Hyperspectral Image Super-Resolution with RGB Image Super-Resolution as an Auxiliary Task

Ke Li[1] Dengxin Dai[2] Luc Van Gool[1,3]
[1]CVL, ETH Zurich, [2] MPI for Informatics, [3]PSI, KU Leuven
{ke.li,vangool}@vision.ee.ethz.ch, ddai@mpi-inf.mpg.de

Abstract

This work studies Hyperspectral image (HSI) super-resolution (SR). HSI SR is characterized by high-dimensional data and a limited amount of training examples. This raises challenges for training deep neural networks that are known to be data hungry. This work addresses this issue with two contributions. First, we observe that HSI SR and RGB image SR are correlated and develop a novel multi-tasking network to train them jointly so that the auxiliary task RGB image SR can provide additional supervision and regulate the network training. Second, we extend the network to a semi-supervised setting so that it can learn from datasets containing only low-resolution HSIs. With these contributions, our method is able to learn hyperspectral image super-resolution from heterogeneous datasets and lifts the requirement for having a large amount of high resolution (HR) HSI training samples. Extensive experiments on three standard datasets show that our method outperforms existing methods significantly and underpin the relevance of our contributions. Our code can be found at `https://github.com/kli8996/HSISR.git`.

1. Introduction

Hyperspectral imaging acquires images across many intervals of the electromagnetic spectrum. It has been applied to numerous areas such as medical diagnosis [37], food quality and safety control [23], remote sensing [22] and object detection [39]. All these applications benefit from analyzing the spectral information coming with HSIs. One obstacle in the way of further unleashing this potential is data acquisition. Acquiring HSIs of high spatial and high spectral resolution at a high frame rate is still a grand challenge. There is still no camera to achieve these three goals at the same time. Cameras for a compromise setting – high spectral but low spatial resolution – are quite common by now, though still expensive. As a result, increasing efforts have been made to advance HSI super-resolution (SR).

While numerous deep learning methods have been developed for improving the resolution of RGB images (RGBIs), methods for HSI SR are fewer. One of the main reasons is the lack of large-scale HSI datasets featuring high-resolution (HR) HSIs. As known, supervised deep learning methods need an enormous amount of training data. This situation, unfortunately, will not be improved in the foreseeable future due to the challenges hyperspectral imaging faces. In this work, we choose a different route and propose to learn HSI SR with an auxiliary task. We find that while it is difficult to collect HR HSIs, it is very easy to collect HR RGB images. It is thus appealing to have a HSI SR method which can learn from the two heterogeneous sources – RGB images and hyperspectral images – for hyperspectral ISR. Our method is designed for this aim.

Although the data distribution is not the same between RGBIs and HSIs, the two SR tasks do share some common goals in integrating information from neighboring spatial regions during the learning. We embrace this observation and formulate both tasks into the same learning framework such that the parameter distribution induced by the RGBI SR task can serve as an effective regularization for our HSI SR task. The challenge lies in the difference in spectral band numbers, e.g. three in RGBIs vs. e.g. 31 or 128 in HSIs. To tackle this problem, we decompose the HSI SR and RGBI SR into a commonly-shared spatial super-resolution task and two specific spectral refinement tasks, and propose a novel spatial-spectral neural network to solve them in a multi-tasking framework. This way, the spatial super-resolution network is shared between the two tasks to increase the total amount of supervision. It is in a similar spirit to other multi-tasking learning methods [46].

While the aforementioned contribution can yield state-of-the-art performance for HSI SR already, we extend the method further to learn from 'unlabeled' low-resolution HSI images as well. Semi-supervised learning (SSL) exploits unlabeled data to reduce over-fitting to the limited amount of labeled data [16, 32, 45, 49, 25]. While good progress has been made, the strategies are mainly designed for high-level recognition tasks. Their applicability to a

low-level dense regression task such as HSI SR has yet to be verified. In this work, we again leverage the success of RGB image (RGBI) SR and propose a cross-model consistency that favors functions giving consistent outputs between super-resolved RGBIs and super-resolved HSIs. Basically, we convert LR HSIs into LR RGB images and pass those through the trained RGBI SR network. In the meanwhile, we pass the LR HSIs through our HSI SR network to get the super-resolved HSIs and convert them to RGBIs with a standard camera response function. We enforce the consistency between the two versions of super-resolved RGBIs. This way, supervision is transferred from the better-trained RGB SR network to our HSI SR network via a second route.

To summarize, this work makes two contributions: 1) a multi-tasking HSI SR method to learn together with an auxiliary RGBI SR task, and 2) A SSL method to learn also from 'unlabeled' LR HSIs. With these contributions, our method sets the new state of the art for hyperspectral image super-resolution.

2. Related Work

Hyperspectral Image Super-Resolution. HSI SR can be grouped into three categories according to their settings: 1) HSI SR from only RGBIs or HR multispectral images (MSIs); 2) Single HSI SR from LR HSIs; and 3) HSI SR from both HR RGBIs (or MSIs) and LR HSIs of the same scene. Our method belongs to the second group.

HSI SR from only RGBIs is a highly ill-posed problem. However, it has gained great traction in recent years due to its simple setup and the well-organized workshop challenges [8]. Similar to other computer vision topics, the trend has shifted from 'conventional' methods such as radial basis functions [40] and sparse coding [7] to deep neural networks [21, 44, 8]. This trend highlights the need for bigger training datasets. Due to the challenge of reconstructing HSI from RGBIs that contain only three bands, there emerges research applying MSIs containing 3-8 bands to reconstruct HSIs [14].

