IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 March 2024, accepted 1 May 2024, date of publication 6 May 2024, date of current version 14 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3396990

RESEARCH ARTICLE

Synthetic Data Pretraining for Hyperspectral Image Super-Resolution

EMANUELE AIELLO[1], MIRKO AGARLA[2], DIEGO VALSESIA[1], (Member, IEEE), PAOLO NAPOLETANO[2], (Member, IEEE), TIZIANO BIANCHI[1], (Member, IEEE), ENRICO MAGLI[1], (Fellow, IEEE), AND RAIMONDO SCHETTINI[2]

[1]Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy
[2]Department of Computer Science, Systems and Communication (DISCO), University of Milano-Bicocca, 20126 Milan, Italy

Corresponding author: Emanuele Aiello (emanuele.aiello@polito.it)

This work was supported in part by the Future Artificial Intelligence Research (FAIR) through the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3—D.D. 1555, 11/10/2022) under Grant PE00000013. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

ABSTRACT Large-scale self-supervised pretraining of deep learning models is known to be critical in several fields, such as language processing, where its has led to significant breakthroughs. Indeed, it is often more impactful than architectural designs. However, the use of self-supervised pretraining lags behind in several domains, such as hyperspectral images, due to data scarcity. This paper addresses the challenge of data scarcity in the development of methods for spatial super-resolution of hyperspectral images (HSI-SR). We show that state-of-the-art HSI-SR methods are severely bottlenecked by the small paired datasets that are publicly available, also leading to unreliable assessment of the architectural merits of the models. We propose to capitalize on the abundance of high resolution (HR) RGB images to develop a self-supervised pretraining approach that significantly improves the quality of HSI-SR models. In particular, we leverage advances in spectral reconstruction methods to create a vast dataset with high spatial resolution and plausible spectra from RGB images, to be used for pretraining HSI-SR methods. Experimental results, conducted across multiple datasets, report large gains for state-of-the-art HSI-SR methods when pretrained according to the proposed procedure, and also highlight the unreliability of ranking methods when training on small datasets.

INDEX TERMS Hyperspectral images, super resolution, synthetic data, self-supervised pretraining, spectral reconstruction.

I. INTRODUCTION

Hyperspectral imaging is a powerful technology that captures images across a wide range of the electromagnetic spectrum, revealing insights unattainable in the visible. This advanced imaging technique has diverse applications, ranging from medical diagnostics [1] and agricultural monitoring to ensure food quality, to remote sensing for environmental analysis [2], [3], as well as military applications. The rich spectral information contained in hyperspectral images (HSIs) enables precise material identification and analysis, making it an invaluable tool in these fields.

However, the design of hyperspectral imagers faces significant trade-offs. To achieve a fine spectral resolution and capture a broad range of wavelengths, compromises in the optical and sensor designs must be made that sacrifice spatial resolution in favor of spectral resolution. Moreover, the sheer amount of data produced for a hyperspectral cube can pose challenges in handling, particularly when a rapid frame rate is desired or in certain applications, such as satellite imaging, where computational and transmission resources are limited.

This limitation in spatial resolution has thus raised interest in hyperspectral image super-resolution (HSI-SR). Super-resolution techniques are well-established in the RGB imaging domain [4], [5], but their adaptation to the HSI domain is not straightforward. Indeed, one would like to

The associate editor coordinating the review of this manuscript and approving it for publication was Bing Li.

extend techniques developed for RGB images to more carefully account for spatio-spectral correlation and the characteristics of infrared bands. However, the primary obstacle is the scarcity of high-resolution hyperspectral datasets, largely due to the prohibitive costs and logistical challenges in collecting such data. Even worse, different instruments may capture different subsets of wavelengths, rendering the creation of larger datasets as collections from multiple cameras problematic. This lack of extensive, high-quality HSI data has slowed down the development and refinement of HSI-SR methods. Most of the current work focuses on the design of novel neural network architectures, potentially exploiting clever priors or layer structures in their operations. On the other hand, it is well known [6], [7] that training on more data is often more impactful than revising architectural design. Moreover, using small datasets, such as the ones in the current literature, poses the risk of producing unreliable scientific results when assessing the merits of a design over another.

In the case of hyperspectral images, collecting large labeled datasets (such as paired HR-LR HS images) for supervised training can be prohibitive or entirely impossible, due to the lack of higher resolution cameras at the desired wavelengths. This calls for the development of self-supervised pretraining techniques that can leverage a wealth of unlabeled data so that the small amount of labeled data can be used much more effectively. While techniques following this idea [8], [9] have led to robust and transferable models in natural language processing as well as other fields, a further complication arises with hyperspectral images, i.e., the overall relative scarcity of publicly available HSI products, even without demanding additional pairing with higher resolution data.

