Continuous Spectral Reconstruction from RGB Images via Implicit Neural Representation

Ruikang Xu¹  Mingde Yao¹  Chang Chen²  Lizhi Wang³  Zhiwei Xiong¹*

University of Science and Technology of China¹
Huawei Noah’s Ark Lab²
Beijing Institute of Technology³

Abstract

Existing methods for spectral reconstruction usually learn a discrete mapping from RGB images to a number of spectral bands. However, this modeling strategy ignores the continuous nature of spectral signature. In this paper, we propose Neural Spectral Reconstruction (NeSR) to lift this limitation, by introducing a novel continuous spectral representation. To this end, we embrace the concept of implicit function and implement a parameterized embodiment with a neural network. Specifically, we first adopt a backbone network to extract spatial features of RGB inputs. Based on it, we devise Spectral Profile Interpolation (SPI) module and Neural Attention Mapping (NAM) module to enrich deep features, where the spatial-spectral correlation is involved for a better representation. Then, we view the number of sampled spectral bands as the coordinate of continuous implicit function, so as to learn the projection from deep features to spectral intensities. Extensive experiments demonstrate the distinct advantage of NeSR in reconstruction accuracy over baseline methods. Moreover, NeSR extends the flexibility of spectral reconstruction by enabling an arbitrary number of spectral bands as the target output.

1. Introduction

Spectral image records more spectrum information than traditional RGB images, which have been proven useful in various vision-based applications, such as classification [18], edge-detection [39], object tracking [48], and segmentation [15]. To acquire spectral images, existing spectral imaging technology relies on either spatial or spectral scanning for capturing the spectral signature to a number of bands, which is of high complexity and time-consuming [9, 16, 44, 45, 50]. As an alternative, spectral reconstruction from RGB images is regarded as an attractive solution owing to the easy acquisition of RGB images [3, 31, 38, 57].

Existing methods [22, 32, 53, 55] for spectral reconstruction aim to learn a discrete mapping from an RGB image to a certain image, reconstructing a specific number of spectral bands from three bands, as the “blue dashes” in Figure 1. However, this representing scheme ignores the continuous nature of spectral signature. In the physical world, the spectral signature is of presented in a continuous form where the high-dimension correlation is naturally hidden in the continuous representation [8, 21]. To approximate the natural representation of the spectral signature, we reformulate the spectral reconstruction as the sampling process, where a number of spectral bands are sampled from the continuous representation, as the “red curve” in Figure 1.

Recent works adopt the concept of implicit neural representation for super-resolution [13] and 3D reconstruction [25, 35] tasks, to guarantee the continuous representation of signals with high fidelity. Inspired by this idea, we aim to learn the continuous representation of the spectral signature for spectral reconstruction and implement it by a continuous and differentiable function parameterized with a neural network [23, 26, 34]. However, it is non-trivial to learn the continuous representation for the spectral image, since the high-dimension correlation has not been explored in the continuous representation [13, 25], which is the key for spectral reconstruction.

This work aims to bridge this gap. We propose Neural Spectral Reconstruction (NeSR) to continuously represent the spectral signature. Specifically, we first take a backbone network as a feature encoder to extract the spatial feature from the RGB input. To enrich the spectral information of
the feature for continuous spectral representation, we propose a Spectral Profile Interpolation (SPI) module and a Neural Attention Mapping (NAM) module. The SPI module encodes the spatial-spectral correlation to the feature leveraging the vertical and the horizontal spectral profile interpolation. The NAM module further enriches the spatial-spectral information by the spatial-spectral-wise attention mechanism. With these two modules, the spatial-spectral correlation is involved in the deep feature for learning a better representation. To reconstruct the spectral image, we take the number of bands as the coordinate information and project the deep feature to the spectral image.

Benefiting from the continuous spectral representation as well as the modeling of spatial-spectral correlation, the advantage of NeSR is three-fold. First, NeSR improves the performance of the existing methods on spectral image reconstruction, in which it brings 11% accuracy improvement with only 2% parameter increase than the corresponding baseline on the NTIRE2020 dataset. Second, NeSR is a plug-and-play component, which can be incorporated into previous methods by replacing the output layer. Third, NeSR can not only improve the accuracy of spectral reconstruction but also effectively reconstruct spectral images with an arbitrary number of bands, which has the practical potential for various downstream applications.

