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

S. Vincenzi*, A. Porrello*, P. Buzzega*, M. Cipriano*
P. Fronte†, R. Cuccu†, C. Ippoliti§, Annamaria Conte§, Simone Calderara*
*AImageLab, University of Modena and Reggio Emilia, §Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise 'G.Caporale', †Progressive Systems Srl

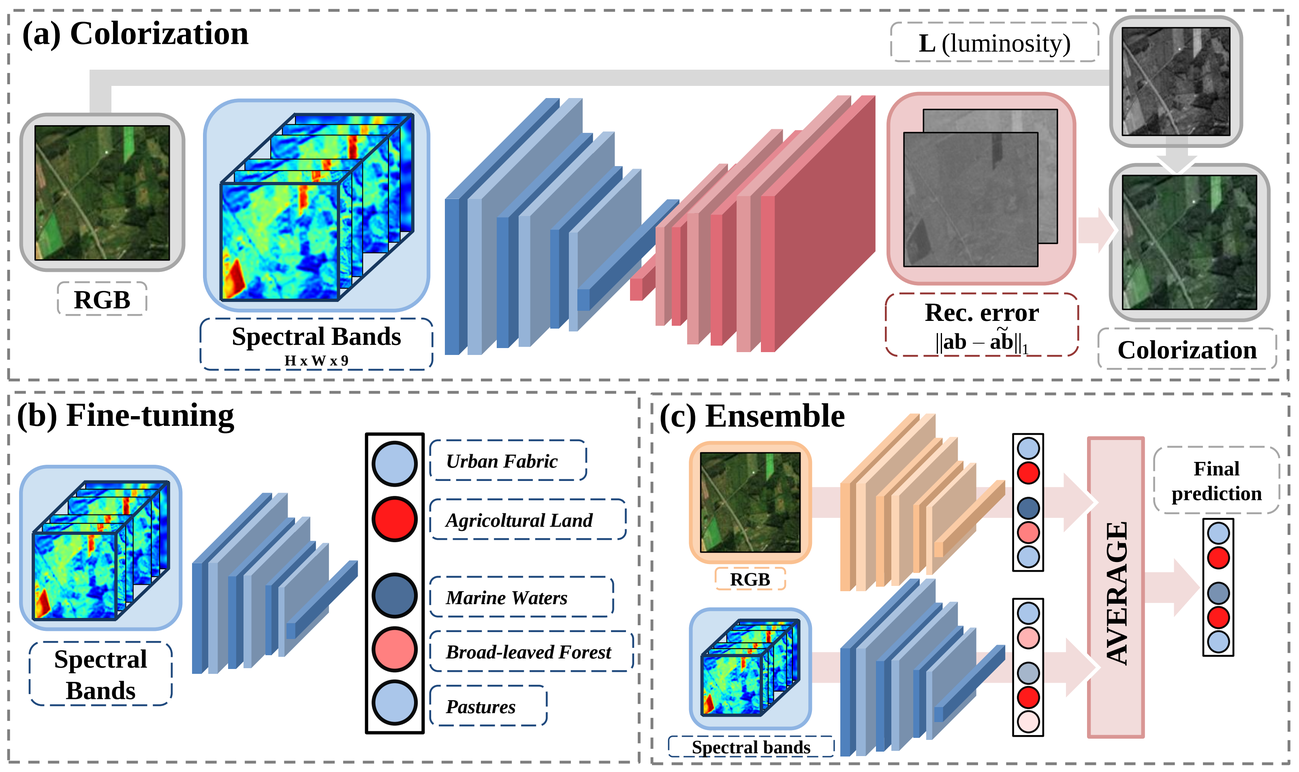
ICPR 2020 — 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION — Milan, Italy 10 | 15 January 2021

1. Introduction

Remote Sensing has become an enabling factor for a broad spectrum of applications such as disaster prevention, vector-borne disease, and climate change. These applications benefit from a higher number of satellitary images, however, acquiring a huge amount of ground truth data is expensive and requires expertise.
A common solution exploits models that are pre-trained on the ImageNet dataset. This approach is limited only to the tasks involving RGB images as input.
We propose a novel representation learning procedure closely relates to **colorization**.
Once the model has reached good capabilities on tile colorization, we use its encoder as a feature extractor, finetuned on a remote sensing task.
The representations learned 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.
We employ the **BigEarthNet** [3] dataset.