In response to this challenge, this paper introduces an innovative approach that pivots on the creation of a large-scale synthetic hyperspectral dataset. Abundant high-resolution RGB data can be found on the Internet and large datasets [10], [11] have already been developed for applications like RGB image generation, restoration, detection, etc. At the same time, spectral reconstruction techniques [12], [13], [14] have recently enjoyed great success in estimating plausible material spectra that extend to the infrared from visible RGB images only. We thus first propose to use spectral reconstruction techniques to transform a large-scale RGB dataset into an HSI dataset with, obviously not perfect, but plausible spectral content and high spatial resolution. Then, a spatial super-resolution pretext task requiring to invert an arbitrary degradation model is set up as a pretraining step. Critically, this does not require further data or annotations, as the LR images are spatially degraded from the available ones by simulating the degradation process. Finally, finetuning with paired real HSI data can be performed.

Experimental results are conducted on multiple datasets and with three state-of-the-art methods for HSI-SR. We report large gains (up to 2dB in MPSNR) in the quality of the super-resolved images when the proposed pretraining approach is followed. Moreover we conduct an ablation experiment (Sec. IV-C) that proves that pretraining with our synthetic dataset leads to better performance than using RGB images as an auxiliary task [7]. We also would like to remark that our results raise questions about the significance of results assessing merits of neural network design obtained on small datasets. In fact, we see that pretraining on the large dataset affects the relative ranking of the state-of-the-art methods. Moreover, we argue that the large-scale pretraining technique we propose could pave the way for development of bigger and more powerful neural network models.

II. BACKGROUND AND RELATED WORK

A. HYPERSPECTRAL IMAGE SUPER RESOLUTION

Hyperspectral Image Super resolution seeks to increase the spatial resolution of hyperspectral images starting from low-resolution observations. Several methods have been developed to solve this task under various settings. This work is focused on the single hyperspectral image super-resolution (SHSR) setting where the LR HS image is the only information available to reconstruct the HR image. This is contrast with other settings in which a co-registered auxiliary image with one or few bands at higher resolution is available as a guide [3], [15], [16]. The SHSR task is generally more interesting due to the wider applicability as it does not require an auxiliary input, as well as more challenging due to its highly ill-posed nature. Several approaches for SHSR have been proposed over the years [7], [17], [18], [19], [20], [21], starting from a pioneering work leveraging a Bayesian prior [17] and, more recently, deep learning methods focused on applying deep neural networks to learn a direct mapping between LR inputs and HR ground truth images. Among them, [22] makes use of 3D convolutions to explore both spatial and spectral correlation. MCNet [23] adopts a mixed convolutional module, that contains a combination of 2D and 3D convolutions to mine spatial features of the hyperspectral image and spectral information in contrast to a more computationally expensive fully 3D-convolutional model. SSPSR [24] introduces a spatial-spectral prior network to fully exploit the spatial information and the correlation between spectra. Moreover, given that hyperspectral data are very scarce and have high dimensionality the authors propose to use grouped convolution to increase the training stability. More recently, HSISR [7] proposes the use of RGB super resolution as an auxiliary task in a multi-task training framework, showing how this can be beneficial to the HSI SR task.

B. HYPERSPECTRAL DATA SCARCITY

The single image super-resolution task is an ill-posed inverse problem that necessitates a strong prior to be effectively regularized. Traditional handcrafted priors like Bayesian approaches [25] and sparse coding [26] are increasingly being replaced by learning-based approaches and neural networks which require large amounts of data for training. This is one of the main challenges in the hyperspectral domain due to the inherent difficulties and cost of data acquisitions. Commonly used datasets [21], [27], [28] usually have only a small number of images, e.g. CAVE [27] contains 20 images for training while NTIRE2020 [21] has 480 images. This limits the applicability and performance of most SHSR methods in real world cases, where better generalization abilities could be achieved if more data were available. Recent work [29] develops a novel data augmentation procedure to enlarge the number of data during the training phase of hyperspectral super resolution methods. On the other hand, some approaches have attempted to exploit the abundance of RGB images, although in a way that is different from the technique proposed in this paper. Yuan et al. [30] train a single-band SR network on natural images and apply it to HSIs in a band wise manner to exploit the spatial correlations learned on RGB data. This is clearly suboptimal as it does not exploit spectral correlation, and might also be challenged in learning features that are specific to each wavelength. Li et al. [31] develop an RGB-induced feature modulation network that exploits features learned from RGB datasets transferring them to the SHSR task. Subsequently, Li et al. [7] proposed a multi-task approach where RGB super-resolution is treated as an auxiliary task to boost the performance of the SHSR task. Their method exploits the correlation between RGB and HS image features for the super-resolution task. Our method is orthogonal and possibly complementary to all the previously proposed methods and models in the SHSR landscape.

C. SPECTRAL RECONSTRUCTION FROM RGB

Spectral reconstruction is the task of estimating the intensity of light at wavelengths beyond those captured, typically extrapolating information in infrared bands from an RGB input. This task requires to model or learn physically-plausible spectral signatures and to use the limited information in the visible, as well as spatial clues, to guess the spectrum of each pixel at the unseen wavelengths. Traditional methods for this task rely on handcrafted hyperspectral priors [32], [33]. More recently learning based approaches ([12], [13], [14] have been used to learn a direct mapping between RGB images and HS images. Among them, one of the most recent and efficient methods is MST++ [14], that exploits a Transformer-based architecture to process inputs in a multi-scale, spectral-wise manner. The method is based on a spectral-wise multi-head self attention as a basic unit, building on the intuition that HSIs are spatially sparse but spectrally self-similar. The model is built with a U-shaped structure to exploit learned features at different granularities.