Single image SR aims to model the relationship between the LR images and HR ones by learning from a collection of examples consisting of pairs of HR images and LR images. Single RGBI SR has achieved remarkable results in the last years. Since the first work of using neural networks for the task [18], progress has been made in making networks deeper and the connections denser [28, 57], using feature pyramids [31], employing GAN losses [34], and modeling real-world degradation effects [24]. As to single HSI SR, there has been great early work [3, 58] as well. However, that is also surpassed by deep learning methods. For instance, Yuan *et al*. [54] trained a single-band SR method on natural image datasets, and applied it to HSIs in a band-wise manner to explore spatial information. The spectral information is explored via matrix factorization afterwards. In order to explore both spatial and spectral correlation at the same time, methods based on 3D Convolutional Networks [38, 35] have been developed. Although 3D CNNs sound like a perfect solution, the computational complexity is very high. To alleviate this, Grouped Convolutions (GCs) with shared parameters have been recently used in [36, 27]. The backbone network of our method is also based on GCs.

Fusion-based methods use HR RGBIs (or MSIs) of the same scene as references to improve the spatial resolution of the LR HSIs [11, 53, 51]. This stream of methods have received more research attention than the former two. Many learning techniques have been applied to this data fusion task including Bayesian inference [5, 6, 56], matrix factorization [33, 17], sparse representation [4, 19], and deep neural networks [41, 50]. The common goal of these methods is to learn to propagate the detailed information in the HR RGBIs (or MSIs) to the target HSIs and fuse them with the fundamental spectral information from LR HSIs. Despite the plethora of fusion algorithms developed, they all assume that the LR HSIs and the HR RGBIs (or MSIs) are very well co-registered [27]. This data registration is a challenge on its own and registration errors will lead to degraded SR results [13, 59].

Learning with Auxiliary Tasks. It is quite a common practice to borrow additional supervision from related auxiliary tasks, when there is insufficient data to learn a task. The common strategy is to learn all the tasks together so that the auxiliary tasks can regularize the optimization. There are normally two assumptions: (1) we only care about the performance of the main task and (2) the supervision for the auxiliary tasks is easier to obtain than that of the main task. Previous work has employed various kinds of self-supervised methods as auxiliary tasks for the main supervised task in a semi-supervised setting [30, 10, 42]. For instance, generative approaches have been explored in [30] and predicting the orientation of image patches is used in [10]. Another related setting is multi-task learning (MTL) [47]. In MTL, the goal is to reach high performance on multiple tasks simultaneously, so all tasks are main tasks and all tasks are auxiliary tasks. While the goal is different, many strategies in MTL such as parameter sharing [9], task consistency [55], and loss balance [15] are useful for learning with auxiliary tasks.

3. Approach

HSIs provide tens of narrow bands and RGB images have three bands. In order to let them share a large part of the overall network, we decouple both the HSI SR task and the RGBI SR task into a spatial super-resolution task and a spectral refinement task. The spatial super-resolution task is designed to enhance the spatial resolution of a general single-channel image, regardless of the spectral frequency

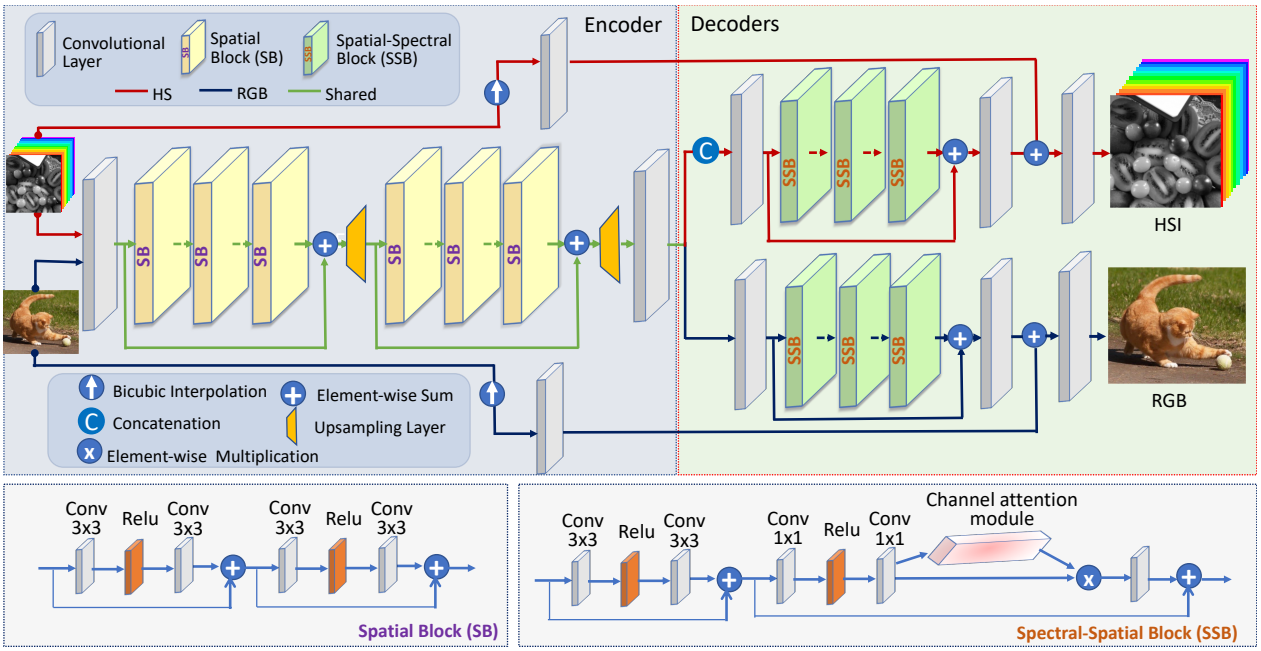


Figure 1: The architecture of our network consisting of a shared encoder and two specific decoders for the two SR tasks.

of that band. This way, it can be used and shared by both of the SR tasks. The spectral refinement networks are task specific – one to refine the spectral signals of the three channels of RGB images and the other to refine the spectral signals of tens of bands for HSIs.