2. Model

Our main goal consists of finding a good initialization for the classifier, in such a way that it can later capture meaningful and robust patterns even in presence of few labeled data.
Our proposal leverages an **encoder-decoder architecture**. Afterward, we exploit the encoder to tackle a downstream task (*e.g.* land cover classification).
Eventually, an ensemble model further refines the final prediction combining the outputs from the two input modalities (RGB and spectral bands).
In more detail, we adopt **ResNet18** [1] as the backbone and the decoder network produces a tensor which yields the pixel-wise predictions in terms of a and b coordinates in the **CIE*Lab*** color space.



Once the encoder-decoder has been trained, we exploit only the encoder, adding a single final linear layer that maps bottleneck features to the classification output space.
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. To further take advantage of color information, we set up an ensemble model at inference time.

3. Baselines comparison

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

We compare our model with the networks presented in [4].
Results show:

- how our ensemble builds upon ResNet-18 outperforms over parametrized networks;
- 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.

4. Land-Cover Classification

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
Ens.	ImagNet+ImageNet	.649	.707	.749	.815	**.904**
Ens.	Color.+ImageNet	**.656**	**.751**	**.778**	**.823**	.896

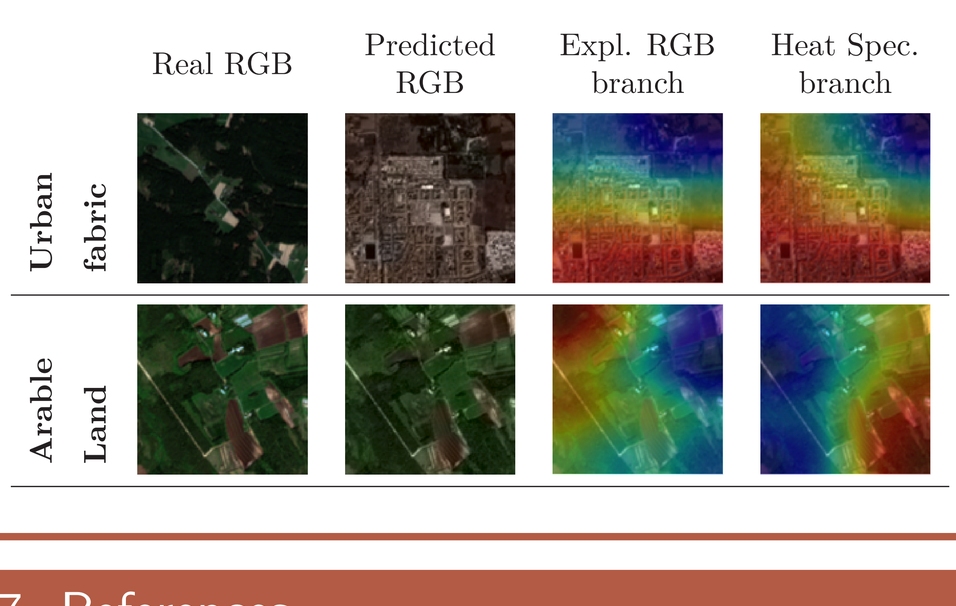
The initialization offered by colorization surpassing the other alternatives, especially in presence of scarce data, thus complying with the goals we have striven for in this work.
The ImageNet pre-training performs well for an RGB input; however, when dealing with the spectral domain, colorization is the sole that rewards the exploitation of spectral bands and justifies their usage in place of RGB.

5. West Nile Disease

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**

The table reports the results achieved on the West Nile Disease case study, framed in the Surveillance plan put in place by the Ministry of Health in Italy.
We provide a simple baseline (*i.e.* "random classifier") that computes predictions by randomly guessing from the class-prior distribution of the training set.
All the networks we trained exceed random guessing. Further, the ensemble model surpassing networks based on a single modality by a consistent margin.

6. Model Explanation



We take advantage of **GradCam** [2] to assess whether the two branches look for different properties within their inputs.
The third and fourth columns in the figure highlight the input regions that have been considered important for predicting the target category.
As one can see, the explanations provided visually diverge.

7. References

[1] H. Kaiming et al. Deep residual learning for image recognition. In *CVPR*, 2016.
[2] R. R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*, 2017.
[3] G. Sumbul et al. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS*, 2019.
[4] G. Sumbul et al. Bigearthnet deep learning models with a new class-nomenclature for remote sensing image understanding. *arXiv preprint arXiv:2001.06372*, 2020.