III. METHOD

In this section, we propose a method to enhance the performance of any state-of-the-art neural network for spatial super-resolution of hyperspectral images. The core idea is to pretrain the neural network with a self-supervised super-resolution task on a very large dataset of synthetically generated high-resolution hyperspectral images. Since very large datasets of hyperspectral images with consistent band characteristics and high spatial resolutions do not exist, we employ spectral reconstruction techniques to convert an RGB dataset into an HSI one. Finally, finetuning with the few real HSI pairs available yields the best model.

A. SYNTHETIC DATA GENERATION

In this phase, we generate synthetic HSI data starting from an RGB dataset by employing a spectral reconstruction technique. Suppose a spectral reconstruction technique is available as a function $\phi : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{H\times W\times B}$, where B is the desired number of bands at the target wavelengths. Then, we use the spectral reconstructor ϕ on all the images of a large-scale RGB dataset $\mathcal{D}_{\text{RGB}}$ to create a synthetic HSI dataset $\mathcal{D}_{\text{HS-synth}}$:

$$\mathcal{D}_{\text{HS-synth}} = \phi(\mathcal{D}_{\text{RGB}}) \tag{1}$$

The quality of the generated synthetic dataset depends on the ability of the spectral reconstruction method to generate physically-plausible as well as spatially-consistent spectra, where each of the generated bands presents features similar to those of real HSI data at the corresponding wavelength, and is positively correlated with the performance of our pretraining procedure. As a note, most spectral reconstruction methods prioritize distortion over perception in the well-known tradeoff [34], leading to spectra that are on average more accurate but do not lie in the distribution of real spectral. It would be interesting to understand if generating data prioritizing being on the real spectral distribution (perception) leads to further improvements in the pretraining framework of this paper, but this is currently outside the scope of this paper and left as future work.

In the experiments presented in this paper we employ the state-of-the-art MST++ [14] neural network as spectral reconstructor ϕ.

B. PRETRAINING PROCEDURE

The procedure explained in the previous section allowed the creation of a large-scale dataset of hyperspectral images $\mathcal{D}_{\text{HS-synth}}$. However, $\mathcal{D}_{\text{HS-synth}}$ is just a collection of unlabeled images, so its use for pretraining purposes requires a definition of a suitable self-supervised pretext task from which features can be learned which are useful for the downstream problem our neural network model seeks to solve. Since this paper addresses the downstream problem of HSI-SR, we propose to use a self-supervised formulation of super-resolution as a pretext task for the pretraining phase. In this task, we degrade the HR synthetic HSIs with an arbitrary degradation model that is similar to the degradation model that generates real LR hyperspectral images from the HR originals. A better match between the degradation model used in the pretraining task and the degradation model of real images would result in a more effective pretraining. However, in general, one resorts to supervised training with paired real LR-HR images because the degradation model is unknown and possibly complex, so it might be difficult to approximate it for the pretraining phase. In this work, we use a simple, but widely used model, consisting of spatial convolution with a lowpass kernel, and decimation by a factor s. In formulas:

$$I_\lambda^{\mathrm{LR}} = \left(K_\lambda * I_\lambda^{\mathrm{HR}}\right)_{\downarrow s} \tag{2}$$

where I_λ^{HR} represents a band at wavelength λ of a high-resolution image in the dataset $\mathcal{D}_{\text{HS-synth}}$, K_λ is the filter kernel, and I^{LR} is the low-resolution image. For simplicity, one can use the bicubic interpolation kernel for K_λ, for all bands. However, if the point spread function of the real optical system is known at each wavelength, then using it for K_λ in this pretext task would provide a better pretext task and, possibly, better downstream performance. The pretext task trains the neural network model with a conventional regression loss, such as L1 or Charbonnier [35], between the super-resolved image obtained from I^{LR} and I^{HR}.

We remark that using a large-scale RGB dataset with high resolution images to obtain $\mathcal{D}_{\text{HS-synth}}$ is desirable because it allows the model to learn how to restore high-frequency patterns during the pretraining phase.

C. FINETUNING PROCEDURE

Subsequent to the pretraining phase, we proceed to the finetuning stage, which follows exactly the same procedure that supervised training would. In this stage, the network, initialized with the pretrained parameters is further trained on real hyperspectral data. In general, a domain gap will exist between the synthetic data and the real data in terms of image features. The finetuning process adapts the network to the characteristics of the real-world data. However, this operation is significantly more data-efficient, as the network already knows how to extract low-level features that are relevant to the super-resolution task. The finetuning stage will also correct discrepancies in the degradation model between the pretext task and the real world.

