PoNet+: A Physical Optimization-based Network with Spectral Grouping for Spectral Recovery

Jiang He, Yi Xiao, Jiajun Xiao, Qiangqiang Yuan, Jie Li, Liangpei

March 30, 2022

1 Team details

- Team name: SGG_RS_Whu
- Team leader name: Jiang He
- Team leader address: School of Geodesy and Geomatics, Wuhan University, Hubei, 430079, China;
  Phone number: +86 16607114843; Email: jiang_he@whu.edu.cn;
- Yi Xiao(xy574475@gmail.com), Jiajun Xiao(wuhansunyat@gmail.com), Qiangqiang Yuan(yqiang86@gmail.com), Jie Li(jli89@sgg.whu.edu.cn), Liangpei Zhang(zlp62@whu.edu.cn)
- School of Geodesy and Geomatics, Wuhan University, China
- User names: hj_whu
- Best scoring entries of the team during development/validation phase: 0.260699857
- Link to the codes: https://github.com/JiangHe96/PoNet_plus
- Link to the restoration results: https://zenodo.org/record/6394053

2 Contribution details

- PoNet+: A Physical Optimization-based Network with Spectral Grouping for Spectral Recovery
- A physical optimization-based spectral recovery methods is unrolled into an end-to-end CNN as our previous work PoNet [1]. Besides, we employed the spectral grouping similar to HSRnet [2]. Detialed contributions are shown in below:

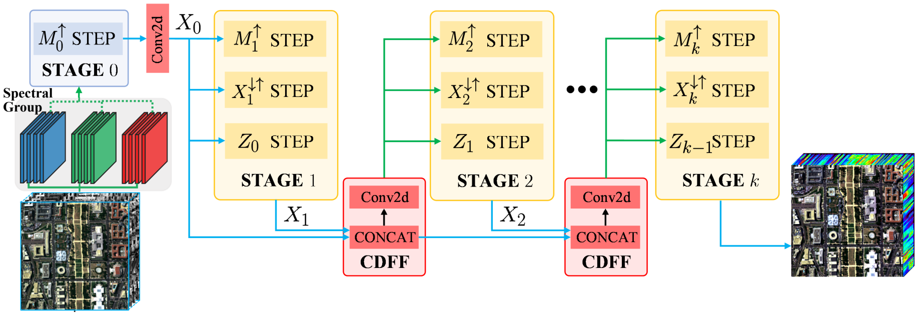

Figure 1: The framework of the proposed PoNet+

2.1 Physical Optimization Unrolling

Let $X \in \mathbb{R}^{W\times H\times C}$ represent the observed HSI, where C is the number of the spectral channels, and W and H are the width and height, respectively. $Y \in \mathbb{R}^{W\times H\times c}$ represents the observed multispectral image, where $c < C$ is the number of multispectral bands, specifically for RGB image, with $c = 3$. Varying in SRF, the sensors obtain different MS or HS data with different bands. A transformation matrix $\Phi \in \mathbb{R}^{c\times C}$can be used to describe the spectral degradation between MS and HS imaging as follows.

$$Y = \Phi X \tag{1}$$

The high-dimension HSIs can be approximately predicted by adopting some priors to a minimization problem to constrain the solution space as follows:

$$\hat{X} = \arg\min_X \|Y - \Phi X\|_2^2 + \lambda \mathcal{R}(X) \tag{2}$$

where λ is a trade-off parameter, and $\mathcal{R}(\cdot)$ is a regularization function. Employ the half-quadratic splitting method with a penalty parameter as μ and solve it by the gradient descent algorithm:

$$\hat{X}_{k+1} = (1 - \epsilon\mu)\, X_k - \epsilon X_k \Phi\Phi^T + \epsilon M_H \Phi^T + \epsilon\mu Z_k \tag{3}$$

$$\hat{Z}_k = Prox(X_k) = \arg\min_Z \|Z - X_k\|_2^2 + \frac{\lambda}{\mu}\mathcal{R}(Z) \tag{4}$$

where ϵ is the optimization stride. As for the Z-subproblem, proximal operators that impose prior knowledge can deal with it.

Unrolling the physical optimization method into CNN, the proposed PoNet+ is shown in Fig. 1.

2.2 Cross-Dimensional Channel Attention

In traditional physical optimization-based algorithms, hyperparameters need to be defined manually and adjust to the optimal through a large number of experiments. Furthermore, in spectral super-resolution, differential treatment should be performed for the hyperparameters of different channels due to the different radiation characteristics.

Pooling is a common operation used in traditional channel attention, which is popular for fast computation and no parameter requirement at the cost of high information loss. Furthermore, traditional channel attention weights the different channels of features separately ignoring the interaction between channels. There have been many works stated that building relationships between every two channels is much of importance. However, when the number of channels is large and attention mechanisms are frequently employed, the problem of computational burden should also be focused on.

Inspired by the above-mentioned points, we proposed a strategy named *Cross-Dimensional Channel Attention* (CDCA) employing 1D and 2D convolutional layers to manage the hyperparameter learning in this paper. 2D convolutional layers are used to extract pixel-by-pixel attention maps. On the other hand, 1D convolutional layers are employed to integrate attention maps for fast computational speed. Details of the proposed module are shown in Fig. 2.

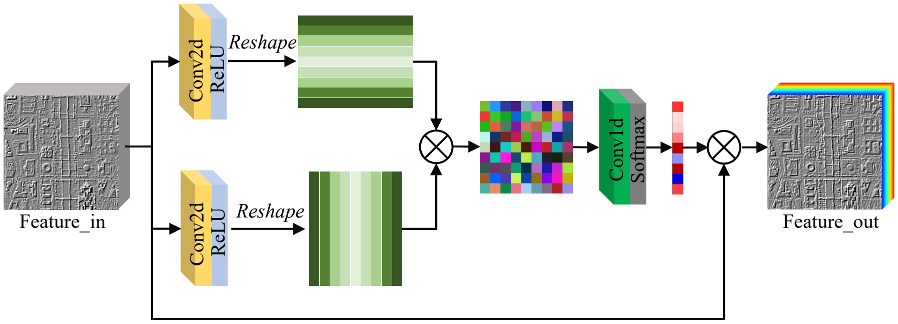

Figure 2: Cross-dimensional channel attention

We adopt two 2D convolutional layers with the kernel size of 1×1 to extract different spectral features $R, S \in \mathbb{R}^{W \times H \times C}$. Attention map $A \in \mathbb{R}^{C \times C}$ between any two channels can be calculated as follows:

$$a_{ij} = R_i S_j^T \tag{5}$$

where a_{ij} measures the attention between the i^{th} and j^{th} bands. R_i and S_j is the reshaped channel. To boost the computational speed, we employ a 1D convolutional layer with the kernel size of k to integrate channel-to-channel attention map A. Then, the final cross-dimensional channel attention-based hyperparameter $P \in \mathbb{R}^{1\times C}$ will be obtained after a softmax layer:

$$p_j = \frac{\exp\left(W^{\mathrm{1d}} * A_j + b^{\mathrm{1d}}\right)}{\sum_{j=1}^{C} \exp\left(W^{\mathrm{1d}} * A_j + b^{\mathrm{1d}}\right)} \tag{6}$$

where W^{1d} and b^{1d} mean the kernel weights and biases for the 1D convolutional layer, and $p_j \in P$ is the parameter for the j^{th} band. In this way, we build a learnable end-to-end CNN by unrolling the physical optimization algorithm, which keeps the advantages of deep learning and physical model-based algorithm.

2.3 Cross-Depth Feature Fusion

In deep learning-based algorithms, the depth makes much sense for the network effect, in other words, the deeper networks get the better results. However, the shallow features are also very important. In the proposed method, we get multiple updated results at different depths. To improve the model memory of shallow features, a strategy named *Cross-Depth Feature Fusion* (CDFF) is proposed as shown in Fig. 1.

Given a set of intermediate results $\{X_0, X_1,\ X_2, \cdots X_{k-1}\}$, PoNet+ firstly concatenates results at different depths:

$$F^C_{k-1} = Concat(X_0, X_1,\ X_2, \cdots X_{k-1}) \tag{7}$$

Then, a convolutional layer with ReLU is also employed to fuse cross-depth features to obtain the input for the next stage.

$$X^{In}_k = ReLU(W^F_{k-1} * F^C_{k-1} + b^F_{k-1}) \tag{8}$$

where X^{In}_k means the input feature for the k^{th} optimization stage. Acquiring various information from cross-depth features, X^{In}_k can represent more comprehensive spectral information from shallow and deep features, which is beneficial to the subsequent optimization.

2.4 Spectral Grouping

Spectral grouping is achieved by grouping bands with spectral relevance according to spectral response functions (SRFs). The spectral grouping is used to avoid reconstruction distortion caused by the excessive spectral difference between different channels. Nevertheless, it seems inevitable that there still will be some differences between bands in the same group. The proposed strategy ensures that intra-group bands reconstruction is determined by the same combination of multispectral channels. By roughly representing spectral relevance from the similarity of imaging according to spectral response functions, SRF-guided convolutional layers don't have to be adjusted for the same sensor, which improves the generalization of this module.

- Note that, PoNet is a universal network presented to address generalized spectral super-resolution, including classical spectral super-resolution (sSR), FusSR, and PansSR. So, we only employed the PoNet in sSR.
- Representative diagram of the method: As shown in Fig. 1.

3 Global Method Description

- Total method complexity: Parameter Number: 682.6K
- Training description: We directly read 900 images into ".h5" file with the original size (512×482). And in each iteration, we just input one image pairs. Training loss is ℓ_1 loss function.
- Testing description: The proposed method is an end-to-end network, we just gnerate the recovered hyperspectral images by feeding the test image into the trained model.
- We got a mean relative absolute error (MRAE) as 0.260699857 in validation phase. Besides, the MRAE in test phase is 0.3060293408 and root mean square error (RMSE) as 0.05071633486.
- The proposed method is a combination between physical optimization-unrolling networks and spectral grouping.
- The proposed solution is based on our previous works PoNet [1] and HSRnet[2].

4 Technical details

- Implementation details: We run our codes in Linux system with 32GB memory and a RTX A5000 GPU. Moreover, the codes are based on Pytorch.
- Human effort required for implementation, training and validation? No
- Training: 0.308751s/ite; 899ites/epoch; Convergence: 80 epochs.
  Whole Training: 6.17 hours.
- Runtime at test per image: 0.286305s
- The proposed method is light-weight and easy to be generalized to other sets.
- We thought the proposed model is light-weight and easy to be reproducted.

5 Other details

- We have no planned submission at NTIRE2022 workshop.
- General comments and impressions: The development of algorithms is extremely rapid.
- We expect there could be a new challenge about spectral recovery of satellite images.
- We suggest that spectral angle mapper (SAM) should also be employed to evaluate the spectral consistency between test results and target images.

References

[1] J. He, Q. Yuan, J. Li, and L. Zhang, "Ponet: A universal physical optimization-based spectral super-resolution network for arbitrary multispectral images," *Information Fusion*, vol. 80, pp. 205–225, 2022.

[2] J. He, J. Li, Q. Yuan, H. Shen, and L. Zhang, "Spectral response function-guided deep optimization-driven network for spectral super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.