In this work, we assume that the relationships between low/high-resolution HSIs and low/high-resolution RGBIs are correlated, so they should be trained together so that RGBI SR can provide additional supervision for HSI SR. This way, the HSI SR method can enjoy training samples of a much more diverse set of scenes especially those that cannot be captured well by current hyperspectral imaging devices such as moving objects. In order to share the spectral super-resolution network by all single bands from the two tasks, we use a grouped convolutional network with group size of 1. The grouping strategy is to divide input HSIs and RGB images into single bands. The architecture of the network is shown in Fig. 1.

3.1. HSI SR with an Auxiliary RGBI SR Task

Given two SR tasks $\mathcal{T}_{\text{HS}}$ and $\mathcal{T}_{\text{RGB}}$, we aim to help improve the learning of a model for $\mathcal{T}_{\text{HS}}$ by using the knowledge contained in $\mathcal{T}_{\text{RGB}}$. In the supervised setting, each task is accompanied by a training dataset consisting of N training samples, i.e., $\mathcal{D}_{\text{HS}} = \{\mathbf{x}_{\text{HS}}^i, \mathbf{X}_{\text{HS}}^i\}_{i=1}^{N_{\text{HS}}}$ and $\mathcal{D}_{\text{RGB}} = \{\mathbf{x}_{\text{RGB}}^i, \mathbf{X}_{\text{RGB}}^i\}_{i=1}^{N_{\text{RGB}}}$, where $\mathbf{x}_{\text{HS}} \in \mathbb{R}^{h_1 \times w_1 \times C}$, $\mathbf{X}_{\text{HS}} \in \mathbb{R}^{H_1 \times W_1 \times C}$, $\mathbf{x}_{\text{RGB}} \in \mathbb{R}^{h_2 \times w_2 \times Z}$, and $\mathbf{X}_{\text{RGB}} \in \mathbb{R}^{H_2 \times W_2 \times Z}$. We denote low-resolution (LR) images by $\mathbf{x}$, high-resolution (HR) images by $\mathbf{X}$, the number of bands of HSIs by C, the number of bands in RGB images by Z (3 here), and the size of the images by h, w, H and W. Given a scaling factor τ, we have $H_i = \tau h_i$ and $W_i = \tau w_i$ for both tasks.

The goal is to train a neural network Φ_{HS} to predict the HR HSI for a given LR HSI: $\mathbf{X}_{\text{HS}} = \Phi_{\text{HS}}(\mathbf{x}_{\text{HS}})$. Different from previous methods, which have a single network for the whole task, our method consists of three blocks: an encoder which is shared by the two SR tasks, and two task-specific decoders to output the final outputs. More specifically, $\Phi_{\text{HS}} = (\Phi^{\text{En}}, \Phi_{\text{HS}}^{\text{De}})$ and $\Phi_{\text{RGB}} = (\Phi^{\text{En}}, \Phi_{\text{RGB}}^{\text{De}})$.

In order to share the same encoder between the two SR tasks to enhance the spatial resolution of all single-band images, we divide $\mathbf{x}_{\text{HS}}$ into C single bands and $\mathbf{x}_{\text{RGB}}$ into 3 single bands. For both tasks, the encoder network Φ^{En} takes one low-resolution single-band image as input and generates one high-resolution single-band image as output, regardless of the spectral frequency of that band. The outputs of all the bands of $\mathbf{x}_{\text{HS}}$ are then concatenated according to their original spectral band position to assemble a high-resolution HSI $\bar{\mathbf{X}}_{\text{HS}} \in \mathbb{R}^{H_1 \times W_1 \times C}$. Similarly, we can assemble a high-resolution RGB image $\bar{\mathbf{X}}_{\text{RGB}} \in \mathbb{R}^{H_1 \times W_1 \times 3}$. There are two upsampling layers to upscale the size of the input to the desired size in a progressive manner. This progressive upsampling has proven useful for both RGBI SR [31] and HSI SR [27]. The reconstructed $\bar{\mathbf{X}}_{\text{HS}}$ is then fed into the decoder network $\Phi_{\text{HS}}^{\text{De}}$ as a whole for spectral refinement in order to generate the final output $\hat{\mathbf{X}}_{\text{HS}}$, which is then compared to the ground truth $\mathbf{X}_{\text{HS}}$ to compute the

loss for HSI SR. Likewise, the spatially enhanced $\bar{\mathbf{X}}_{\text{RGB}}$ is fed into the decoder network $\Phi^{\text{De}}_{\text{RGB}}$ to generate the final HR RGB estimate $\hat{\mathbf{X}}_{\text{RGB}}$, which is then compared to the ground truth $\mathbf{X}_{\text{RGB}}$ to compute the loss for RGBI SR. For the refinement decoders, all the bands are fed directly to learn both short-range and long-range spectral correlations to refine the results.

In order to have a modular design, the three subnetworks have the same basic architecture. The encoder network for spatial super-resolution is composed of a sequence of Spatial Block (SB) modules. The SB module has two identical basic cells connected in a sequence, each consists of one 3×3 convolutional layer, followed by a Relu and another 3×3 convolutional layer. There is also skip connection for each of this basic cell. Please see the bottom-left panel of Fig. 1 for its structure. The two decoders designed for spectral refinement of the two SR tasks have the same architecture and each consist of a sequence of three spectral-spatial block (SSB) modules. The SSB module was proposed in [27] as a basic building block for their HSI SR network. Each SSB has a Spatial Residual Module and a Spectral Attention Residual Module. Two Convolutional layers (the first one followed by a ReLu layer) with 3x3 filters are used in the Spatial Residual Module to capture spatial correlations. Two Convolutional layers (the first one again followed by a Relu layer) with 1x1 filters are used in the Spectral Attention Residual Module to capture spectral correlations. Please refer to the bottom-right panel of Fig. 1 for the architecture of the SSB module.