IV. EXPERIMENTS

A. SETTING

a: MODELS AND SAMPLING

We evaluate the proposed pretraining solution on three state-of-the-art methods for hyperspectral super-resolution, MCNet [23], SSPSR [24] and HSISR [7]. Our study investigates super-resolution factors of ×4 and ×8. For ×4, we train on non-overlapping 64 × 64 pixel patches cropped from the original images, while for ×8 we use larger 128 × 128 pixel patches. Both sets of patches are degraded via bicubic interpolation to create their corresponding low-resolution HSI counterparts.

b: DATASETS

We evaluate the state-of-the-art algorithms on three main datasets commonly used for benchmarking hyperspectral super-resolution, namely, the CAVE dataset [27], the Harvard dataset [28], and the NTIRE 2020 dataset [21] The images in the CAVE and NTIRE 2020 datasets consist of 31 bands spanning from 400 nm to 700 nm, with intervals of 10 nm. The images in the Harvard dataset consist of 31 bands but range from 420 nm to 720 nm. The CAVE dataset comprises 32 images, each with dimensions of 512 × 512 pixels. For the evaluation, we allocate 20 images for training and 10 for testing. Regarding the Harvard dataset, it comprises a total of 50 images, with 40 allocated for training and 10 for testing. The NTIRE 2020 dataset consists of 480 images, we assign 400 for training and 80 for testing.

TABLE 1. Quantitative results (×4 super-resolution).

Dataset	Method	Pretext	Finetune	MPSNR↑	RMSE↓	ERGAS↓
NTIRE2020	HSISR [7]	-	✓	38.9642	0.0150	2.0650
		✓	-	35.1876	0.0224	3.0351
		✓	✓	**39.8843**	**0.0137**	**1.8886**
	SSPSR [24]	-	✓	38.0740	0.0164	2.2539
		✓	-	34.9169	0.0226	3.1501
		✓	✓	**39.5264**	**0.0142**	**1.9592**
	MCNet [23]	-	✓	38.0248	0.0168	2.2834
		✓	-	40.0617	**0.0132**	1.8379
		✓	✓	**40.0631**	**0.0132**	**1.8368**
	Bicubic	-	-	34.7401	0.0235	3.1901
CAVE	HSISR [7]	-	✓	42.7645	0.0114	**3.3346**
		✓	-	38.5010	0.0176	5.1675
		✓	✓	**42.7746**	**0.0112**	3.3374
	SSPSR [24]	-	✓	40.9131	0.0144	4.0406
		✓	-	34.9800	0.0251	7.9823
		✓	✓	**42.2938**	**0.0118**	**3.5755**
	MCNet [23]	-	✓	40.7385	0.0146	4.1659
		✓	-	41.0221	0.0136	4.0295
		✓	✓	**43.5819**	**0.0105**	**3.0634**
	Bicubic	-	-	38.7380	0.0185	5.2719
Harvard	HSISR [7]	-	✓	**40.9317**	0.0132	3.0128
		✓		34.7720	0.0236	5.5637
		✓	✓	40.1527	**0.0130**	**2.9041**
	SSPSR [24]	-	✓	**40.3209**	0.0142	3.2274
		✓		33.4518	0.0327	7.7063
		✓	✓	39.9613	**0.0132**	**2.9660**
	MCNet [23]	-	✓	40.1873	0.0147	3.2606
		✓		38.8096	0.0151	3.3738
		✓	✓	**40.3471**	**0.0127**	**2.8224**
	Bicubic	-	-	38.8975	0.0167	3.8069

For the super-resolution pretraining task, we employ a subset of the Large Scale Dataset for Image Restoration (LSDIR) [6]. The dataset is composed of 87,141 RGB images, where we randomly select 20,000 and 5,000 images for the train and test set, respectively. The images are resized to match the resolution of 512 × 512 pixels. The synthetic HSI dataset $\mathcal{D}_{\text{HS-synth}}$ is obtained following the procedure presented in III-A.

c: EVALUATION METRICS

To assess the performance of all methods, we employ three commonly used metrics: Root Mean Squared Error (RMSE), which measures the average squared difference between

