv preprint arXiv:1902.06162*, 2019.

[23] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[24] Z. Lin, Y. Chen, X. Zhao, and G. Wang, "Spectral-spatial classification of hyperspectral image using autoencoders," in *2013 9th International Conference on Information, Communications & Signal Processing*. IEEE, 2013, pp. 1–5.

[25] X. Ma, H. Wang, and J. Geng, "Spectral–spatial classification of hyperspectral image based on deep auto-encoder," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 9, pp. 4073–4085, 2016.

[26] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[27] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1422–1430.

[28] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–84.

[29] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6874–6883.

[30] A. Deshpande, J. Lu, M.-C. Yeh, M. Jin Chong, and D. Forsyth, "Learning diverse image colorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6837–6845.

[31] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European Conference on Computer Vision*. Springer, 2016, pp. 577–593.

[32] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European conference on computer vision*. Springer, 2016, pp. 649–666.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[34] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019, pp. 1920–1929.

[35] 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*, vol. 12, no. 7, pp. 2217–2226, 2019.

[36] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 197–209, 2018.

[37] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.

[38] J. Feranec, T. Soukup, G. Hazeu, and G. Jaffrain, *European landscape dynamics: CORINE land cover data*. CRC Press, 2016.

[39] G. Sumbul, J. Kang, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, and B. Demir, "Bigearthnet deep learning models with a new class-nomenclature for remote sensing image understanding," *arXiv preprint arXiv:2001.06372*, 2020.

[40] A. Tran, B. Sudre, S. Paz, M. Rossi, A. Desbrosse, V. Chevalier, and J. C. Semenza, "Environmental predictors of west nile fever risk in europe," *International journal of health geographics*, vol. 13, no. 1, p. 26, 2014.

[41] D. Bisanzio, M. Giacobini, L. Bertolotti, A. Mosca, L. Balbo, U. Kitron, and G. M. Vazquez-Prokopec, "Spatio-temporal patterns of distribution of west nile virus vectors in eastern piedmont region, italy," *Parasites & vectors*, vol. 4, no. 1, p. 230, 2011.

[42] A. Conte, L. Candeloro, C. Ippoliti, F. Monaco, F. De Massis, R. Bruno, D. Di Sabatino, M. L. Danzetta, A. Benjelloun, B. Belkadi *et al.*, "Spatio-temporal identification of areas suitable for west nile disease in the mediterranean basin and central europe," *PloS one*, vol. 10, no. 12, 2015.

[43] S. Vincenzi, A. Porrello, P. Buzzega, A. Conte, C. Ippoliti, L. Candeloro, A. Di Lorenzo, A. C. Dondona, and S. Calderara, "Spotting insects from satellites: modeling the presence of culicoides imicola through deep cnns," *arXiv preprint arXiv:1911.10024*, 2019.

[44] P. Colangeli, S. Iannetti, A. Cerella, C. Ippoliti, and A. Di, "Sistema nazionale di notifica delle malattie degli animali," *Veterinaria Italiana*, vol. 47, no. 3, pp. 291–301, 2011.

[45] F. Riccardo, F. Monaco, A. Bella, G. Savini, F. Russo, R. Cagarelli, M. Dottori, C. Rizzo, G. Venturi, M. Di Luca *et al.*, "An early start of west nile virus seasonal transmission: the added value of one heath surveillance in detecting early circulation and triggering timely response in italy, june to july 2018," *Eurosurveillance*, vol. 23, no. 32, 2018.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[47] G. Prathap and I. Afanasyev, "Deep learning approach for building detection in satellite multispectral imagery," in *2018 International Conference on Intelligent Systems (IS)*. IEEE, 2018, pp. 461–465.

[48] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.