We construct the whole network with standard Convolutional Layers, SBs, SSBs, Upsampling Layers and Concatenation Operations. There are also skip connections at multiple scales to facilitate the information flow. The input LR images are also scaled to the desired size via Bicubic Interpolation and fused with the network output for residual learning. The complete network is shown in Fig. 1. We employ the PixelShuffle [43] operator for the upsampling layer. Given a scaling factor τ, the first upsampling layer upscales the features $\tau/2$ times and the second one handles the remaining $\times 2$ factor. The internal features of all SBB modules are limited to 256 in this work. The filter size of all Convolutional Layers, except for those 1×1 filters in the Spectral Attention Residual Module of SBBs, are set to 3×3.

3.2. Semi-Supervised HSI SR

While training with auxiliary RGB SR task can greatly improve the performance, it is still interesting to investigate whether the method can also learn from an additional collection of low-resolution hyperspectral images. This setting is interesting because capturing low-resolution hyperspectral images is much easier than capturing high-resolution hyperspectral images. This is especially true as modern snapshot HS cameras that captures LR HSIs at high frame rate are becoming more and more accessible. This means that methods that can learn further with low-resolution hyperspectral images are practically very useful. In the literature, there has been a diverse sets of methods developed for semi-supervised learning (SSL) based on techniques such as entropy minimization and pseudo-labels generation. However, they are mostly designed for high-level recognition tasks and cannot be applied to HSI SR directly.

In this work, we propose a new SSL method specifically for HSI SR. For this purpose, we again leverage the fact that RGBI SR is a better-addressed problem, given that it has a large amount of training data and it predicts only three channels. The method works as follows: given an image $\mathbf{x}_{\text{HS}}$, we convert it to an RGB image $\tilde{\mathbf{x}}_{\text{RGB}}$ with the camera response function of a standard RGB camera:

$$\mathbf{f} : \tilde{\mathbf{x}}^{(i,j)}_{\text{RGB}} = \mathbf{f} * \mathbf{x}^{(i,j)}_{\text{HS}}, \tag{1}$$

where $*$ is a convolution operation. The operation is to integrate the spectra signatures into R, G, and B channels with the response function of a standard camera. Note that this conversion is widely used in the literature of spectral image super-resoluton [20]. The response function of Canon 1D Mark 3 [26] is used in this work but the method works with the response functions of other cameras.

The original HSI $\mathbf{x}_{\text{HS}}$ and the converted RGB image $\tilde{\mathbf{x}}_{\text{RGB}}$ are then fed into the HSI SR network Φ_{HS} and the RGBI SR network Φ_{RGB}, respectively, to generate the super-resolved results:

$$\hat{\mathbf{X}}_{\text{HS}} = \Phi_{\text{HS}}(\mathbf{x}_{\text{HS}}), \tag{2}$$

and

$$\hat{\tilde{\mathbf{X}}}_{\text{RGB}} = \Phi_{\text{RGB}}(\tilde{\mathbf{x}}_{\text{RGB}}). \tag{3}$$

$\hat{\mathbf{X}}_{\text{HS}}$ is then converted to an RGB image by using the same camera response function:

$$\hat{\tilde{\mathbf{X}}}^{(i,j)}_{\text{RGB}} = \mathbf{f} * \hat{\mathbf{X}}^{(i,j)}_{\text{HS}}. \tag{4}$$

Finally, a consistency loss $L_{\text{ssl}}(\hat{\tilde{\mathbf{X}}}_{\text{RGB}}, \tilde{\hat{\mathbf{X}}}_{\text{RGB}})$ is computed between the two HR RGB results. This consistency makes a good use of low-resolution HSIs and high-resolutoin RGB images. It transfers supervision from the RGB side to the HSI side. The diagram of this method is shown in Fig. 2.

3.3. Loss Function

The overall loss for our SR tasks is:

$$\begin{aligned}\mathcal{L}^{\text{Total}} = \mathcal{L}^{\text{HS}}(\mathbf{X}_{\text{HS}}, \hat{\mathbf{X}}_{\text{HS}}) + \mathcal{L}^{\text{RGB}}(\mathbf{X}_{\text{RGB}}, \hat{\mathbf{X}}_{\text{RGB}}) \\ + \mathcal{L}^{\text{SSL}}(\hat{\tilde{X}}_{\text{RGB}}, \tilde{\hat{\mathbf{X}}}_{\text{RGB}}).\end{aligned} \tag{5}$$

The main loss is augmented by the two auxiliary losses which are optional but highly beneficial.

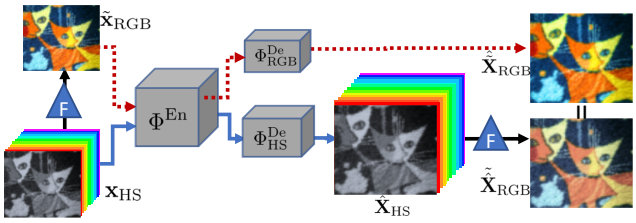


Figure 2: The pipeline of our semi-supervised learning.