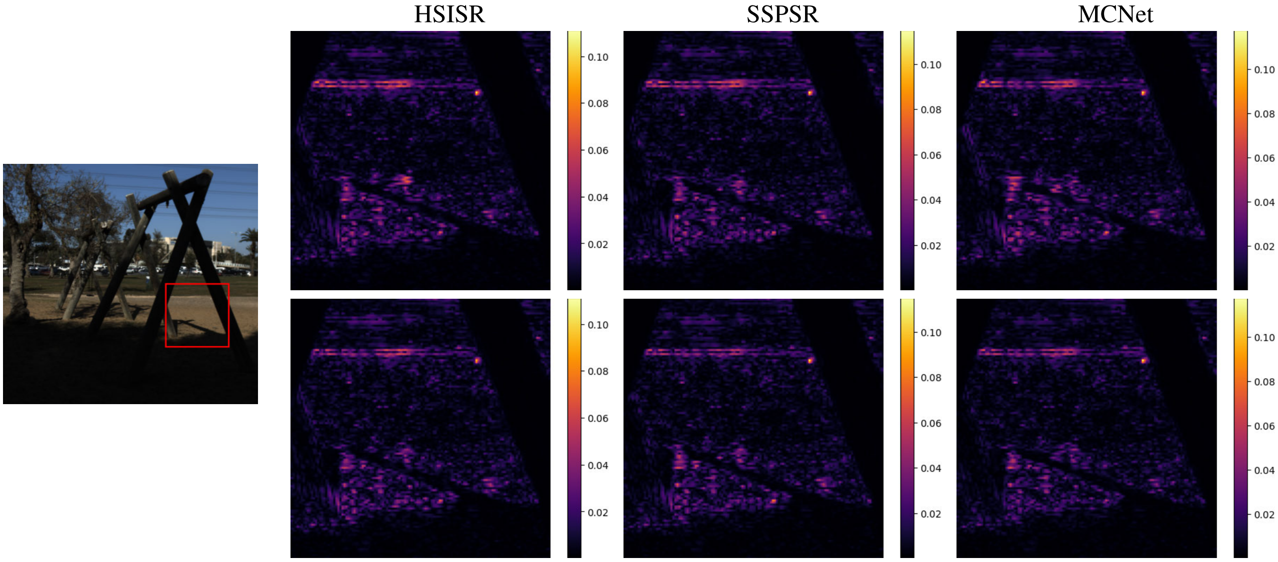

FIGURE 1. Mean Absolute Error visualization for different methods with and without the proposed pretraining strategy on an NTIRE2020 test image (RGB bands shown on the left). The first row shows baseline methods (left-to-right MPSNR: 40.06 dB, 39.71 dB, 40.12 dB), while the second row shows synthetic pretraining followed by finetuning (left-to-right MPSNR: 40.66 dB, 40.76 dB, 40.64 dB).

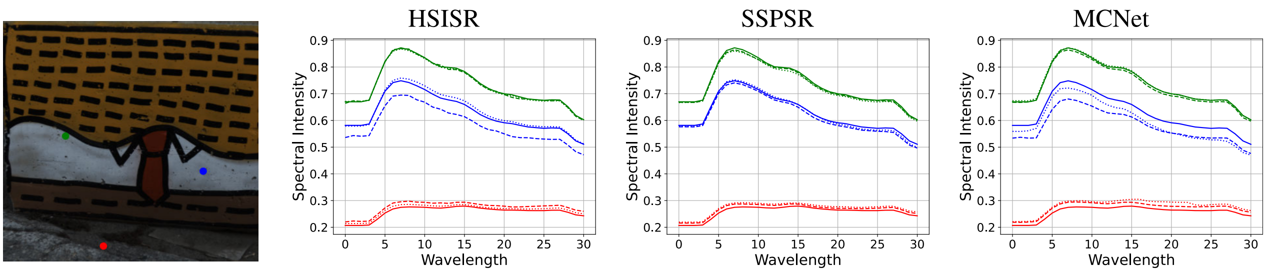

FIGURE 2. Visualization of spectra of three pixels from a super-resolved image from the NTIRE2020 test set. Ground truth: continuous line, Baseline: dashed line, Pretraining+Baseline (ours): dotted line. Best viewed zoomed.

predicted and actual values:

$$\mathrm{RMSE} = \sqrt{\frac{1}{NB}\sum_{i=1}^{N}\sum_{\lambda=1}^{B}(I_{i,\lambda}^{\mathrm{true}} - I_{i,\lambda}^{\mathrm{pred}})^2}; \tag{3}$$

N is the total number of pixels in each of the B bands, $I_{i,\lambda}^{\mathrm{true}}$ and $I_{i,\lambda}^{\mathrm{pred}}$ are the values of the i-th pixel in the λ-th band for the ground truth and predicted images, respectively;

Erreur Relative Globale Adimensionnelle de Synthese (ERGAS), a dimensionless indicator of overall reconstruction error frequently used in HSI fusion:

$$\mathrm{ERGAS} = 100s\sqrt{\frac{1}{B}\sum_{\lambda=1}^{B}\left(\frac{\mathrm{RMSE}_\lambda}{\mu_\lambda}\right)^2}; \tag{4}$$

RMSE_λ is the RMSE for each band, s represents the upsampling factor (e.g., 4 for $\times 4$ upsampling) and μ_λ is the mean value for the spectral band.

Multi-scale Peak Signal-to-Noise Ratio (MPSNR) provides a composite measure of the reconstruction fidelity.

$$\mathrm{PSNR}_\lambda = 10\log_{10}\left[\frac{\mathrm{MAX}_\lambda}{\mathrm{MSE}_\lambda}\right]; \tag{5}$$

where MAX_λ is the maximum possible value in band λ (e.g., 255 for 8-bit images). The MPSNR is the average of the PSNR_λ over all bands.

d: IMPLEMENTATION DETAILS

In the pretraining stage, we follow the author's implementation of each method using our synthetically generated LSDIR dataset. We train each model for 4 epochs and select the model with the lowest RMSE on the validation set. Then, the pretrained model is used as the starting configuration for the next training phase that involves the three selected datasets. For this phase, we still use the authors' implementations for all the methods.

The original version of the HSISR method exploits auxiliary RGB images and semi-supervised learning, as described by the authors [7]. For our experiments in Sec. IV-B, we keep the procedure for the baseline HSISR assessment, while we remove it when we use the proposed pretraining.

B. EXPERIMENTAL RESULTS

We evaluate state-of-the-art methods in the $\times 4$ and $\times 8$ super-resolution setups, presenting the results of each model

TABLE 2. Quantitative results (×8 super-resolution).