The contributions are summarized as follows:

1. We propose NeSR to learn the continuous representation of the spectral image via implicit neural representation.
2. We propose the SPI module and the NAM module to enrich the information of the features from RGB inputs for continuous spectral representation by exploring the spatial-spectral correlation.
3. Extensive experiments demonstrate that NeSR outperforms the existing state-of-the-art methods in reconstructing spectral images with a specific number of bands.
4. NeSR firstly gives a solution to reconstruct spectral images with an arbitrary number of spectral bands while keeping high accuracy.

2. Related Works

2.1. Implicit Neural Representation

The implicit neural representation aims to model an object as a continuous and differentiable function that maps coordinates and the deep feature to the corresponding signal, which is parameterized by the deep neural network. Recent works demonstrate its potential for modeling the surfaces [10, 20, 35], the shapes [6, 24], and the appearance of the 3D objects [28, 29]. Mildenhall et al. introduce the implicit neural representation for synthesizing novel views of complex scenes using a sparse set of input views, named Neural Radiance Filed (NeRF) [25]. Compared with explicit 3D representations, implicit neural representation can capture subtle details. With the great success of NeRF [25], a large number of works extend the implicit neural representation to other applications, such as 3D-aware generalization [11, 40], pose estimation [37, 52], relighting [7, 36], and composition [27, 30].

To get a more general representation, recent works share an implicit space for different objects or scenes [13, 14, 34], instead of learning an independent implicit neural representation for each object. The implicit space can be generated with an auto-encoder architecture. For example, Sitzmann et al. [33] propose a meta-learning-based method for sharing the implicit space. Mescheder et al. [23] propose to generate a global latent space of given images as input, and use an occupancy function conditioning to perform the 3D reconstruction. Chen et al. [13] utilize local implicit image function for representing natural and complex images. To the best of our knowledge, we address the spectral reconstruction from RGB images by the implicit neural representation for the first time.

2.2. Spectral Image Reconstruction

Most existing spectral imaging systems rely on either spatial or spectral scanning [16, 44, 45]. These methods need to capture the spectral information of single points or bands separately, and then scan the whole scene to get a spectral image. It is difficult to measure information from scenes with moving content and hence is unsuitable for real-time operation. Reconstructing spectral images from RGB images is a relatively low-cost and convenient approach to acquiring spectral images. However, spectral reconstruction is a severely ill-posed problem, since much information is lost after integrating the spectral radiance into RGB values.

Many methods have been proposed for spectral reconstruction from RGB images. Early works leverage sparse coding methods to recover the lost spectral information from RGB images. Arad et al. [3] first exploit spectral prior from vast amounts of data to create a sparse dictionary, which facilitates the spectral reconstruction. Later, Aeschbacher et al. [1] further improve the performance of Arad’s method leveraging A+ framework [41] from super-resolution. Alternatively, Akhtar et al. [2] utilize Gaussian processes and clustering instead of dictionary learning. With the great success of convolution neural networks in computer vision tasks, a large number of methods have been proposed [38, 53, 57]. Xiong et al. [49] propose a unified deep learning framework for spectral reconstruction from both RGB and compressive measurements. Shi et al. [32] propose two improvements to HSCNN for boosting the performance called HSCNN+. One is to introduce a deep residual network named HSCNN-R, and the other is to replace the residual block with the dense block with a novel feature fusion scheme named HSCNN-D. Li et al. [22] propose an adaptive weighted attention network (AWAN) to explore the
camera spectral sensitivity prior to the reconstructing accuracy improvement. Zhao et al. [55] propose a hierarchical regression network (HRNet) for reconstructing spectral images from RGB images. However, these methods reconstruct the spectral images by the discrete mapping, which fixes the number of output spectral bands. In this work, we propose NeSR to learn a continuous representation for spectral reconstruction, which not only improves accuracy for reconstructing spectral images with a specific number of bands but also enables spectral reconstruction with an arbitrary number of spectral bands.

3. Neural Spectral Reconstruction

3.1. Overview

The overview of NeSR is shown in Figure 2 (a). Given an input image, we cascade a feature encoder to extract the spatial feature \( M_{in} \in \mathbb{R}^{\lambda_{in} \times H \times W} \) at first. Then, NeSR takes the deep feature \( M_{in} \) and the target number of spectral bands \( \lambda_{out} \) as inputs and reconstructs the spectral image with the target number of bands as the output, denoted as

\[
Y = \mathcal{F}(M_{in}, \lambda_{out}), \tag{1}
\]

where \( Y \) and \( \mathcal{F} \) stand for the reconstructed spectral image with the target number of bands and our NeSR, respectively. Since the target of the spectral reconstruction task is the reconstructed image with a specific number of bands, it is straightforward to take the number of spectral bands as coordinate information for representing the spectral image. Another common choice is representing the spectral image with normalized coordinates, however, we find that the performance of this scheme is inferior to our method (please refer to the supplemental material for more details).

Specifically, the spatial feature \( M_{in} \) has the spatial information but is unable to fully explore spectral information which is the key to the continuous representation of spectral images. To this end, we propose the SPI module and the NAM module. The SPI module is designed to explore the spatial-spectral correlation, which interpolates the feature to the target number of spectral bands leveraging the dedicated spectral profile operation. The NAM module is designed for global attention on spatial and spectral dimensions, which explores the similarity of different channels according to the spatial-spectral-wise attention mechanism. The proposed two modules generate a new deep feature on the target number of spectral bands \( M_{out}^* \in \mathbb{R}^{\lambda_{out} \times H \times W} \), where the spatial-spectral correlation is involved for learning a better continuous spectral representation.

After encoding the spatial-spectral correlation to the feature \( M_{out}^* \), each latent code of the feature is fed to a multilayer perceptron (MLP) to predict the corresponding spectral signature. Iterating over all the locations of the target spectral image, NeSR can reconstruct the spectral image with the desired number of spectral bands while keeping high fidelity.

3.2. Spectral Profile Interpolation

In this subsection, we introduce the SPI module, which explores the spatial-spectral correlation of the spectral image from a new perspective. The spatial-spectral correlation is an important characteristic of spectral representation [43, 46]. To explore this correlation, we propose the
concept of spectral profile inspired by other high-dimension image reconstruction tasks [47,58].

We first give a definition of Spectral Profile (SP) as follows, consider a spectral image as a 3D volume $I(h,w,λ)$, where $λ$ stands for the spectral dimension, the vertical SP $P_υ(h,λ)$ and the horizontal SP $P_h^*(w,λ)$ are the slices generated when $w = w'$ and $h = h'$, respectively. Since much information is lost after integrating the spectral radiance into RGB values, we explore the spatial-spectral correlation of the spectral image in the feature domain. The flow diagram of the SPI module is shown in Figure 2 (b).

In the SPI module, given the deep feature $M_{in} \in \mathbb{R}^{λ_{in} \times H \times W}$, we first convert it into the vertical and horizontal SPs to explore the correlations of the spatial and spectral dimension, denoted as $P_υ(h,λ) \in \mathbb{R}^{H \times λ_{in}}$ and $P_h^*(w,λ) \in \mathbb{R}^{W \times λ_{in}}$. Then, the upsampled vertical SPs $P_υ^* (h, λ) \in \mathbb{R}^{H \times λ_{out}}$ and the upsampled horizontal SPs $P_h^*(w,λ) \in \mathbb{R}^{W \times λ_{out}}$ can be generated by the bilinear interpolation, which can be denoted as

$$
P_υ^*(h, λ) = BI (P_υ(h, λ)),
$$
$$
P_h^*(w, λ) = BI (P_h^*(w, λ)),
$$

where $BI(\cdot)$ stands for the bilinear interpolation. After converting the above upsampled SPs back to the feature maps, we obtain the intermediate results $M_{out}^h, M_{out}^w \in \mathbb{R}^{λ_{out} \times H \times W}$. To generate the 3D deep feature for implicit neural representation, we fuse the features with the concatenation operation, and utilize the 3D convolution layers with LeakyReLU, which can be denoted as

$$
M_{out} = Conv_{3D} \left( [M_{out}^h, M_{out}^w] \right),
$$

where $M_{out} \in \mathbb{R}^{C \times λ_{out} \times H \times W}$ stands for the output feature of the SPI module, which is viewed as $λ_{out} \times H \times W$ latent codes evenly spread in the 3D domain. $Conv_{3D}(\cdot)$ and $[\cdot,\cdot]$ denote the 3D convolution layers with LeakyReLU and the concatenation operation, respectively.

As a result, the spatial-spectral correlation can be fully explored by the SPI module for leaining the spectral representation. The effectiveness of the SPI module is demonstrated in Section 4.4.

### 3.3. Neural Attention Mapping

In this subsection, we introduce the NAM module to enrich spatial-spectral information for continuous spectral representation by a new attention mechanism. Exploring the spatial-spectral information of the deep feature is important for spectral reconstruction. To this end, inspired by the success of Transformer in natural language processing [42] and computer vision [12, 17, 56], we propose the spatial-spectral-wise attention mechanism, which explores the interactions of different channels leveraging the spectral-spatial information. The overview of the NAM module is shown in Figure 2 (c).

Different from the standard Transformer block, which receives a 1D token embedding [17] as input, we take a 3D tensor as input. To handle the 3D tensor, we first reshape the 3D feature $M_{out}$ to the 2D token embedding, and then map the embedding with an MLP, denoted as

$$
v = MLP \left( Reshape (M_{out}) \right),
$$

where $MLP(\cdot)$ and $Reshape(\cdot)$ stand for the MLP and the Reshape operation, respectively. $v \in \mathbb{R}^{(λ_{out} \times H \times W) \times C}$ denotes the reshaped token embedding. The token embedding $v$ is fed to the spatial-spectral attention block to enrich the spatial-spectral information of the feature.

In computing attention, we first map the token embedding $v$ into key and memory embeddings for modeling the deep correspondences. Different from the self-attention mechanism, we compute the attention maps of different channels to capture the interactions among channels to enrich the information of the feature. Benefiting from that the token embedding $v$ compresses the spectral-spatial information to one dimension, the attention maps can fully explore the spectral-spatial information to compute the similarity among channels, denoted as

$$
Q, K, V = vW_q, vW_k, vW_v,
$$
$$
v^* = V \otimes \text{SoftMax} (Q^T \otimes K),
$$

where $Q, K, V \in \mathbb{R}^{(λ_{out} \times H \times W) \times C}$ stand for query, key and value to generate the attention, and $W_q, W_k, W_v$ stand for the corresponding weights, respectively. $\otimes$ and $T$ denote the batch-wise matrix multiplication and the batch-wise transposition, respectively. Finally, we map the output token embedding $v^*$ by an MLP, and then reshape the embedding to the 3D feature, denoted as

$$
M_{out}^* = \text{Reshape} \left( MLP (v^*) \right),
$$

where $M_{out}^* \in \mathbb{R}^{C \times λ_{out} \times H \times W}$ stands for the output of the NAM module.

Thanks to the attention mechanism, the proposed NAM module can enrich spatial-spectral information and we demonstrate the effectiveness of the module in Section 4.4.

### 3.4. Loss Function

In this paper, we adopt the Mean Relative Absolute Error (MRAE) as the loss function, which is defined as

$$
\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left( \left| Y^{(i)} - I^{(i)}_{GT} \right| / (I^{(i)}_{GT} + \varepsilon) \right),
$$

where $Y^{(i)}$ and $I^{(i)}_{GT}$ stand for the $i^{th}$ $(i = 1, ..., N)$ pixel of the reconstructed and ground truth spectral images, respectively. We set $\varepsilon = 1 \times 10^{-8}$ due to zero points in the ground truth spectral image.
Table 1. Quantitative comparison on the NTIRE2020, CAVE and NTIRE2018 datasets for 31-band spectral reconstruction. Red and blue indicate the best and the second best performance, respectively.

| Methods        | NTIRE2020 Clean | NTIRE2020 Real | CAVE | NTIRE2018 Clean | NTIRE2018 Real | Param (M) |
|----------------|-----------------|----------------|------|-----------------|----------------|-----------|
|                | MRAE RMSE       | MRAE RMSE      | MRAE RMSE | MRAE RMSE      | MRAE RMSE      | MRAE RMSE | MRAE RMSE | Param (M) |
| BI             | 0.16566 0.04551 | 0.17451 0.04307 | 5.74425 0.16886 | 0.12524 0.01941 | 0.15862 0.02375 | -         |
| HSCNN-R        | 0.04128 0.01515 | 0.07278 0.01892 | 0.19613 0.03553 | 0.01943 0.00389 | 0.03437 0.00621 | 1.20       |
| HSCNN-R + NeSR | 0.04012 0.01487 | 0.07191 0.01841 | 0.15353 0.03304 | 0.01611 0.00564 | 0.03134 0.00590 | 1.34       |
| HRNet          | 0.04007 0.01401 | 0.06756 0.01821 | 0.17219 0.02987 | 0.01521 0.00368 | 0.02985 0.00571 | 31.70      |
| HRNet + NeSR   | 0.03552 0.01314 | 0.06683 0.01717 | 0.17021 0.02921 | 0.01462 0.00355 | 0.02971 0.00569 | 31.88      |
| AWAN           | 0.03641 0.01385 | 0.06883 0.01711 | 0.19156 0.03752 | 0.01226 0.00555 | 0.03121 0.00577 | 28.59      |
| AWAN + NeSR    | 0.03216 0.00994 | 0.06646 0.01687 | 0.13124 0.02776 | 0.01112 0.00355 | 0.03019 0.00576 | 29.29      |

Figure 3. Visualization of the error maps of different methods on “ARAD_HS_0465” of the NTIRE2020 “Clean” dataset.

4. Experiments

4.1. Comparison to State-of-the-art Methods

To quantitatively evaluate the effectiveness of the proposed method, we first compare NeSR with state-of-the-art methods in reconstructing spectral images with a specific number of spectral bands.

Datasets. In this work, we select three datasets as the benchmark for training and evaluation, including NTIRE2020 [5], CAVE [51] and NTIRE2018 [4]. NTIRE2020 and NTIRE2018 are the benchmarks for the hyperspectral reconstruction challenges in NTIRE2020 and NTIRE2018, respectively. Both datasets consist of two tracks: “Clean” and “Real World”, in which each spectral image consists of 31 successive spectral bands ranging from 400 nm to 700 nm with a 10 nm increment. We abbreviate the “Real World” track as “Real”. CAVE is a 16-bit spectral image dataset, which contains 32 different images with a spatial resolution of 512 × 512 pixels and 31 successive spectral bands ranging from 400 nm to 700 nm with a 10 nm increment. For training, the image pairs of three datasets are randomly cropped to 64 × 64 and normalized to range [0, 1]. The training sets of the NTIRE2020 and NTIRE2018 datasets and the randomly picked 28 images from the CAVE dataset are used for training. We use the validation sets of the NTIRE2020 and NTIRE2018 datasets and the remaining four images from the CAVE dataset as the test datasets. We utilize the MRAE and Root Mean Square Error (RMSE) as metrics to quantitatively evaluate the reconstructed spectral images.

Implementation Details. Our proposed method is trained on spectral and RGB images on the three datasets separately. We employ a 4-layer MLP with ReLU activation to reconstruct spectral images at the end of NeSR, and the hidden dimensions are (128, 128, 256, 256). Adam optimizer is utilized with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initially set to $1 \times 10^{-4}$ and is later downscaled by a factor of 0.5 after every $2 \times 10^4$ iterations till $1 \times 10^5$ iterations. All the experiments in this paper are conducted in PyTorch 1.1 and trained on NVIDIA GeForce GTX1080Ti GPUs.

Comparison Methods. To verify the superiority of our method, we compare NeSR with different baselines, including one classical method (bilinear interpolation, BI) and three deep-learning-based methods (HSCNN-R [32], HRNet [55] and AWAN [22]). In NeSR, we set the feature ex-
Methods | 61 bands (5nm) | 31 bands (10nm) | 16 bands (20nm) | 11 bands (30nm) | 7 bands (50nm)
--- | --- | --- | --- | --- | ---
BI | 0.14842 0.02868 | 0.15573 0.02941 | 0.16611 0.03052 | 0.19012 0.03293 | 0.20214 0.03276
Sparse coding |  |  |  |  | 0.20984 0.03276
AWAN (-D) | 0.02182 0.00528 | 0.02298 0.00534 | 0.02313 0.00544 | 0.02412 0.00565 | 0.02535 0.00578
AWAN (-S) | 0.02182 0.00528 | 0.02212 0.00530 | 0.02241 0.00543 | 0.02347 0.00567 | 0.02484 0.00571
AWAN + NeSR | 0.01982 0.00483 | 0.02012 0.00502 | 0.02079 0.00523 | 0.02196 0.00541 | 0.02206 0.00544

Table 2. Quantitative comparison on the ICVL dataset for spectral reconstruction with arbitrary bands. Red and blue indicate the best and the second best performance, respectively.

Figure 4. Visualization of the error maps of different methods on “BGU_HS_00044” of the ICVL dataset with 31 bands.

4.2. Continuous Spectral Reconstruction

Previous methods of spectral reconstruction fix the number of output bands. Different from existing methods, by leveraging the continuous spectral representation, NeSR can reconstruct the spectral images with an arbitrary number of spectral bands and maintain high fidelity using a single model.

**Comparison Methods.** Since there are few available methods that could reconstruct spectral images with different bands using a single model, we set up three methods as baselines in this section (See Table 2). BI: We interpolate RGB images following the prior work [53, 57], which operates the interpolation function along the spectral dimension, resulting in the spectral image with the desired band number. Sparse coding: Sparse coding method [3] reconstructs spectral images leveraging a sparse dictionary of spectral signatures and their corresponding RGB projections, which can reconstruct spectral images with desired output band numbers. Deep-based: we select AWAN [22] as a representative deep-learning-based baseline for comprehensive
comparison. Because the current deep-learning model cannot reconstruct spectral images with an arbitrary number of bands, we separately train five models for each spectral band number (7/11/16/31/61 bands), denoted as AWAN(-S). We also design a two-step strategy, denoted as AWAN(-D), to reconstruct spectral images by first reconstructing the spectral images with a large number of spectral bands (e.g., 61 band) and then downsampling the spectral images to the target number of spectral bands (e.g., 31 band).

**Quantitative Evaluation.** We compare the quantitative results among NeSR and the aforementioned baselines. As shown in Table 2, reconstructing the spectral image with an arbitrary number of bands by BI shows poor performance, because RGB image is integrated on spectral dimension and direct interpolation between RGB channels cannot correspond to any spectral wavelength. The sparse coding method [3] also has a limited performance since it relies on the sparse dictionary but lacks the powerful representation of the deep neural networks. Our continuous spectral representation (AWAN+NeSR) can reconstruct spectral images into arbitrary bands with only one model, which surpasses the aforementioned methods. Compared with AWAN(-S), NeSR achieves 9.2%, 9.1%, 7.2%, 6.4%, and 11.2% decrease in terms of MRAE on the different numbers of spectral bands. The comparison results demonstrate that NeSR can effectively reconstruct spectral images with an arbitrary number of bands leveraging the continuous spectral representation.

**Qualitative Evaluation.** To give a visual comparison of different methods, we display the error maps of one representative scene from the ICVL dataset with the 31 spectral bands in Figure 4. From the visualization, we can see that NeSR can provide better recovery results and higher reconstruction fidelity than other methods. We also show the spectral curves of different methods on different spectral band numbers in Figure 5, which shows that our method provides a lower spectral error than other methods. The qualitative comparison demonstrates that NeSR can reconstruct spectral images with an arbitrary number of spectral bands while keeping high fidelity.

### 4.3. Towards a Flexible Learning Architecture

Since NeSR aims to learn the continuous representation for reconstructing spectral images, we apply it to reconstruct high accuracy and high spectral-resolution images from low spectral-resolution images, other than RGB images, to validate its wide application scenarios. As shown in Table 3, we reconstruct the spectral image with 16 bands to 31 and 61 bands by a single model. Experiments are

| Methods                | 16–31 bands | 16–61 bands |
|------------------------|-------------|-------------|
| BI                     | MRAE        | RMSE        |
|                        | 0.01521     | 0.00578     |
| AWAN                   | 0.01285     | 0.00271     |
| AWAN + NeSR            | 0.01239     | 0.00259     |

Table 3. Quantitative results of reconstructing spectral images from spectral images.
4.4. Ablation Studies

**Influence of the Components.** We validate the effectiveness of the SPI module and the NAM module by adding them to the basic network step by step. The results of the ablation experiment are shown in Table 4. We use AWAN cascaded with an MLP as the basic network and conduct an experiment on the NTIRE2020 “Clean” dataset. Since the deep feature lacks the spatial-spectral correlation, the baseline shows a limited reconstruction fidelity without the SPI and NAM modules. When inserting the SPI module into the baseline, the reconstruction error decreases from 0.03587 to 0.03339, and the NAM module also contributes to 0.00123 decrease in terms of MRAE. Moreover, from the error maps and the spectral curves in Figure 7, we can see that the reconstruction error decreases as inserting the SPI and NAM modules into the basic network. The quantitative and visual comparison results demonstrate that our proposed modules are effective to enrich the spectral information and explore spatial-spectral correlation for learning the continuous representation of spectral images.

**Influence of the Attention Mechanism.** To validate the effectiveness of our spectral-spatial-wise attention, we compare it with spectral-wise and spatial-wise attention [19,54]. We use AWAN as the feature encoder and maintain the SPI module and conduct an experiment on the NTIRE2020 “Clean” dataset. For each model, we only replace the attention mechanism of the NAM module. The results of the ablation experiment are shown in Table 5. Since the spatial-wise attention and spectral-wise cannot fully explore the spectral-spatial correlation, they show a limited reconstruction accuracy. Compared with the spatial-wise attention, our attention makes the reconstruction error decreases from 0.03312 to 0.03216 in term of MRAE. The experiment results demonstrate that our spectral-spatial-wise attention mechanism is effective to capture the interactions of different channels for boosting performance.

### Table 4. Ablation on the proposed modules.

| Methods             | MRAE   | RMSE   |
|---------------------|--------|--------|
| AWAN                | 0.03587| 0.01214|
| AWAN + SPI          | 0.03339| 0.01128|
| AWAN + SPI + NAM    | 0.03216| 0.00994|

### Table 5. Ablation on the attention mechanism.

| Methods             | MRAE   | RMSE   |
|---------------------|--------|--------|
| Spectral-Wise       | 0.03304| 0.01012|
| Spatial-Wise        | 0.03312| 0.01126|
| Spatial-Spectral-Wise| 0.03216| 0.00994|

5. Limitations

It is difficult to apply the proposed method to real-time applications, since generating a spectral image from a trained NeSR requires querying an MLP thousands of times (for thousands of pixels). But given the maturity of parallel computing and speeding up algorithms, we expect this to be an engineering challenge and not a fundamental limitation. Please see the supplementary material for comparisons and analyses of computational cost.

6. Conclusion

In this paper, we propose NeSR to learn the continuous representation for reconstructing the spectral image from the RGB image. NeSR takes the target number of spectral bands and the deep feature as inputs, and reconstructs the spectral image with the desired band numbers. We propose the SPI and the NAM to enrich the spectral information of the deep feature by exploring the spatial-spectral correlation. Extensive experiments have demonstrated that NeSR brings significant accuracy improvement over the corresponding baselines with little parameter increase. Moreover, NeSR can effectively reconstruct spectral images with an arbitrary number of spectral bands leveraging the continuous spectral representation, which is practical to suit various downstream applications while keeping high accuracy.
References

[1] Jonas Aeschbacher, Jiqing Wu, and Radu Timofte. In defense of shallow learned spectral reconstruction from rgb images. In ICCV, 2017. 2
[2] Naveed Akhtar and Ajmal Mian. Hyperspectral recovery from rgb images using gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(1):100–113, 2018. 2
[3] Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In ECCV, 2016. 1, 2, 6, 7
[4] Boaz Arad, Ohad Ben-Shahar, and Radu Timofte. Ntire 2018 challenge on spectral reconstruction from rgb images. In CVPRW, 2018. 5
[5] Boaz Arad, Radu Timofte, Ohad Ben-Shahar, Yi-Tun Lin, and Graham D Finlayson. Ntire 2020 challenge on spectral reconstruction from an rgb image. In CVPRW, 2020. 5
[6] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In CVPR, 2020. 2
[7] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In ICCV, 2021. 2
[8] Guangzhi Cao, Leonardo R Bacheva, and Charles A Bouman. The sparse matrix transform for covariance estimation and analysis of high dimensional signals. IEEE Transactions on Image Processing, 20(3):625–640, 2010. 1
[9] Xun Cao, Hao Du, Xin Tong, Qionghai Dai, and Stephen Lin. A prism-mask system for multispectral video acquisition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(12):2423–2435, 2011. 1
[10] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, JulianStraub,Steven Lovegrove, and RichardNewcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In ECCV, 2020. 2
[11] Eric R. Chan, Marco Monteiro, Pettr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In CVPR, 2021. 2
[12] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In CVPR, 2021. 4
[13] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In CVPR, 2021. 1, 2
[14] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In CVPR, 2019. 2
[15] Phuong D Dao, Kiran Mantripragada, Yuhong He, and Faisal Z Qureshi. Improving hyperspectral image segmentation by applying inverse noise weighting and outlier removal for optimal scale selection. ISPRS Journal of Photogrammetry and Remote Sensing, 171:348–366, 2021. 1
[16] Michael Descour and Eustace Dereniak. Computed-tomography imaging spectrometer: experimental calibration and reconstruction results. Applied optics, 34(22):4817–4826, 1995. 1, 2
[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2020. 4
[18] Jun Fu, Jing Liu, Jie Jiang, Yong Li, Yongjun Bao, and Hanqing Lu. Scene segmentation with dual relation-aware attention network. IEEE Transactions on Neural Networks and Learning Systems, 32(6):2547–2560, 2021. 1
[19] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In CVPR, 2019. 8
[20] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In CVPR, 2020. 2
[21] Oleg Kuybeda, David Malah, and Meir Barzohar. Rank estimation and redundancy reduction of high-dimensional noisy signals with preservation of rare vectors. IEEE Transactions on Signal Processing, 55(12):5579–5592, 2007. 1
[22] Jiaojiao Li, Chaoxiong Wu, Rui Song, Yunsong Li, and Fei Liu. Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from rgb images. In CVPRW, 2020. 1, 2, 5, 6
[23] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In CVPR, 2019. 1, 2
[24] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In CVPR, 2019. 2
[25] Ben Mildenhall, Pratl P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020. 1, 2
[26] Michael Niemeyer and Andreas Geiger. Giraffe: representing scenes as compositional generative neural feature fields. In CVPR, 2021. 1
[27] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In CVPR, 2021. 2
[28] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In CVPR, 2020. 2
[29] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In CVPR, 2019. 2
[30] Julian Ost, Fahim Mannan, Nils Thayer, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In CVPR, 2021. 2
[31] Antonio Robles-Kelly. Single image spectral reconstruction for multimedia applications. In ACM MM, 2015. 1
[32] Zhan Shi, Chang Chen, Zhiwei Xiong, Dong Liu, and Feng Wu. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In CVPRW, 2018. 1, 2, 5