In order to capture both spatial and spectral correlation of the SR results, we follow [27] and combine the L1 loss and the spatial-spectral total variation (SSTV) loss [1]. SSTV is used to encourage smooth results in both spatial domain and spectral domain and it is defined as:

$$\mathcal{L}_{\mathrm{SSTV}} = \frac{1}{N}\sum_{n=1}^{N}(||\nabla_{\mathrm{h}}\hat{\mathbf{X}}^n||_1+||\nabla_{\mathrm{w}}\hat{\mathbf{X}}^n||_1+||\nabla_{\mathrm{c}}\hat{\mathbf{X}}^n||_1), \tag{6}$$

where ∇_{h}, ∇_{w}, and ∇_{c} compute gradient along the horizontal, vertical and spectral directions, resp. The loss is:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{\mathrm{SSTV}}. \tag{7}$$

A joint training with all losses together works well in principle by stacking multiple types of data samples in a single mini-batch. However, that will heavily limits the size of the training data for each loss. In this work, we adopt an alternating training strategy; that is to train with each of the three losses in turn in every iteration. In our implementation, the weights for all losses are set to 1. The contributions of different terms are balanced or controlled by altering the number of mini-batches for that loss in each iteration. The influence of these numbers are studied in Sec. 4.2.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our method on three public datasets: CAVE dataset [52], Harvard dataest [12], and NTIRE 2020 dataset [8]. Images in CAVE and NTIRE 2020 dataset have 31 bands ranging from 400 nm to 700 nm at a step of 10 nm. Images in Harvard dataset contain 31 bands as well but range from 420 nm to 720 nm. The CAVE dataset contains 32 images of 512 x 512 pixels. We use 20 images for training and 10 images for testing. We evaluated in a supervised setting and a semi-supervised setting. For our semi-supervised setting, only 5 high-resolution training images are used and for the remaining 15 images, only their downsampled low-resolution images are available. For the Harvard dataset, there are 50 images in total. We use 40 for training and 10 for test. For the semi-supervised setting, 6 high-resolution images and the down-sampled low-resolution images of the remaining 34 images are used for training. For NTIRE 2020, there are 480 images. We use 400 images for training and 80 images for test. For the semi-supervised case, we further split the 400 images into 100 and 300; for the former, their high-resolution images can be used and for the latter only the downsampled low-resolution images are available. For the auxiliary RGBI SR task, we adopt the DIV2K Dataset [2]. Because the resolution of DIV2K is much higher than our HSIs, we first downsample them by a factor of $\times 2$ and take these downsampled images as our HR RGB images. After cropping, it leads to $137,430$ image patches of 64×64 pixels. This is about 34, 10, and 6 times larger than CAVE, Harvard and NTIRE datasets, respectively.

Methods. We compare the proposed method to four state-of-the-art HSI SR methods: GDRRN [36], 3DFCNN [38], SSPSR [27], and MCNet [35]. We use the same training data for all methods and use the default training settings given by the authors of these methods. Bicubic interpolation is also introduced as a baseline.

Evaluation Metrics. We follow the literature and evaluate the performance of all methods under three standard metrics [27]. They are root mean squared error (RMSE), erreur relative globale adimensionnelle de synthese (ERGAS) [48], and peak signal-to-noise ratio (PSNR). For PSNR of the reconstructed HSIs, their mean values of all spectral bands are reported as MPSNR. ERGAS are widely used in HSI fusion task.

Parameters. In this work, we focus on scaling factor $\times 4$ and $\times 8$. For the case of $\times 4$, we crop the images into patches of 64×64 pixels without overlapping to collect the training data. For $\times 8$, we use patches of 128×128 pixels. Those patches are then downsampled via Bicubic interpolation to obtain the corresponding LR HSI patches. The choice of value for other parameters are studied in Sec. 4.2.

Training Details. We use ADAM optimizer [29] and train all variants of our method for 20 epoches. This is a small number compared to the ones used by the comparison methods. For instance, GDRRN [36] trains for 30 epoches, 3DFCNN [38] trains for 200 epoches, SSPSR [27] for 40 epoches, and MCNet [35] for 200 epoches. We find that 20 epoches are sufficient to give good results for our method, and believe a larger number probably can further push the numbers up. The initial learning rate of all our methods is set to 10^{-4} and is reduced by a factor of 0.3 after every 3 epoches. As to the batch size, 16 is used for all experiments except for the case when the SSL loss is added. For that 8 is used due to the limit of GPU memory.

4.2. Ablation Study

We analyze the parameter choices of our method in this section. Experiments are conducted on the CAVE dataset in the semi-supervised setting.

Amount of RGB data. The number of mini-batches for

#(Mini-Batches)	0	1	2	3	4	5	6	8	10
RMSE ↓	0.01451	0.01357	0.01329	0.01309	0.01308	0.01305	0.01315	0.01315	0.01317

Table 1: Performance as a function of the number of mini-batches for RGBI SR loss.

Methods	Components		CAVE			Harvard			NTIRE		
	RGBSR	SSL	RMSE ↓	MPSNR ↑	ERGAS ↓	RMSE ↓	MPSNR↑	ERGAS ↓	RMSE ↓	MPSNR↑	ERGAS ↓
Ours			0.0144	40.8385	4.0345	0.0146	40.4666	3.1712	0.0154	38.3149	2.2069
Ours	✓		0.0118	42.3575	3.0128	0.0134	40.7579	3.0769	0.0150	38.7229	2.1189
Ours	✓	✓	**0.0114**	**42.7645**	**3.3346**	**0.0132**	**40.9317**	**3.0128**	**0.0150**	**38.9642**	**2.065**
Bicubic	-	-	0.0185	38.7380	5.2719	0.0167	38.8975	3.8069	0.0235	34.7401	3.1901
GDRRN [36]	-	-	0.0246	36.2775	7.0043	0.0160	38.6953	4.3031	0.0197	36.0793	2.8175
3DFCNN [38]	-	-	0.0173	38.3928	6.7055	0.0157	39.3441	3.6172	0.0208	35.6630	2.8246
SSPSR [27]	-	-	0.0144	40.9131	4.0406	0.0142	40.3209	3.2274	0.01636	38.0740	2.2539
MCNet [35]	-	-	0.0146	40.7385	4.1659	0.01468	40.1873	3.26059	0.0168	38.0248	2.2834

Table 2: Results of all methods on the CAVE, Harvard, and NTIRE datasets in the semi-supervised setting for the $\times 4$ case.

loss $\mathcal{L}^{\text{HS}}(\mathbf{X}_{\text{HS}}, \hat{\mathbf{X}}_{\text{HS}})$ in each iteration is fixed to 1. The number of mini-batches for loss $\mathcal{L}^{\text{SSL}}(\tilde{\hat{X}}_{\text{RGB}}, \tilde{\hat{\mathbf{X}}}_{\text{RGB}})$ is set to 3. This is decided by the ratio of the size of low-resolution hyperspectral dataset to the size of the high-resolution training set. We have studied the influence of the amount of used RGB images on the performance, *i.e*. the data used for loss $\mathcal{L}^{\text{RGB}}(\mathbf{X}_{\text{RGB}}, \hat{\mathbf{X}}_{\text{RGB}})$. We evaluate over a large range of values. The results are shown in Table 1. The performance increases first with the number of mini-batches and then decreases with it. This is because at the beginning as the amount of RGB data increases, more supervision is added to the system and the spatial super-resolution network is better trained. However, when too much RGB data (supervision) is introduced, the network is more trained for the auxiliary task RGBI SR than for the primary task HSI SR, thus the performance of the primary task will drop. We fix the number of batches for RGB data to 3 as it is a good trade-off between performance and computational time.

4.3. Main Results

We first present the results in the semi-supervised setting. The results of all competing methods and all variants of our method on the CAVE, Harvard, and NTRIE dataset are shown in Table 2 and in Table 3. The results in this table and other results in supplemental material show that our method outperforms all other state-of-the-art methods significantly and consistently over all datasets and under all evaluation metrics.

The good performance of our base model is mainly due to the use of decoupled spatial-spectral networks for this task. Due to the design, each sub-network can just focus on one task: spatial enhancement or spectral refinement. The network of [38] is quite shallow, probably because 3D convolution based methods are computationally heavy in general. We believe that this is the reason why their method does not give top results. When compared to the very recent method MCNet [35], our base model also performs better in almost all cases.

The proposed contributions, namely training with the auxiliary task RGBI SR and the semi-supervised learning method based on cross-model consistency, both contribute significantly to the final results. Learning with the auxiliary task RGBI SR can work on its own. The SSL component needs to be used together with the auxiliary task RGBI SR. The results show that our SSL method can provide further improvement on top of the auxiliary RGBI SR method, and when the two components are combined together, we get the best performance.

When more supervision is given such as in the fully-supervised setting, the conclusions we drawn in the semi-supervised setting hold as shown in Table 4. The results show that the proposed components are very effective and can be applied to situations with varying amount of HR HSIs. We show visual results of our method and other competing methods in Fig. 3 and Fig.4. The figure shows that our method generate few errors than all competing methods.

4.4. Discussion

The superior performance shows that our method is able to learn from heterogeneous datasets rather than only from pairs of low-resolution and high-resolution HR HSIs. This greatly increases the amount of training data that can be used for HSI SR and can also include training samples for scenes such as moving objects that cannot be captured easily with the current hyperspectal imaging devices. We would like to point out that while the general concepts of learning with auxiliary tasks and semi-supervised learning are well known, the challenge and novelty lie in defin-

Methods	Components RGBSR	SSL	CAVE RMSE ↓	CAVE MPSNR ↑	CAVE ERGAS ↓	Harvard RMSE ↓	Harvard MPSNR↑	Harvard ERGAS ↓	NTIRE RMSE ↓	NTIRE MPSNR↑	NTIRE ERGAS ↓
Ours			0.0241	35.8976	7.1154	0.0221	36.6527	4.8522	0.0232	32.8287	4.0434
Ours	✓		0.0215	37.1387	6.1442	0.0205	37.1859	4.5575	0.0269	33.3306	3.8548
Ours	✓	✓	**0.0206**	**37.3532**	**6.0027**	**0.0201**	**37.3546**	**4.5448**	**0.0263**	**33.4557**	**3.8437**
Bicubic	-	-	0.0304	34.2221	8.4350	0.0249	35.7409	5.4772	0.0396	29.9589	5.4594
GDRRN [36]	-	-	0.0347	32.9363	9.8554	0.0238	35.6441	5.7287	0.0359	30.6723	5.1265
3DFCNN [38]	-	-	0.0292	32.9024	16.7265	0.0237	36.0551	5.2192	0.3857	9.1753	6.1624
SSPSR [27]	-	-	0.0248	35.8896	7.0394	0.0228	36.4563	4.9978	0.0326	31.7896	4.4952
MCNet [35]	-	-	0.0280	34.3116	10.2985	0.0234	36.3921	5.0572	0.0327	31.9629	4.4169

Table 3: Results of all methods on the CAVE, Harvard, and NTIRE datasets in the semi-supervised setting for the $\times 8$ case.

(a) Ground Truth & Metrics (b) Bicubic Interpolation (c) 3DFCNN [38] (d) SSPSR [27] (e) Ours

Figure 3: Exemplar results of our method and two competing methods trained in the semi-supervised setting on the CAVE dataset: top row for the super-resolved results and bottom row for the error maps.

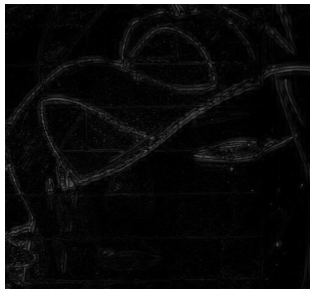
Methods	Components RGBI SR	CAVE RMSE ↓	CAVE MPSNR ↑	CAVE ERGAS ↓	Harvard RMSE ↓	Harvard MPSNR ↑	Harvard ERGAS ↓
Ours		0.0118	42.3836	3.45912	0.01356	40.91025	3.0104
Ours	✓	**0.0105**	**43.3242**	**3.1182**	**0.0128**	**41.0589**	**2.9649**
GDRRN [36]	-	0.0162	39.7470	4.5268	0.0148	39.6275	3.6793
3DFCNN [38]	-	0.0158	39.2178	5.4179	0.0151	39.6627	3.4773
SSPSR [27]	-	0.0124	42.1378	3.5514	0.0135	40.8149	3.0500
MCNet [35]	-	0.0124	42.2597	3.5624	0.0140	40.5922	3.1052

Table 4: Results of all methods on the CAVE and Harvard datasets in the fully-supervised setting for the $\times 4$ case.

ing proper auxiliary tasks for a new primary task developing effective SSL methods for a new task. This learning with auxiliary RGBI SR becomes possible because of our special network design – decoupling the spatial super-resolution task and the spectral refinement task and use two sub-networks for them. It is also known that many seemingly-related auxiliary tasks yield no improvement or even degrade the performance of the main task [42]. Therefore, developing effective methods for learning with auxiliary task is not always straightforward.

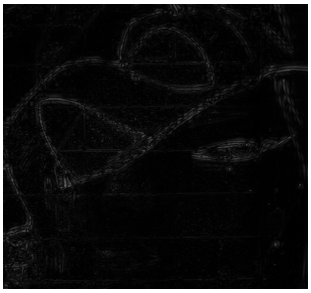
(a) Ground Truth

(b) Bicubic Interpolation



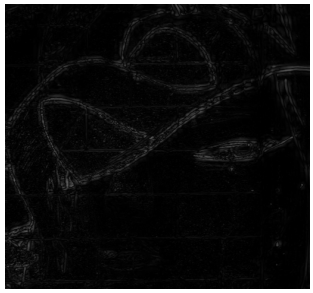
(c) Error of Bicubic

(d) GDRRN [36]



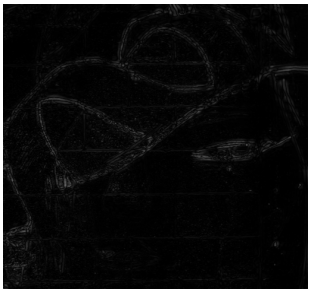
(e) Error of GDRRN

(f) 3DFCNN [38]



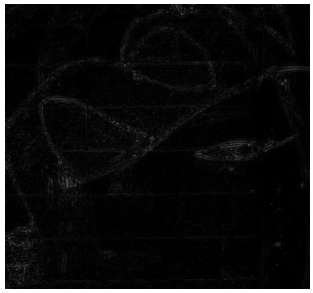
(g) Error of 3DFCNN

(h) SSPSR [27]



(i) Error of SSSPSR

(j) MCNet [35]



(k) Error of MCNet

(l) Ours

(m) Error of Ours

Figure 4: Exemplar results of $\times 8$ by our method and all comparison methods. The error is L2 distance to the ground-truth pixel values, averaged over the three bands.

5. Conclusion

In this paper, we have proposed a new method for hypserspectral image (HSI) super-resolution (SR). We build a deep convolutional network that decouple the task HSI SR into two sub-tasks: spatial super-resolution of single spectral band and joint spectral refinement over all bands. The method yields the state-of-the-art results. To further improve it, we have proposed two more contributions. First, we extend the network such that the HSI SR task can be trained together with an auxiliary RGB image SR task to gain more supervision. Second, the network is extended to also learn from datasets with LR HSIs only. The contributions greatly increase the amount of training data that HSI SR methods can use. Extensive experiments show that all the contributions are useful for the performance.

References

[1] H. K. Aggarwal and A. Majumdar. Hyperspectral image denoising using spatio-spectral total variation. *IEEE Geoscience and Remote Sensing Letters*, 13(3):442–446, 2016.

[2] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[3] T. Akgun, Y. Altunbasak, and R. M. Mersereau. Super-resolution reconstruction of hyperspectral images. *IEEE Transactions on Image Processing*, 14(11):1860–1875, 2005.

[4] Naveed Akhtar, Faisal Shafait, and Ajmal Mian. Sparse spatio-spectral representation for hyperspectral image super-resolution. In *ECCV*, 2014.

[5] N. Akhtar, F. Shafait, and A. Mian. Bayesian sparse representation for hyperspectral image super resolution. In *CVPR*, 2015.

[6] Naveed Akhtar, Faisal Shafait, and Ajmal Mian. Hierarchical beta process with gaussian process prior for hyperspectral image super resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, 2016.

[7] Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *ECCV*, 2016.

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[9] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound prediction with binaural sounds. In *European Conference on Computer Vision (ECCV)*, 2020.

[10] Lucas Beyer, Xiaohua Zhai, Avital Oliver, and Alexander Kolesnikov. S4L: self-supervised semi-supervised learning. In *ICCV*, 2019.

[11] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379, 2012.

[12] A. Chakrabarti and T. Zickler. Statistics of Real-World Hyperspectral Images. In *CVPR*, 2011.

[13] C. Chen, Y. Li, W. Liu, and J. Huang. Sirf: Simultaneous satellite image registration and fusion in a unified framework. *IEEE Transactions on Image Processing*, 24(11):4213–4224, 2015.

[14] Wenjing Chen, Xiangtao Zheng, and Xiaoqiang Lu. Hyperspectral image super-resolution with self-supervised spectral-spatial residual network. *Remote Sensing*, 13(7):1260, 2021.

[15] R. Cipolla, Y. Gal, and A. Kendall. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018.

[16] Dengxin Dai and Luc Van Gool. Ensemble projection for semi-supervised image classification. In *ICCV*, 2013.

[17] R. Dian, L. Fang, and S. Li. Hyperspectral image super-resolution via non-local sparse tensor factorization. In *CVPR*, pages 3862–3871, 2017.

[18] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016.

[19] W. Dong, F. Fu, G. Shi, X. Cao, J. Wu, G. Li, and X. Li. Hyperspectral image super-resolution via non-negative structured sparse representation. *IEEE Transactions on Image Processing*, 25(5):2337–2352, 2016.

[20] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang. Hyperspectral image super-resolution with optimized rgb guidance. In *CVPR*, 2019.

[21] S. Galliani, Charis Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler. Learned spectral super-resolution. *ArXiv*, abs/1703.09470, 2017.

[22] Alexander F.H. Goetz. Three decades of hyperspectral remote sensing of the earth: A personal view. *Remote Sensing of Environment*, 113:S5 – S16, 2009.

[23] A.A. Gowen, C.P. O'Donnell, P.J. Cullen, G. Downey, and J.M. Frias. Hyperspectral imaging – an emerging process analytical tool for food quality and safety control. *Trends in Food Science & Technology*, 18(12):590 – 598, 2007.

[24] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *CVPR*, 2020.

[25] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Köring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *CVPR*, 2021.

[26] J. Jiang, D. Liu, J. Gu, and S. Süsstrunk. What is the space of spectral sensitivity functions for digital color cameras? In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, 2013.

[27] J. Jiang, H. Sun, X. Liu, and J. Ma. Learning spatial-spectral prior for super-resolution of hyperspectral imagery. *IEEE Transactions on Computational Imaging*, 6:1082–1096, 2020.

[28] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[30] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NeurIPS*. 2014.

[31] W. Lai, J. Huang, N. Ahuja, and M. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017.

[32] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.

[33] C. Lanaras, E. Baltsavias, and K. Schindler. Hyperspectral super-resolution by coupled spectral unmixing. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[34] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[35] Qiang Li, Qi Wang, and Xuelong Li. Mixed 2d/3d convolutional network for hyperspectral image super-resolution. *Remote Sensing*, 12(10), 2020.

[36] Y. Li, Lei Zhang, C. Ding, Wei Wei, and Y. Zhang. Single hyperspectral image super-resolution with grouped deep recursive residual network. *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, 2018.

[37] Guolan Lua and Baowei Fei. Medical hyperspectral imaging: a review. *Journal of Biomedical Optics*, 2014.

[38] Shaohui Mei, Xin Yuan, Jingyu Ji, Yifan Zhang, Shuai Wan, and Qian Du. Hyperspectral image spatial super-resolution via 3d full convolutional neural network. *Remote Sensing*, 9(11), 2017.

[39] N. M. Nasrabadi. Hyperspectral target detection : An overview of current and future challenges. *IEEE Signal Processing Magazine*, 31(1):34–44, 2014.

[40] Rang M. H. Nguyen, Dilip K. Prasad, and Michael S. Brown. Training-based spectral reconstruction from a single rgb image. In *ECCV*, 2014.

[41] Ying Qu, Hairong Qi, and Chiman Kwan. Unsupervised sparse dirichlet-net for hyperspectral image super-resolution. In *CVPR*, 2018.

[42] Baifeng Shi, Judy Hoffman, Kate Saenko, Trevor Darrell, and Huijuan Xu. Auxiliary task reweighting for minimum-data learning. In *NeurIPS*, 2020.

[43] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[44] Z. Shi, C. Chen, Z. Xiong, D. Liu, and F. Wu. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[45] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

[46] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[47] Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, Dengxin Dai, , and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey, 2020.

[48] L. Wald. Data fusion: definitions and architectures: fusion of images of different spatial resolutions. In *Presses des MINES*, 2002.

[49] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020.

[50] Qi Xie, Minghao Zhou, Qian Zhao, Deyu Meng, Wangmeng Zuo, and Zongben Xu. Multispectral and hyperspectral image fusion by MS/HS fusion net. In *CVPR*, 2019.

[51] Jize Xue, Yong-Qiang Zhao, Yuanyang Bu, Wenzhi Liao, Jonathan Cheung-Wai Chan, and Wilfried Philips. Spatial-spectral structured sparse low-rank representation for hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 30:3084–3097, 2021.

[52] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar. Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, 2010.

[53] N. Yokoya, C. Grohnfeldt, and J. Chanussot. Hyperspectral and multispectral data fusion: A comparative review of the recent literature. *IEEE Geoscience and Remote Sensing Magazine*, 5(2):29–56, 2017.

[54] Y. Yuan, X. Zheng, and X. Lu. Hyperspectral image super-resolution by transfer learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(5):1963–1974, 2017.

[55] Amir Zamir, Alexander Sax, Teresa Yeo, Oğuzhan Kar, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas Guibas. Robust learning through cross-task consistency. In *CVPR*. 2020.

[56] Lei Zhang, Jiangtao Nie, Wei Wei, Yanning Zhang, Shengcai Liao, and Ling Shao. Unsupervised adaptation learning for hyperspectral imagery super-resolution. In *CVPR*, 2020.

[57] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution.

[58] Y. Zhao, Jinxiang Yang, Qingyong Zhang, L. Song, Y. Cheng, and Q. Pan. Hyperspectral imagery super-resolution by sparse representation and spectral regularization. *EURASIP Journal on Advances in Signal Processing*, 2011:1–10, 2011.

[59] Y. Zhou, A. Rangarajan, and P. D. Gader. An integrated approach to registration and fusion of hyperspectral and multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3020–3033, 2020.