Dataset	Method	Pretext	Finetune	MPSNR↑	RMSE↓	ERGAS↓
NTIRE2020	HSISR [7]	-	✓	**33.4557**	**0.0263**	3.8437
		✓	-	31.3115	0.0342	4.7410
		✓	✓	33.4447	0.0279	**3.8028**
	SSPSR [24]	-	✓	31.7896	0.0326	4.4952
		✓	-	30.6348	0.0367	5.0694
		✓	✓	**33.2168**	**0.0285**	**3.8903**
	MCNet [23]	-	✓	31.9629	0.0327	4.4169
		✓	-	31.6321	0.0336	4.6053
		✓	✓	**33.4515**	**0.0279**	**3.7922**
	Bicubic	-	-	29.9589	0.0396	5.4594
CAVE	HSISR [7]	-	✓	37.3532	0.0206	6.0027
		✓	-	35.0913	0.0248	7.9551
		✓	✓	**37.7347**	**0.0197**	**5.8021**
	SSPSR [24]	-	✓	35.8896	0.0248	7.0394
		✓	-	34.5132	0.0269	8.0961
		✓	✓	**37.6007**	**0.0202**	**5.9358**
	MCNet [23]	-	✓	34.3116	0.0280	10.2985
		✓	-	35.3778	0.0228	**5.0354**
		✓	✓	**37.8668**	**0.0198**	5.7969
	Bicubic	-	-	34.2221	0.0304	8.4350
Harvard	HSISR [7]	-	✓	**37.3546**	**0.0201**	**4.5448**
		✓	-	33.8785	0.0263	6.4212
		✓	✓	36.1885	0.0208	4.5457
	SSPSR [24]	-	✓	**36.4563**	0.0228	4.9978
		✓	-	33.6097	0.0266	6.5786
		✓	✓	35.9873	**0.0212**	**4.6509**
	MCNet [23]	-	✓	**36.3921**	0.0234	5.0572
		✓	-	35.3778	0.0228	5.0354
		✓	✓	36.3761	**0.0203**	**4.4320**
	Bicubic	-	-	35.7409	0.0249	5.4772

both with and without pretraining using our synthetic data, followed by finetuning on the target dataset. Table 1 and Table 2 present the results for the ×4 and ×8 scenarios, respectively. For each model and dataset, three experiments are reported: i) "finetune only" is the baseline, i.e., the model as published in the literature; ii) "pretext only" is when only the pretraining phase on the synthetic dataset is performed without finetuning on the target dataset; iii) "pretext+finetuning" is the full method with pretraining on synthetic data and finetuning on the target dataset.

We can first notice that the domain gap between the synthetic and real datasets can, in general, limit the performance of using only the pretraining approach without finetuning, albeit some cases (e.g., MCNet on NTIRE2020) already report an improvement over the baseline. In general, pretraining on the synthetically generated data followed by finetuning provides the best results, sometimes with large margins, only occasionally not reporting an improvement over all the three metrics.

We can also notice that, while HSISR [7] is generally considered the state-of-the-art approach, providing the best results in the baseline setting, this is no longer true after large-scale pretraining. Indeed, MCNet after pretraining and

TABLE 3. Impact of the number of pretraining data, on HSISR finetuned with NTIRE2020 dataset (×4 Super-Resolution).

# pretext images	MPSNR↑	RMSE↓	ERGAS↓
2500	39.7836	0.01381	1.9134
5000	39.8645	0.01374	1.9021
10000	39.8777	**0.01366**	1.8933
20000	**39.8843**	0.01369	**1.8886**

TABLE 4. Effectiveness of auxiliary training tasks (HSISR, × 4 super-resolution).

Task	MPSNR↑	RMSE↓	ERGAS↓
None	38.3149	0.0154	2.2069
RGB-SR+SSL [7]	38.9642	0.0150	2.0650
Proposed	**39.8843**	**0.0137**	**1.8886**

finetuning seem to display the best overall performance. This points out a limitation of the current literature in assessing the merits of model design on small datasets, which may lead to unreliable results, as we demostrate.

Fig. 1 reports a visual comparison for one non-cherrypicked image from the NTIRE2020 test set. We visualize the mean absolute error for the different methods with and without the proposed pretraining strategy. Moreover, in Figure 2 we plot the spectra of randomly selected pixels in the super-resolved image for each method. The proposed pretraining approach yields models that are able to more faithfully reproduce the original spectrum.

C. ABLATION STUDIES

In this section, we first study the impact of the amount of synthetic data used during the proposed pretraining stage. For this experiment we pretrain the same model (HSISR [7]) with a variable number of synthetic data and we finetune each pretrained model on NTIRE2020 dataset, results are reported in Table 3. Our experiments show that increasing the number of data improves the performance with a diminished return over 10K synthetic images. We hypothesize that this may be due to the limited representational capacity of current architectures, being designed to work with a smaller amount of data.

Then, we evaluate the effectiveness of the proposed pretraining strategy vis-à-vis an alternative approach using RGB images as an auxiliary task, i.e., the procedure used in [7]. Table 4 shows the performance of the HSISR architecture under three different conditions. First, training without any auxiliary task reports the worst performance across all metrics. The semi-supervised procedure with auxiliary RGB images of [7] improves performance (about +0.6 dB in MPSNR), but it can be noticed that the proposed pretraining strategy is the most effective (about +1.5 dB improvement in MPSNR).

V. CONCLUSION AND DISCUSSION

In this study, we have demonstrated the significant impact of large-scale synthetic data pretraining in the realm of hyperspectral image super-resolution. Our approach, leverages models for spectral reconstruction to create a large HSI dataset from RGB images. When employed for a pretraining phase with a suitable pretext task, large improvements in the quality of super-resolved images have been observed on a number of datasets and state-of-the-art models. This work not only presents a viable solution to the data limitation in HSI SR but also sets a precedent for future research in synthetic hyperspectral data. We hope that our methodology will inspire further exploration and innovative applications in the field of hyperspectral imaging, extending beyond super-resolution tasks to a broader spectrum of problems.

ACKNOWLEDGMENT

(Emanuele Aiello and Mirko Agarla contributed equally to this work.)

REFERENCES

[1] G. Lu and B. Fei, "Medical hyperspectral imaging: A review," *J. Biomed. Opt.*, vol. 19, no. 1, Jan. 2014, Art. no. 010901.

[2] M. J. Khan, H. S. Khan, A. Yousaf, K. Khurshid, and A. Abbas, "Modern trends in hyperspectral image analysis: A review," *IEEE Access*, vol. 6, pp. 14118–14129, 2018.

[3] D. Wu and D.-W. Sun, "Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review—Part I: Fundamentals," *Innov. Food Sci. Emerg. Technol.*, vol. 19, pp. 1–14, Jul. 2013.

[4] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin Transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 1833–1844.

[5] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[6] Y. Li, K. Zhang, J. Liang, J. Cao, C. Liu, R. Gong, Y. Zhang, H. Tang, Y. Liu, D. Demandolx, R. Ranjan, R. Timofte, and L. Van Gool, "LSDIR: A large scale dataset for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 1775–1787.

[7] K. Li, D. Dai, and L. Van Gool, "Hyperspectral image super-resolution with RGB image super-resolution as an auxiliary task," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 4039–4048.

[8] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A. G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum, "A cookbook of self-supervised learning," 2023, *arXiv:2304.12210*.

[9] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[11] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 25278–25294.

[12] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler, "Learned spectral super-resolution," 2017, *arXiv:1703.09470*.

[13] Z. Xiong, Z. Shi, H. Li, L. Wang, D. Liu, and F. Wu, "HSCNN: CNN-based hyperspectral image recovery from spectrally undersampled projections," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 518–525.

[14] Y. Cai, J. Lin, Z. Lin, H. Wang, Y. Zhang, H. Pfister, R. Timofte, and L. V. Gool, "MST++: Multi-stage spectral-wise transformer for efficient spectral reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 744–754.

[15] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022.

[16] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[17] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3631–3640.

[18] J. Li, R. Cui, B. Li, R. Song, Y. Li, Y. Dai, and Q. Du, "Hyperspectral image super-resolution by band attention through adversarial learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4304–4318, Jun. 2020.

[19] O. Sidorov and J. Y. Hardeberg, "Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3844–3851.

[20] J. Jiang, C. Wang, X. Liu, K. Jiang, and J. Ma, "From less to more: Spectral splitting and aggregation network for hyperspectral face super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 266–275.

[21] B. Arad et al., "NTIRE 2020 challenge on spectral reconstruction from an RGB image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1806–1822.

[22] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3D full convolutional neural network," *Remote Sens.*, vol. 9, no. 11, p. 1139, Nov. 2017.

[23] Q. Li, Q. Wang, and X. Li, "Mixed 2D/3D convolutional network for hyperspectral image super-resolution," *Remote Sens.*, vol. 12, no. 10, p. 1660, May 2020.

[24] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial–spectral prior for super-resolution of hyperspectral imagery," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1082–1096, 2020.

[25] H. Irmak, G. B. Akar, and S. E. Yuksel, "A MAP-based approach for hyperspectral imagery super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2942–2951, Jun. 2018.

[26] H. Huang, J. Yu, and W. Sun, "Super-resolution mapping via multi-dictionary based sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3523–3527.

[27] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.

[28] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. CVPR*, Jun. 2011, pp. 193–200.

[29] N. Aburaed, M. Q. Alkhatib, S. Marshall, J. Zabalza, and H. Al Ahmad, "Hyperspectral data scarcity problem from a super resolution perspective: Data augmentation analysis and scheme," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2023, pp. 5057–5060.

[30] Y. Yuan, X. Zheng, and X. Lu, "Hyperspectral image superresolution by transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1963–1974, May 2017.

[31] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "RGB-induced feature modulation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023, Art. no. 5512611.

[32] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural RGB images," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 19–34.

[33] J. Wu, J. Aeschbacher, and R. Timofte, "In defense of shallow learned spectral reconstruction from RGB images," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 471–479.

[34] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6228–6237.

[35] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods," *Int. J. Comput. Vis.*, vol. 61, no. 3, pp. 1–21, Feb. 2005.

EMANUELE AIELLO received the bachelor's degree in electronics and communications engineering and the master's degree (Hons.) in telecommunications from Politecnico di Torino, where he is currently pursuing the Ph.D. degree in artificial intelligence. He is also a Teaching Assistant with Politecnico di Torino. He has gained practical experience through prestigious internships, as a Research Scientist Intern at Meta. His research interest includes multimodal deep learning.

MIRKO AGARLA received the bachelor's and master's degrees in computer science from the University of Milano-Bicocca. He is currently pursuing the Ph.D. degree in artificial intelligence with Politecnico di Torino. He is also a Teaching Assistant and a Student Mentor with the University of Milano-Bicocca. He was a Research Assistant with the IDIAP Research Institute, Huawei Research, and the University of Milano-Bicocca in the field of AI for cutting-edge research with practical applications. His primary research interests include quality control in Industry 4.0, with a focus on object detection, defect segmentation, and hyperspectral imaging. In addition, his research covers image and video processing techniques, addressing quality enhancement, super-resolution, and defenses against adversarial attacks.

DIEGO VALSESIA (Member, IEEE) received the Ph.D. degree in electronic and communication engineering from Politecnico di Torino, in 2016. He is currently an Assistant Professor with the Department of Electronics and Telecommunications (DET), Politecnico di Torino. His main research interests include the processing of remote sensing images and deep learning for inverse problems in imaging. He is a member of the EURASIP Technical Area Committee for Signal and Data Analytics for Machine Learning and a member of the ELLIS Society. He was a recipient of the IEEE ICIP 2019 Best Paper Award and the IEEE Multimedia 2019 Best Paper Award. He is an Associate Editor of IEEE Transactions on Image Processing, for which he received the 2023 Outstanding Editorial Board Member Award.

PAOLO NAPOLETANO (Member, IEEE) has been an Associate Professor of computer science with the Department of Informatics, Systems and Communication, University of Milano-Bicocca, since 2021. His main research interests include artificial intelligence, machine and deep learning, computer vision, pattern recognition, intelligent sensors, biological signal processing, and human–machine systems. He is a member of the European Laboratory for Learning and Intelligent Systems (ELLIS Society). He is an Associate Editor of IEEE Journal of Biomedical and Health Informatics, *Neurocomputing* (Elsevier), *IET Signal Processing*, *Sensors* (MDPI), and *Smart Cities* (MDPI). He is the Chair of the IEEE CTSoc Machine Learning, Deep Learning and AI in CE (MDA) Technical Committee (TC).

TIZIANO BIANCHI (Member, IEEE) received the M.Sc. degree (Laurea) in electronic engineering and the Ph.D. degree in information and telecommunication engineering from the University of Florence, Italy, in 2001 and 2005, respectively. From 2005 to 2012, he was a Research Assistant with the Department of Electronics and Telecommunications, University of Florence. In 2012, he joined Politecnico di Torino as an Assistant Professor, where he is currently an Associate Professor. He has authored more than 100 papers in international journals and conference proceedings. His research interests include multimedia security technologies, multimedia forensics, and the processing of remote-sensing images. He was a recipient of the IEEE Multimedia 2019 Best Paper Award and the 2021 and 2022 Best Associate Editor Award of the *Journal of Visual Communication and Image Representation*. He is currently an Associate Editor of IEEE Transactions on Information Forensics and Security and a Senior Area Editor of the *Journal of Visual Communication and Image Representation*.

ENRICO MAGLI (Fellow, IEEE) received the M.Sc. and Ph.D. degrees from the Politecnico di Torino, Italy, in 1997 and 2001, respectively. He is currently a Full Professor with Politecnico di Torino, where he leads the Image Processing and Learning Group, performing research in the fields of deep learning for image and video processing, image compression, and image forensics for multimedia and remote sensing applications. He is a fellow of the ELLIS Society for the Advancement of Artificial Intelligence in Europe. He was a recipient of the IEEE Geoscience and Remote Sensing Society 2011 Transactions Prize Paper Award, the IEEE ICIP 2015 Best Student Paper Award (as a Senior Author), the IEEE ICIP 2019 Best Paper Award, the IEEE Multimedia 2019 Best Paper Award, and the 2010 and 2014 Best Associate Editor Award of IEEE Transactions on Circuits and Systems for Video Technology. He is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology and *EURASIP Journal on Image and Video Processing*. He has been an IEEE Distinguished Lecturer, from 2015 to 2016.

RAIMONDO SCHETTINI is currently a Full Professor with the University of Milano-Bicocca, Italy, leading the Imaging and Vision Laboratory. With more than 30 years of experience, he has published extensively in color imaging and image processing, supervised numerous Ph.D. students, and led research projects in collaboration with prominent companies. He holds fellowships with the International Association of Pattern Recognition (IAPR), Asia-Pacific Artificial Intelligence Association (AAIA), and International Artificial Intelligence Industry Alliance (AIIA). He is listed on Stanford University's World Ranking Scientists List. He serves as the Editor-in-Chief for the *Journal of Imaging* (MDPI).

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement