博士論文

Deep Unsupervised Hyperspectral Image Super-resolution

(教師無し深層学習によるハイパースペクトル画像の超解像度)

令和　5 年　9 月

LIU ZHE

山口大学大学院創成科学研究科

Deep Unsupervised Hyperspectral Image Super-resolution

Author:
Zhe LIU

Supervisor:
Prof. Xian-Hua HAN

Graduate School of Sciences and Technology for Innovation
Natural Science

YAMAGUCHI UNIVERSITY

Abstract

Graduate School of Sciences and Technology for Innovation
Natural Science

Doctor of Philosophy

Deep Unsupervised Hyperspectral Image Super-resolution

by Zhe LIU

Hyperspectral (HS) imaging can capture the detailed spectral signature of each spatial location of a scene and leads to better understanding of different material characteristics than traditional imaging systems. However, existing HS sensors can only provide low spatial resolution images at a video rate in practice. Thus reconstructing high-resolution HS (HR-HS) image via fusing a low-resolution HS (LR-HS) image and a high-resolution RGB (HR-RGB) image with image processing and machine learning technique, called as hyperspectral image super resolution (HSI SR), has attracted a lot of attention. Existing methods for HSI SR are mainly categorized into two research directions: mathematical model based method and deep learning based method. Mathematical model based methods generally formulate the degradation procedure of the observed LR-HS and HR-RGB images with a mathematical model and employ an optimization strategy for solving. Due to the ill-posed essence of the fusion problem, most works leverage the hand-crafted prior to model the underlying structure of the latent HR-HS image, and pursue a more robust solution of the HR-HS image. Recently, deep learning-based approaches have evolved for HS image reconstruction, and current efforts mainly concentrated on designing more complicated and deeper network architectures to pursue better performance. Although impressive reconstruction results can be achieved compared with the mathematical model based methods, the existing deep learning methods have the following three limitations. 1) They are usually implemented in a fully supervised manner, and require a large-scale external dataset including the degraded observations: the LR-HS/HR-RGB images and their corresponding HR-HS ground-truth image, which are difficult to be collected especially in the HSI SR task. 2) They aim to learn a common model from training triplets, and are undoubtedly insufficient to model abundant image priors for various HR-HS images with rich contents, where the spatial structures and spectral characteristics have considerable difference. 3) They generally assume that the spatial and spectral degradation procedures for capturing the LR-HS and HR-RGB images are fixed and known, and then synthesize the training triplets to learn the reconstruction model, which would produce very poor recovering performance for the observations with different degradation procedures. To overcome the above limitations, our research focuses on proposing the unsupervised learning-based framework for HSI SR to learn the specific prior of an under-studying scene without any external dataset. To deal with the observed images captured under different degradation procedures, we further automatically learn the spatial blurring kernel and the camera spectral response function (CSF) related to the specific observations, and incorporate them with the above unsupervised framework to build a high-generalized blind unsupervised HSI SR paradigm.

Moreover, Motivated by the fact that the cross-scale pattern recurrence in the natural images may frequently exist, we synthesize the pseudo training triplets from the degraded versions of the LR-HS and HR-RGB observations and themself, and conduct supervised and unsupervised internal learning to obtain a specific model for the HSI SR, dubbed as generalized internal learning. Overall, the main contributions of this dissertation are three-fold and summarized as follows:

1. A deep unsupervised fusion-learning framework for HSI SR is proposed. Inspired by the insights that the convolution neural networks themself possess large amounts of image low-level statistics (priors) and can more easy to generate the image with regular spatial structure and spectral pattern than noisy data, this study proposes an unsupervised framework to automatically generating the target HS image with the LR-HS and HR-RGB observations only without any external training database. Specifically, we explore two paradigms for the HS image generation: 1) learn the HR-HS target using a randomly sampled noise as the input of the generative network from data generation view; 2) reconstructing the target using the fused context of the LR-HS and HR-RGB observations as the input of the generative network from a self-supervised learning view. Both paradigms can automatically model the specific priors of the under-studying scene by optimizing the parameters of the generative network instead of the raw HR-HS target. Concretely, we employ an encoder-decoder architecture to configure our generative network, and generate the target HR-HS image from the noise or the fused context input. We assume that the spatial and spectral degradation procedures for the under-studying LR-HS and HR-RGB observation are known, and then can produce the approximated version of the observations by degrading the generated HR-HS image, which can intuitively used to obtain the reconstruction errors of the observation as the loss function for network training. Our unsupervised learning framework can not only model the specific prior of the under-studying scene to reconstruct a plausible HR-HS estimation without any external dataset but also be easy to be adapted to the observations captured under various imaging conditions, which can be naively realized by changing the degradation operations in our framework.

2. A novel blind learning method for unsupervised HSI SR is proposed. As described in the above deep unsupervised framework for HSI SR that the spatial and spectral degradation procedures are required to be known. However, different optical designs of the HS imaging devices and the RGB camera would cause various degradation processes such as the spatial blurring kernels for capturing LR-HS images and the camera spectral response functions (CSF) in the RGB sensors, and it is difficult to get the detailed knowledge for general users. Moreover, the concrete computation in the degradation procedures would be further distorted under various imaging conditions. Then, in real applications, it is hard to have the known degradation knowledge for each under-studying scene. To handle the above issue, this study exploits a novel parallel blind unsupervised approach by automatically and jointly learning the degradation parameters and the generative network. Specifically, according to the unknown components, we propose three ways to solve different problems: 1) a spatial-blind method to automatically learn the spatial blurring kernel in the capture of the LR-HS observation with the known CSF of the RGB sensor; 2) a spectral-blind method to automatically learn the CSF transformation matrix in the capture of the HR-RGB observation with known burring kernel in the HS imaging device; 3) a complete-blind method to simultaneously learn both spatial blurring kernel and CSF matrix. Based on our previously proposed unsupervised framework, we particularly design the special convolution layers for parallelly realizing the spatial and spectral degradation procedures, where the layer

parameters are treated as the weights of the blurring kernel and the CSF matrix for being automatically learned. The spatial degradation procedure is implemented by a depthwise convolution layer, where the kernels for different spectral channel are imposed as the same and the stride parameter is set as the expanding scale factor, while the spectral degradation procedure is achieved with a pointwise convolution layer with the output channel 3 to produce the approximated HR-RGB image. With the learnable implementation of the degradation procedure, we construct an end-to-end framework to jointly learn the specific prior of the target HR-HS images and the degradation knowledge, and build a high-generalized HSI SR system. Moreover, the proposed framework can be unified for realizing different versions of blind HSI SR by fixing the parameters of the implemented convolution as the known blurring kernel or the CSF, and is highly adapted to arbitrary observation for HSI SR.

3. A generalized internal learning method for HSI SR is proposed. Motivated by the fact that natural images have strong internal data repetition and the cross-scale internal recurrence, we further synthesize labeled training triplets using the LR-HS and HR-RGB observation only, and incorporate them with the un-labeled observation as the training data to conduct both supervised and unsupervised learning for constructing a more robust image-specific CNN model of the under-studying HR-HS data. Specifically, we downsample the observed LR-HS and HR-RGB image to their son versions, and produce the training triplets with the LR-HS/HR-RGB sons and the LR-HS observation, where the relation among them would be same as among the LR-HS/HR-RGB observations and the HR-HS target despite of the difference in resolutions. With the synthesized training samples, it is possible to train a image-specific CNN model to achieve the HR-HS target with the observation as input, dubbed as internal learning. However, the synthesized labeled training samples usually have small amounts especially for a large spatial expanding factor, and the further down-sampling on the LR-HS observation would bring severe spectral mixing of the surrounding pixels causing the deviation of the spectral mixing levels at the training phase and test phase. Therefore, these limitations possibly degrade the super-resolved performance with the naive internal learning. To mitigate the above limitations, we incorporate the naive internal learning with our self-supervised learning method for unsupervised HSI SR, and present a generalized internal learning method to achieve more robust HR-HS image reconstruction.

Acknowledgements

During my Ph.D. course, I have experienced confusion, loss, and depression on many late nights. But, I have been more enriched, grown, and experienced, and I am grateful for the warmth and assistance I have received. I have many people to thank.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Prof. Xian-Hua Han, for guiding me with patience and wisdom through my Ph.D. course. I feel immensely fortunate to have been Prof. Han's first Ph.D. student. I am grateful for the care and support she has provided me with since my arrival at Yamaguchi University. I remember vividly how Prof. Han patiently introduced me to the campus life and research seminars, making it possible for me to adjust quickly to the unfamiliar environment. Under Prof. Han's guidance, I gained the confidence to present my research findings logically and became a fearless presenter. Additionally, she assisted me in securing a scholarship from the Japanese Ministry of Education, Culture, Sports, Science, and Technology, which allowed me to focus on my research without financial concerns. The four years I spent at Yamaguchi University were the most rewarding of my life, and I feel blessed to have had Prof. Han as my mentor and guide. Prof. Han's dedication to research, professional experience, and meticulous attitude have inspired me to engage with my discipline actively. I would also like to express my gratitude to Prof. Han for her efforts in lab construction and safety, as well as for her hard work in organizing and hosting group meetings to create a warm and excellent research environment for everyone. Of all the teachers who have taught me, Prof. Han is the one I admire and respect the most. I can honestly say that Prof. Han has changed my life and been my benefactor. I am extremely fortunate to have met such a wonderful teacher in my life. The best way to learn is to learn from the best, and Prof. Han is the best teacher I have ever had. Thank you for everything you have done for me, Prof. Han. Best wishes to my beloved teacher!

I would like to express my sincere gratitude to my lab mates for their invaluable assistance during these four years. We have shared our doubts about the project and improved our research skills together. Even if the research task is heavy, I can still laugh and relax because of your presence, and every day in the lab I am grateful for the reward and growth. And I would like to thank my international friends and Japanese friends I met in Yamaguchi for giving me support and warmth in a foreign country. I wish you all good health, good luck and good cheerfulness each day.

I would like to thank TV series such as Journey to the West, My Own Swordsman, Empresses in the Palace: The Legend of Zhen Huan, The Story of Ming Lan and The Story of the Cooking Squad, which accompanied me through countless sleepless nights, taught me many life lessons with humorous or profound performances, and gave me the strength to survive every moment of collapse.

I would like to express my deepest gratitude to the unconditional understanding and support of my family, I was able to follow my heart's choice at every fork in the road. And I would like to thank my besties for always being there to lend an ear when I needed to bounce off some blue-sky thinking. Your support has meant the world to me, and I am truly blessed to have you in my life.

Finally, here is to being footloose and fancy-free! I hope I myself can find true peace and be rich.

Contents

Abstract **iii**

Acknowledgements **vii**

1 Introduction **1**
1.1 Hyperspectral Imaging Technology 1
1.2 Applications of Hyperspectral Images 2
1.3 Limitations of Hyperspectral Images 3
1.4 Resolution Enhancement of Hyperspectral Images 5
1.4.1 Spatial Resolution Enhancement 5
1.4.2 Spectral Resolution Enhancement 6
1.4.3 Fusion Enhancement 7
1.5 Contributions of the Dissertation 8

2 Hyperspectral Image Super-Resolution **11**
2.1 Mathematical Model-based Super-Resolution 13
2.2 Deep Learning-based Super-Resolution 15
2.3 Related Work 17
2.3.1 Deep Image Prior 17
2.3.2 Internal Learning 18
2.3.3 Self-Supervised Learning 20

3 Unsupervised Learning of Hyperspectral Image Prior **21**
3.1 Problem formulation 21
3.2 Motivation 22
3.2.1 Common Prior Learning with Supervision 23
3.2.2 Specific Prior Learning with Un-supervision 23
3.3 Architecture of the generative neural network 24
3.4 Input data to the generative neural network 25
3.4.1 The Noise Input 26
3.4.2 The Fusion Context Input 26
3.4.3 The RGB Input 27
3.5 Loss Function 30
3.6 Experiment Results 30
3.6.1 Experimental Setting 30
Datasets 30
Evaluation Metrics 31
Implementation Details 31
3.6.2 Performance Evaluation 32
Comparison with traditional optimization-based methods . . . 32
Comparison with deep learning-based methods 34
Ablation study 37
Perceptual Quality 39

3.7 Conclusion 40

4 Unsupervised Blind Learning of Hyperspectral Image Super-Resolution 45
4.1 Traditional Non-Blind Spatial/Spectral Degradation 45
4.2 Learnable Blind Method 46
4.2.1 Non-blind 47
4.2.2 (Semi-)Blind 47
4.3 Experiment Results 49
4.3.1 Comparison with (Semi-)Blind Methods 49
4.3.2 Comparison with SoTA Methods 50
4.3.3 Ablation Study 51
Different Settings in the Proposed ISGM (blind) Method 51
4.3.4 Perceptual Quality 53
4.4 Conclusion 54

5 Unsupervised Internal Learning of Hyperspectral Image Super-Resolution 57
5.1 Deep Supervised External Learning 58
5.2 Zero-Shot Learning 59
5.3 Deep Internal Learning 59
5.4 Proposed Image-Specific Generative Model of Generalized Internal Learning 60
5.4.1 Motivation 60
5.4.2 Unsupervised Internal Learning: UIL 62
5.4.3 Pseudo-Supervised Internal Learning: SIL 63
5.4.4 Image-Specific Generative Network of Generalized Internal Learning: ISGM (GIL) 63
5.5 Experiment Results 64
5.5.1 Comparisons with the State-of-the-art Methods 65
5.5.2 Ablation Study 65
5.5.3 Perceptual Quality 68
5.6 Conclusion 68

6 Conclusion 71

List of Figures

1.1 Illustrated examples of hyperpsectral images and RGB images. 2
1.2 Applications of hyperspectral images 3
1.3 The contribution of this dissertation 8

3.1 Conceptual diagram of the proposed image-specific generative network (ISGM) for unsupervised deep HSI SR. 25
3.2 Proposed framework of deep RGB-guided image-specific generative network (ISGM (RGB)). FT Conv Module denotes feature transfer convolution module. Spatial Degradation module is implemented by a depth-wise convolution layer while Spectral Degradation module is conducted by a point-wise convolution layer. 29
3.3 Example images in the CAVE and Harvard datasets. 31
3.4 Recovered HR-HS image of the ′Balloon′ sample in the CAVE dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 8. 40
3.5 Recovered HR-HS image of the ′Balloon′ sample in the CAVE dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 16. 40
3.6 Recovered HR-HS image of the ′Balloon′ sample in the CAVE dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 32. 41
3.7 Recovered HR-HS image of the ′img1′ sample in the Harvard dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 8. 41
3.8 Recovered HR-HS image of the ′img1′ sample in the Harvard dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 16. 42
3.9 Recovered HR-HS image of the ′img1′ sample in the Harvard dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 32. 42

3.10 Visual compared results of one representative image: paints from the CAVE dataset with the traditional optimization-based method: CSU [93] and NSSR [46], the supervised deep learning-based methods: DHSIS [57], and the un-supervised deep learning-based methods: uSDN [77], DHP [101] for spatial expanding factor: 16. (a) Visualized results of spectral band 16: 550 nm. (b) Visualized results of spectral band 31: 700 nm. 43

3.11 Visual compared results of one representative image: imgb4 from the Harvard dataset with the traditional optimization-based method:CSU [93] and NSSR [46], the supervised deep learning-based methods: DHSIS [57], and the un-supervised deep learning-based methods: uSDN [77], DHP [101] for spatial expanding factor: 16. (a) Visualized results of spectral band 16: 550 nm. (b) Visualized results of spectral band 31: 700 nm. 44

3.12 Sam results (spectral band 16: 550nm) of one representative image: paints from the CAVE dataset and imgb4 from the Harvard dataset with the traditional optimization-based method: CSU [93] and NSSR [46], the supervised deep learning-based methods: DHSIS [57], and the un-supervised deep learning-based methods: uSDN [77], DHP [101] for spatial expanding factor: 16. 44

4.1 Concept of two parallel blocks for the degradation 48

4.2 Visual difference results of the representative images: Paints in the CAVE dataset of mathematical optimization-based methods: CSU [45], NSSR [46] and deep learning-based methods: uSDN [77], DUFL [109], DHSIS [57] and our method on the CAVE dataset with the up-scale factor 16. 54

4.3 Visual difference results of the representative images: imgb4 in the Harvard dataset of mathematical optimization-based methods: CSU [45], NSSR [46] and deep learning-based methods: uSDN [77], DUFL [109], DHSIS [57] and our method on the CAVE dataset with the up-scale factor 16. 55

4.4 The recovered spectral curve of one pixel in the representative images: Paints in the CAVE dataset of mathematical optimization-based methods: CSU [45], NSSR [46] and deep learning-based methods: uSDN [77], DUFL [109], DHSIS [57] and our method on the CAVE dataset with the up-scale factor 16. 55

4.5 The recovered spectral curve of one pixel in the representative images: imgb4 in the Harvard dataset of mathematical optimization-based methods: CSU [45], NSSR [46] and deep learning-based methods: uSDN [77], DUFL [109], DHSIS [57] and our method on the CAVE dataset with the up-scale factor 16. 56

5.1 Conceptual diagram of the proposed ISGM (GIL) method. 62

5.2 Visualization error map results of an example image: paints from CAVE dataset compared with unsupervised prior-based methods: CSU [93], NSSR [46], supervised deep learning methods: DHSIS [57], unsupervised deep learning-based methods: DHP [59], uSDN [77], GDD [81] and the proposed method. 68

5.3 Visualization error map results of an example image: imgb4 from Harvard dataset compared with unsupervised prior-based methods: CSU [93], NSSR [46], supervised deep learning methods: DHSIS [57], unsupervised deep learning-based methods: DHP [59], uSDN [77], GDD [81] and the proposed method. 69

5.4 Visualization SAM results of example images from CAVE and Harvard datasets compared with unsupervised prior-based methods: CSU [93], NSSR [46], supervised deep learning methods: DHSIS [57], unsupervised deep learning-based methods: DHP [59], uSDN [77], GDD [81] and the proposed method. 69

List of Tables

3.1 Comparison results between state-of-the-art traditional optimization-based methods methods (GOMP [95], MF [50], and SNNMF [66]) and ISGM (noise) method in the CAVE and Harvard Datasets with the Expanding Factor: 32. 33
3.2 Comparison results between state-of-the-art traditional optimization-based methods methods and ISGM (noise, fusion, RGB) and ISGM+ (noise) method in the CAVE and Harvard Datasets with the Expanding Factor: 8 and 16. 34
3.3 Comparison results between unsupervised methods with learned image priors (DHIP [59] and uSDN [77]), and fully supervised deep-learning methods (PNN [99], 3D-CNN [100], and CMHF-net [75]) and ISGM (noise) method in the CAVE and Harvard Datasets with the Expanding Factor: 32. 35
3.4 Comparison with the deep non-blind learning-based methods including the supervised approaches: SSFNet [56], ResNet [102], DHSIS [57] and the un-supervised approaches: uSDN [77], DHP [101], GDD [81] on the CAVE and Harvard datasets for both spatial expanding factors: 8 and 16. 36
3.5 Ablation study of the evaluation results using our proposed method: ISGM+ (noise) with different weights α values of 0.3, 0.5 and 0.7 in the CAVE and Harvard datasets for both spatial expanding factors: 8, 16 and 32. 37
3.6 Ablation study of our proposed ISGM (fusion) method with different α in the loss function (Eq. 3.6) in the CAVE dataset for upscale factor: 8. 38
3.7 Ablation studies of different block numbers of the generative network and loss terms on CAVE dataset with the up-scale factor 8. 38
3.8 Ablation studies of different network inputs on CAVE dataset with the up-scale factor 8. 39

4.1 Comparison with the (semi-)blind methods including the unsupervised semi-blind approaches: UAL [82] and the semi-blind implementation of NSSR [46] on the CAVE and Harvard datasets for both spatial expanding factors: 8 and 32. 50
4.2 Compared results with the SoTA methods including mathematical optimization-based and deep learning-based methods on both CAVE and Harvard datasets with the up-scale factor 16. 51
4.3 Compared results of two representative images from both CAVE and Harvard datasets with the SoTA methods including the u^2-MDN [110] method for the up-scale factor 32. 52
4.4 Compared results with the SoTA methods on the NUS dataset with the up-scale factor 16. 52

4.5 Compared results using different settings: the input of the generative network (noise or the fused context: $Z_{\{XY\}}$) and the spatial degradation kernels including the known kernel (Lanczos), and the unknown kernels: the widely used Gaussian kernels with different standard deviations and automatically learned kernel, in our proposed DSSH method for both CAVE and Harvard datasets with the spatial expanding factors: 8 and 16. 53
4.6 Ablation study with different α in the loss function (Eq. 4.7) on the CAVE dataset for upscale factor: 8. 54

5.1 Compared evaluation results between unsupervised prior-based methods: BSR [96], CSU [93], GOMP [96], SNNMF [97], NSSR [46], supervised deep learning methods: SSFNet [56], ResNet [102], DHSIS [57], unsupervised deep learning-based methods: uSDN [77], DHP [59], DUFL [109], GDD [81] and our proposed method with the scale factors: 8 and 16 in CAVE and Harvard datasets. 66
5.2 Ablation studies with the three learning strategies: UIL, SIL and ISGM (GIL) and w/o data augmentation of the UIL for different upscale factors: 4, 8, 16 in both CAVE and Harvard datasets. 67
5.3 Ablation studies for network architectures. 68

Chapter 1

Introduction

In recent decades, hyperspectral imaging technologies have made significant progress in providing hyperspectral images for applications ranging from agriculture and astronomy to surveillance and medicine. However, the images acquired by existing hyperspectral imaging systems are difficult to provide all possible required detail distributions in all domains, such as spatial, temporal or spectral, at the same time. Current research efforts have focused on improving the spatial resolution of hyperspectral images. Hyperspectral image super-resolution is an effective means to increase the spatial resolution of hyperspectral images. In this dissertation, we focus on how to achieve hyperspectral image super-resolution. Thus, we give a brief overview of hyperspectral imaging technologies, hyperspectral images and their applications. Then we briefly introduce different super-resolution methods and summarize their basic principles.

1.1 Hyperspectral Imaging Technology

Hyperspectral imaging technology captures detailed information about the composition and physical properties by measuring the reflected or emitted radiation in many narrow, contiguous wavelength bands. It is a cutting-edge scientific tool that has revolutionized the way we gather and analyze data about the natural world. This technology works by capturing detailed information about an object or scene across a broad range of wavelengths, creating a highly accurate and detailed image that can reveal key information about its composition and properties. At its core, hyperspectral imaging uses specialized cameras and spectrometers to capture light across the electromagnetic spectrum, from visible light to infrared and beyond. This data is then processed and analyzed to create an image that displays the unique spectral signatures of different materials in the scene. This allows scientists and researchers to identify specific materials, minerals, and chemicals, and to study the interactions between different substances in the environment.

Specifically, hyperspectral imaging sensors can capture the detailed distribution of the spectral direction in each spatial location of a scene and provide more abundant spectral information of the object than a generic color image. As a result, hyperspectral (HS) images have been widely used in various fields, such as remote sensing [1,2], mineral exploration [3], computer vision [4] and medical diagnosis [5], and show promising performances. However, in real imaging systems, the acquired images generally have critical tradeoffs between spatial and spectral resolutions due to the limited amount of incident energy and the interference of imaging environments. Then, the existing HS imaging systems usually provide data only with a large number of spectral bands but the low resolution in the spatial domain. On the other hand, the multispectral imaging systems or general color image sensors can obtain high spatial resolution but with a small number of spectral bands (e.g., the

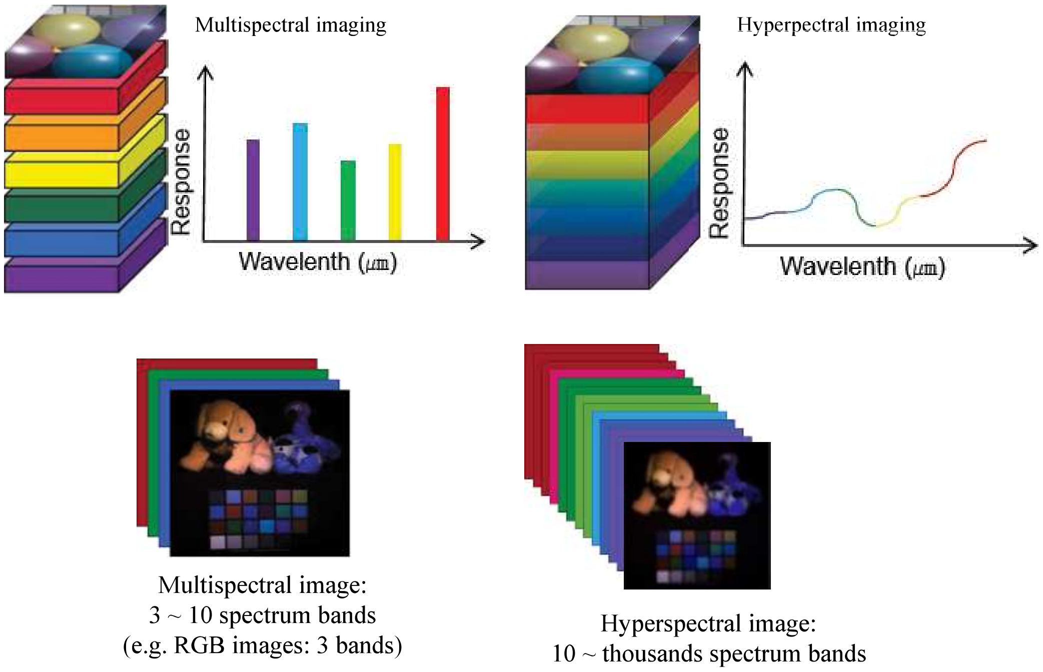


FIGURE 1.1: Illustrated examples of hyperpsectral images and RGB images.

standard RGB image). In Fig. 1.1, we can see examples of hyperpsectral images and RGB images. In conclusion, hyperspectral imaging technology is a highly innovative and valuable tool that has revolutionized the way we gather and analyze data about the natural world. Whether it is used to monitor crops, track wildlife, or detect enemy targets, hyperspectral imaging provides us with a wealth of information that can help us make more informed decisions and improve our understanding of the world around us. Despite its challenges, hyperspectral imaging technology is poised to play an increasingly important role in a wide range of fields, and its impact will continue to be felt for years to come.

1.2 Applications of Hyperspectral Images

Hyperspectral images are a type of remote sensing data that provide detailed information about the composition and properties of objects and materials. These images are captured by sensors that collect data across a wide range of the electromagnetic spectrum, typically from visible light to shortwave infrared. The result is a large number of narrow, contiguous bands of data that provide highly specific information about the materials and substances present in the scene being imaged.

One of the most important applications of hyperspectral imaging technology is in the field of remote sensing (see as Fig. 1.2 (a)). By using satellite and airborne platforms, scientists are able to gather data about large areas of land and ocean, including remote and inaccessible regions. This has important implications for a wide range of applications, including agriculture, environmental monitoring, and resource management. For example, hyperspectral imaging can be used to monitor crop growth and health, detect environmental pollutants and changes in land use, and track the movement of wildlife.

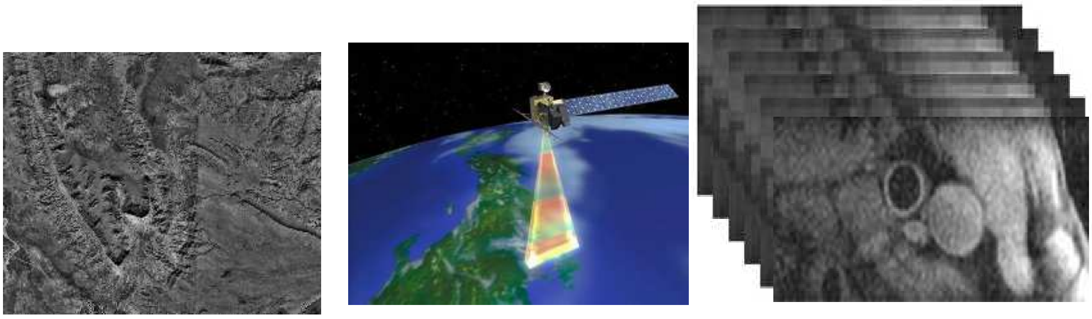
Another important application of hyperspectral images is in agriculture. By analyzing the spectral characteristics of crops, farmers can identify problems such as nutrient deficiencies, pests, and diseases. This information can then be used to optimize fertilizer application and make informed decisions about pest control, ultimately leading to improved crop yields and reduced waste.

Other key applications of hyperspectral images is in mineral and resource exploration. By analyzing the unique spectral signatures of minerals, geologists and mining companies can identify the presence and distribution of valuable minerals, such as copper, iron, and gold, in the subsurface. This information can be used to guide drilling and excavation, reducing the time and cost required for mineral exploration.

Hyperspectral images also have important applications in environmental monitoring and management (see as Fig. 1.2 (b)). By analyzing the spectral characteristics of forests, wetlands, and other ecosystems, scientists can monitor changes in vegetation cover and track the progression of land-use changes, such as deforestation and urbanization. This information can be used to better understand the impacts of human activities on the environment and make informed decisions about conservation and land-use planning.

In addition, hyperspectral images have a wide range of military and security applications. By analyzing the spectral signatures of objects and materials, military and intelligence agencies can identify and track the movement of weapons, vehicles, and personnel. This information can be used to support military operations and enhance national security. Hyperspectral images is also increasingly being used in the medical field, where it is being used to diagnose and monitor a range of health conditions, including skin cancer and other diseases (see as Fig. 1.2 (c)).

Overall, hyperspectral images are a valuable tool for a wide range of applications, from mineral exploration and agriculture to environmental monitoring and military operations. Their ability to provide highly specific information about the composition and properties of objects and materials makes them a valuable resource for a variety of industries and disciplines.



(a) Remote sensing (b) Environmental monitoring (c) Medical diagnose

FIGURE 1.2: Applications of hyperspectral images

1.3 Limitations of Hyperspectral Images

Hyperspectral images, also known as hyperspectral data or hyperspectral cubes, are a type of data product obtained from hyperspectral imaging technology that captures information about the object across a wide range of spectral bands. Despite the vast amount of information they provide, hyperspectral images have limitations that need to be considered when interpreting and analyzing the data.

One of the main limitations of hyperspectral images is their high cost of the specialized cameras and spectrometers required to capture the data. Hyperspectral sensors are expensive to build, launch, and maintain, and the data acquisition process is time-consuming and requires specialized equipment. This makes it challenging for researchers and scientists to access and use hyperspectral data in their work and can limit the widespread adoption of the technology and limit its use to only a few well-funded institutions.

Hyperspectral images also face limitations in terms of atmospheric correction. The Earth's atmosphere can cause significant errors in the data acquired from hyperspectral sensors, making it necessary to apply atmospheric correction algorithms to the data before analysis. However, these algorithms are not always accurate, and errors can remain in the data, affecting the accuracy of the results.

Additionally, the data size of hyperspectral images is very large, making it challenging to process and analyse. The data must be pre-processed, and algorithms must be developed to extract useful information from the data, adding another layer of complexity to the analysis process. Processing and analysing the vast amounts of data generated by hyperspectral imaging can be a time-consuming and computationally intensive process, requiring specialized software and high-performance computing resources.

The biggest limitation is the limited spatial resolution of hyperspectral images. Unlike other hyperspectral imaging technologies, such as Landsat or Sentinel, hyperspectral sensors have a lower spatial resolution, which means that objects on the ground are often blurred and appear as larger pixels. This can make it difficult to accurately identify and distinguish between different objects and materials on the ground. However, this high spectral resolution comes at the cost of lower spatial resolution, which is the ability to distinguish fine features and details in the image. The spatial resolution of a hyperspectral image is determined by the size of the pixels, which can be affected by various factors such as the altitude of the platform, the field of view of the sensor, and the available bandwidth. Low spatial resolution in hyperspectral images can have significant impacts on the accuracy and reliability of the derived information, particularly in applications that require high spatial accuracy, such as urban planning, land cover mapping, and crop monitoring. This paper will discuss some of the effects of low spatial resolution in hyperspectral images, including:

(1) Blurring of features: Low spatial resolution can cause features in the image to become blurred, making it difficult to distinguish between different objects or surfaces. This can lead to errors in identifying and classifying different materials, especially in complex and densely populated areas.

(2) Loss of fine details: Hyperspectral images with low spatial resolution can miss important details that are crucial for accurate analysis and interpretation, such as small structures, vegetation patterns, and subtle changes in land use. This can limit the ability to detect and monitor changes over time, and to identify potential problems and opportunities.

(3) Reduced accuracy of quantitative analysis: Low spatial resolution can also affect the accuracy of quantitative analysis, such as calculating the surface reflectance, emissivity, or radiance of a target. The spectral signature of an object can be altered by the mixing of signals from adjacent pixels, resulting in biased or uncertain results.

(4) Increased noise and data reduction: To reduce the amount of data that needs to be transmitted and stored, hyperspectral images are often resampled to lower

spatial resolution, which can result in increased noise and loss of information. This can also make it more difficult to detect and remove artifacts from the image, such as noise, interference, or atmospheric effects.

In conclusion, hyperspectral images provide a wealth of information about the physcial object and are a valuable tool for scientific research and environmental monitoring. However, their high cost, atmospheric correction limitations, and large data size present challenges that need to be considered while analyzing and interpreting the data. Advances in hyperspectral imaging technology and data processing algorithms are expected to overcome these challenges in the near future. Despite these limitations, low spatial resolution in hyperspectral images can have most significant impacts on the quality and usefulness of the derived information. Therefore, it is important to carefully consider the spatial resolution requirements of each application, and to use appropriate techniques and algorithms to overcome the limitations of low spatial resolution. We will disscuss about the enhancement methods to improve the spatial resolution of hypestral images in the next section.

1.4 Resolution Enhancement of Hyperspectral Images

Resolution enhancement of hyperspectral images refers to the process of improving the spatial resolution of a hyperspectral image. Hyperspectral images are images captured in multiple wavelength bands, and they are typically of lower spatial resolution compared to traditional color images. Resolution enhancement techniques aim to improve the spatial detail of the image by increasing its resolution. There are several methods for enhancing the resolution of hyperspectral images, including deconvolution techniques, and feature-based methods, super-resolution algorithms. In this dissertation, we mainly focus on super-resolution algorithms.

In a word, the goal of resolution enhancement is to improve the accuracy and detail of the information that can be obtained from a hyperspectral image. This information can be used in a variety of applications, including remote sensing, agriculture, and environmental monitoring.

1.4.1 Spatial Resolution Enhancement

Spatial resolution enhancement of hyperspectral images is a process of increasing the level of detail and sharpness in the spatial dimension of hyperspectral images. This can be achieved using various techniques that aim to either upsample or super-resolve the spatial information in the hyperspectral data. Some common techniques are used for spatial resolution enhancement of hyperspectral images, such as interpolation [6] and multi-resolution analysis [7]. For interpolation, this technique involves increasing the resolution of an image by filling in the missing values between the existing pixels. Interpolation techniques include bicubic, bilinear, and nearest-neighbor methods [8]. Multi-resolution analysis involves decomposing the hyperspectral image into multiple levels of spatial frequency components. These components are then processed to enhance the spatial resolution of the image. Overall, the choice of spatial resolution enhancement technique depends on the specific application and the characteristics of the hyperspectral data being processed. It is important to carefully evaluate the results of any spatial resolution enhancement technique to ensure that the spectral information of the hyperspectral data is not compromised during the process.

1.4.2 Spectral Resolution Enhancement

Spectral resolution enhancement of hyperspectral images is the process of increasing the spectral resolution of an image, which involves obtaining additional spectral bands between the original spectral bands of the image. The spectral resolution is determined by the number of spectral bands in the image, and the narrower the spectral bands, the higher the spectral resolution. There are several methods for spectral resolution enhancement of hyperspectral images, including Interpolation [9], band sharpening [10], spectral unmixing [11], hyperspectral image fusion [12]. Interpolation method [13] involves using interpolation techniques to estimate the values of additional spectral bands between the original spectral bands. The most commonly used interpolation techniques for spectral resolution enhancement are linear interpolation, cubic interpolation, and spline interpolation. Band sharpening method [14] involves using a high-resolution multispectral image to sharpen the spectral bands of a lower-resolution hyperspectral image. The multispectral image is used to estimate the high-frequency information in the hyperspectral image, which is then used to sharpen the spectral bands. Spectral unmixing method [11] involves decomposing the hyperspectral image into its constituent spectral signatures and then using these signatures to estimate the additional spectral bands. Spectral unmixing is based on the assumption that the hyperspectral image is a linear mixture of the constituent spectral signatures. Hyperspectral image fusion method [15] involves fusing multiple hyperspectral images with different spectral resolutions to obtain an image with higher spectral resolution. The fusion can be performed in the spectral domain, spatial domain, or both. It is important to note that spectral resolution enhancement can increase the information content of the image, but it can also increase the noise and artifacts in the image. Therefore, it is important to carefully choose the appropriate method and to tune its parameters to obtain the best possible results.

Feature-based methods are another class of techniques for spectral resolution enhancement of hyperspectral images that exploit the statistical dependencies between the spectral and spatial features of the image. These methods aim to recover the high-frequency information in the spectral domain that is lost due to the limited spectral resolution of the imaging system. The most commonly used feature-based methods for spectral resolution enhancement of hyperspectral images are various. Sparse representation-based method [16] involves representing the hyperspectral image as a sparse linear combination of a small number of basis vectors. The basis vectors capture the spectral and spatial features of the image. The method aims to recover the high-frequency information in the spectral domain by solving an optimization problem that promotes sparsity in the representation. Non-local means-based method [17] involves exploiting the redundancy in the spatial and spectral features of the image to estimate the missing high-frequency information in the spectral domain. The method uses a patch-based approach to estimate the spectral values of the missing high-frequency bands. Principal component analysis-based method [18] involves projecting the hyperspectral image onto a lower-dimensional space that captures the most significant spectral and spatial features of the image. The method aims to recover the high-frequency information in the spectral domain by projecting the image back onto the high-dimensional space. Feature-based methods can provide good results, but they require careful tuning of parameters and a large amount of training data for deep learning-based methods. The choice of the appropriate method and parameters depends on the specific application and the characteristics of the hyperspectral data.

1.4.3 Fusion Enhancement

Fusion enhancement is a technique for improving the quality and information content of hyperspectral images by combining them with other types of data, such as multispectral or panchromatic images. The basic idea behind fusion enhancement is to leverage the complementary information provided by different types of data to improve the spatial and spectral resolution, enhance the contrast and texture, and reduce the noise and artifacts of the hyperspectral image. Multi-resolution analysis-based methods [19] involves decomposing the hyperspectral and other types of data into different frequency bands using multi-resolution analysis techniques such as wavelet or curvelet transforms. The method combines the information from the different frequency bands to enhance the spatial and spectral resolution of the hyperspectral image. Principal component analysis-based method [20] involves using principal component analysis (PCA) [21] to extract the most significant spatial and spectral features from the hyperspectral and other types of data. The method combines the principal components of the different data sources to enhance the contrast and texture of the hyperspectral image. Non-negative matrix factorization-based method [22] involves using non-negative matrix factorization (NMF) to extract the pure spectral components from the hyperspectral and other types of data. The method combines the spectral components of the different data sources to enhance the spectral resolution of the hyperspectral image. Deep learning-based method [23] involves using deep neural networks to learn the mapping between the hyperspectral and other types of data. The networks are trained using large amounts of labeled data to extract the complementary information from the different data sources and enhance the quality and information content of the hyperspectral image. However, fusion enhancement techniques should carefully select and preprocess the input data, as well as careful tune of parameters for the different methods.

Feature-based methods for fusion enhancement of hyperspectral images aim to exploit the complementary information of different data sources by selecting and combining relevant features from the input data. The basic idea behind feature-based methods is to extract the most informative and discriminative features from each data source and use them to enhance the quality and information content of the hyperspectral image. Principal component analysis (PCA) based method [24] involves using PCA to extract the most significant spatial and spectral features from the hyperspectral and other types of data. The method combines the principal components of the different data sources to enhance the contrast and texture of the hyperspectral image. Independent component analysis (ICA) based method [25] involves using ICA to extract the statistically independent components from the hyperspectral and other types of data. The method combines the independent components of the different data sources to enhance the spectral and spatial resolution of the hyperspectral image. Sparse representation based method [26] involves representing the hyperspectral and other types of data as a sparse linear combination of some overcomplete dictionary, such as wavelet or curvelet transforms. The method selects the most relevant atoms from the dictionary and combines them to enhance the quality and information content of the hyperspectral image. Joint subspace learning based method [27] involves learning a common subspace that captures the shared and complementary information of the hyperspectral and other types of data. The method uses techniques such as canonical correlation analysis (CCA) or joint non-negative matrix factorization (JNMF) to learn the joint subspace and enhance the quality and information content of the hyperspectral image. But, the appropriate features and preprocessing of the input data, as well as careful tuning of parameters for the different methods are challenges in feature-based methods.

Super-resolution algorithms are commonly used in fusion enhancement methods that aim to increase the spatial or spectral resolution of hyperspectral images by fusing them with other types of data that have higher spatial or spectral resolution. The basic idea behind super-resolution algorithms is to use the high-resolution information of the other data sources to enhance the spatial or spectral resolution of the hyperspectral image. Multi-image super-resolution method [28] involves fusing multiple low-resolution images of the same scene, which are acquired from different viewpoints or at different times, to generate a high-resolution image. The method uses techniques such as patch-based methods, super-resolution convolutional neural networks, or deep learning-based methods to learn the mapping between the low-resolution and high-resolution images and generate a high-resolution image. Hyperspectral and panchromatic image fusion method [29] involves fusing the hyperspectral and panchromatic images, which have different spatial resolutions, to generate a high-resolution hyperspectral image. The method uses techniques such as intensity-hue-saturation (IHS) transformation, wavelet-based methods, or sparse representation-based methods to fuse the information of the two data sources and generate a high-resolution hyperspectral image. Hyperspectral and multispectral image fusion method [30] involves fusing the hyperspectral and multispectral images, which have different spectral resolutions, to generate a high-resolution hyperspectral image. The method uses techniques such as regression-based methods, tensor-based methods, or deep learning-based methods to learn the mapping between the two data sources and generate a high-resolution hyperspectral image.

1.5 Contributions of the Dissertation

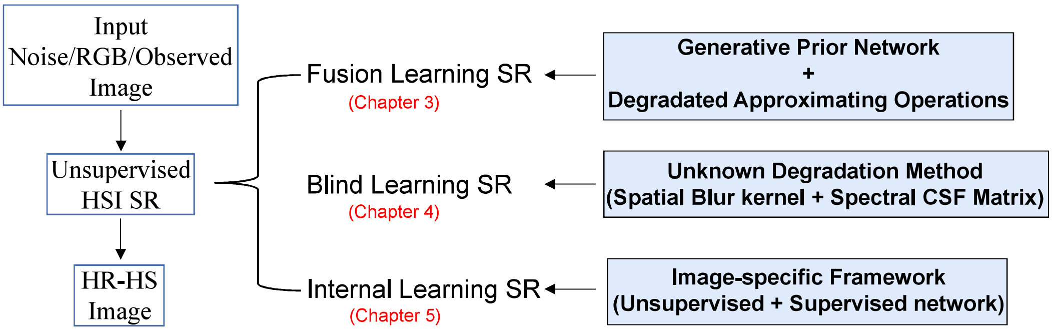


FIGURE 1.3: The contribution of this dissertation

First, we proposed a deep unsupervised fusion-learning method for HSI SR is proposed. In detail, we investigate a framework of hyperspectral image prior with one noise image or fusion context images as the network input for generating a latent HR-HS image using only the observed LR-HS and HR-RGB images without previous preparation of any other training triplets. Based on the fact that a convolutional neural network (CNN) architecture is capable of capturing a large number of low-level statistics (priors) of images, our deep unsupervised fusion-learning method promote the automatic learning of underlying priors of spatial structures and spectral attributes in a latent HR-HS image using only its corresponding degraded observations. Specifically, we investigated the parameter space of a generative neural network used for learning the required HR-HS image to minimize the reconstruction errors of the observations using mathematical relations between data. Moreover, special convolutional layers for approximating the degradation operations between observations and the latent HR-HS image are specifically to construct an end-to-end unsupervised learning framework for HS image super-resolution. The noise input leads to large space of parameter search, and lack of spatial and spectral information to generate a well reconstructed HR-HS image. To improve this, we next leverage the fused context of the observations for providing the insights of the specific spatial and spectral priors in the network learning.

Second, we proposed a (semi-)blind learning method for unsupervised HSI SR is proposed. The above deep unsupervised fusion-learning method for HSI SR uses the designed loss function formulated by the observed LR-HS and HR-RGB images only, and recovers the latent HR-HS image in a non-blind way with the known spatial degradation operation and spectral degradation CSF, which lacks of generalization in real scenario. Motivated by this improvement potential, we exploited a (semi-)blind learning method for unsupervised HSI SR, which is capable of reconstructing the HR-HS image from the observations not only with the known spatial and spectral degradation operations but also with the unknown spatial or spectral degradation operations or both unknown. In detail, the semi-blind learning can be divided in two ways, which are implemented by spatial blind only and the spectral blind only, respectively. For example, the unsupervised adaptation is capable of learning the spatial degradation operation of the observed LR-HS image but can only deal with the observed HR-HS image with known CSF, and thus it would be categorized as semi-blind paradigm for possibly learn the spatial degradation operations only in the observed LR-HS image. In addition, our proposed method can also be implemented in a complete blind setting (both unknown spatial down-sampling kernel for LR-HS image and the unknown CSF for HR-RGB image).

Finally, we proposed a generalized internal learning method for unsupervised HSI SR is proposed. Inspired by the fact that natural images have strong internal data repetition and the cross-scale internal recurrence, we employed a generalized internal learning method for unsupervised HSI SR, and aimed to learn an image-specific CNN model for each under-studying HR-HS data. With regard to naively adopting the internal spatial recurrence, the down-sampling operation on the observations usually causes severe spectral mixing of the surrounding pixels, and thus the deviation of the spectral mixing levels at the training phase and test phase would be great large. This domain shift in HSI SR possibly degrades the super-resolved performance in real experiments. To overcome the above limitations, we present a generalized internal learning method combined with self-supervised method for unsupervised HSI SR, which extracts the training triplets from the down-sampled versions of the observations and the LR-HS image to train a specific CNN model for the under-studying scene.

Chapter 2

Hyperspectral Image Super-Resolution

In hyperspectral (HS) imaging, three-dimensional cubic data with decades or hundreds of wavelength bands are captured. Each spatial point (pixel) contains a high dimensional vector for recording the light intensity at different wavelengths. Thus, the images acquired using the HS imaging technology contain not only abundant spatial structures but also detailed spectral signature and well suited to the substantial and high-performance analysis of imaged scenes. Having the advantages of detailed spectral distribution, the HS images were successfully used in various applications, such as remote sensing [2], food inspection [31–33], image classification [34–36] and object detection [37–39], and medicine [40–42], and are capable of achieving high-performance gain compared with other common RGB images. However, due to the radiant collection for each narrow-spectrum band in HS imaging sensors, less radiant energy per pixel and per spectral band measurement of an image scene is anticipated compared with the RGB imaging sensors. To ensure sufficient signal-to-noise ratio, the photo collection must be conducted in a much larger spatial region. This implies that the spatial resolution must be sacrificed to obtain detailed spectral information. Therefore, there is a trade-off between spatial and spectral resolution in real imaging sensors. This means that an HS sensor usually captures low spatial resolution and detailed spectral distribution (high spectral resolution) images. In contrast, common RGB sensors can provide much higher spatial resolution images but only with RGB color information. There are still some difficulties in acquiring high-resolution data in both spatial and spectral domains from commercial imaging sensors. Therefore, extensive research is necessary to fuse a low-resolution HS image (LR-HS) with the corresponding HR-RGB (multispectral) image for generating an HR-HS image using image processing and machine learning techniques. These fusion methods for generating HR-HS images are in general referred to as hyperspectral image super-resolution (HSI SR) methods [43].

Due to its ill-posed nature (the unknown number of variables in a latent HR-HS image is much larger than the known number of variables in the observations), multispectral and hyperspectral image fusion is a very challenging task. Most previously reported methods generally leverage various hand-crafted image priors to regularize the mathematical degradation model between observations and the latent HR-HS image. They also explore different optimization strategies to achieve the optimal solution. However, since the number of the unknown variables in the HSI-SR problem is much larger than the number of the known variables in the observed LR-HS and HR-RGB images, there are many solutions via directly solving the ill-posed problem with the formulated mathematical model. To narrow the solution space for providing a more plausible HR-HS image, a lot of previous work leverages the hand-crafted priors to characterize the underlying structure of latent HR-HS images and regularize the mathematical model [44]. The used hand-crafted priors in HSI SR scenario have evolved under modeling various aspects of the latent HR-HS image from the physical property of spectral signatures (etc. spectral unmixing model) [45], sparse representation [46], total variation [47] to similarity structure exploring [48] and have proven that it can achieve significant performance gain. Based on the physical properties of the observed spectrum in HS images, one research direction is the investigation of effective representation of high-dimensional spectral vectors such as matrix factorization and spectral unmixing [49]. The spectral representation models were evolved from the fact that HS observations can generally be expressed as a weighted linear combination of the reflectance function basis and their corresponding fraction coefficients [50]. These models achieve acceptable reconstruction performance. Subsequently, based on the significant success of sparse representation in natural image processing, several research studies imposed a sparsity constraint on the spectral representation [51] as prior knowledge, and attempted to model the spatial structure and local spectral characteristics by automatically learning the spectral dictionary from the observed HR-RGB and LR-HS images. Additionally, based on possible low-dimensionality of the spectral space, a low-rank image prior technique was also adopted for exploring the intrinsic spectral correlation of a latent HR-HS image. This technique proved capable of reducing the spectral distortion to some extent [52]. Moreover, some recent research studies extensively exploited the similarity between image priors of global spatial structures and local spectral structures to further boost reconstruction performance [53, 54]. Although the integration of various hand-crafted image priors, such as physical spectral mixing, mathematical sparsity of spectral representation, low-rank property, and similarity resulted in significant progress regarding the HSI SR performance gain, discovering an optimal image prior for a specific scene is still an extremely difficult task due to the configuration and texture diversity in both the spatial and spectral domains. Different priors may be desired for the scenes with various characteristics, and to hammer out a proper prior to a specific scene remains to be an art.

Recently, the deep convolutional neural network (DCNN) has achieved remarkable success in various computer vision tasks. The successful application of deep convolutional neural networks (DCNN) in various computer vision tasks allowed HSI SR to rely on the DCNN's powerful learning capabilities [55] for robustly reconstructing a latent HR-HS image, and demonstrated impressive performance using various neural network architectures [56]. The HSI SR has manifested the DCNN scheme can effectively capture the intrinsic characteristics of the latent HS images. In contrast to the traditional optimization methods, the deep learning (DL) method is capable of automatically learning the underlying image priors in a latent HR-HS image instead of exploiting hand-crafted image priors. Using previously collected training samples consisting of LR-HS, HR-RGB images, and their corresponding HR-HS images, an HS image prediction model in the training phase can be constructed, and the corresponding HR-HS image can be efficiently estimated from its low-quality observations for learning optimal network parameters [57]. However, most of the current DL methods are implemented in a fully supervised manner. For optimal network parameter learning, large-scale training triplets must be collected in advance. However, this is a difficult task, especially in the HSI SR scenario due to the high-cost of capturing the HR-HS images. Moreover, the fully supervised DL paradigm usually suffers from insufficient generalization in real applications and separate HS image prediction models for different HS datasets must be learned. More recently, Ulyanov et al. [58] proposed a deep image prior (DIP) learning neural network and stated that a convolutional neural network itself is capable of capturing a large number of low-level image statistics for well reconstructing a natural image, which can be successfully applied to several image restoration tasks. Sidorov et al. [59] extended the idea of DIP into deep hyperspectral prior (DHP), and adopted denoising, inpainting and super-resolution for hyperspectral images. However, the DHPs only leverage the observed LR-HS image for network training and cannot efficiently learn both the spatial structure and spectral attribute priors for reconstructing a latent HR-HS image. As mentioned above, it is tough to obtain large-scale training samples in the HS image reconstruction scenario, especially the label samples of HR-HS images. On the one hand, self-supervised learning is a general learning framework via resorting to surrogate tasks that can be formulated using only unsupervised data. Self-supervised techniques have been used for various applications in a broad range of computer vision topics [60], and manifest feasibility in different vision tasks.

2.1 Mathematical Model-based Super-Resolution

In recent years, hyperspectral image reconstruction has been actively investigated in the computer vision and computational photography research community, and substantial improvement has been achieved. This work mainly focuses on hyperspectral image super resolution (HSI SR) by fusing the available LR-HS and HR-RGB images obtained from commercial imaging sensors. In this section, the related research work is briefly reviewed.

The HSI SR problem, following the fusion paradigm of the observed LR-HS and HR-RGB images (fusion-based HSI SR), is closely related to a multi-spectral (MS) image pan-sharpening task, where the goal is to merge an LR-MS image with its corresponding HR wide-band panchromatic image [61]. Numerous methods for MS pan-sharpening, mainly including multi-resolution analysis approaches [62] and component substitution methods [63], were proposed. The fusion-based HSI SR problem can be treated as several pan-sharpening sub-problems, where each band of the HR-MS (RGB) image can be considered to be a panchromatic image. However, this simplification cannot fully use the spectral correlation and usually leads to significant spectral distortion in the recovered HR-HS image. In remote sensing, pan-sharping techniques aim to generate an HR multispectral (HR-MS) image via fusing an LR multispectral (LR-MS) image with an HR panchromatic image [64], which is closely related to our investigated HSI SR problem solution. The HS/RGB image fusion problem can follow the simple pan-sharping method by treating it as several pan-sharpening sub-problem which means that each color (R, G, B) band of the HR-RGB image acts as a panchromatic image. Although many pan-sharping methods have been evaluated, this heuristic approach dramatically suffers from high spectral distortion due to the insufficient spectral information in a single panchromatic image.

Mathematical model-based methods of hyperspectral super-resolution use mathematical models to generate high-resolution hyperspectral images from low-resolution images. These methods require a prior knowledge of the hyperspectral data. In mathematical model-based methods of hyperspectral super-resolution, handcrafted priors refer to the use of prior knowledge about the characteristics of hyperspectral images to guide the process of generating a high-resolution image. These priors are often built into the mathematical model used in the super-resolution process. One of the most commonly used handcrafted priors is sparsity, which assumes that the hyperspectral image has a sparse representation in some domain, such as wavelet or Fourier domain. This prior is often used in sparse representation-based methods of hyperspectral super-resolution. The sparsity prior assumes that only a few basis vectors from the dictionary are required to represent the hyperspectral image accurately. This prior can be used to constrain the optimization problem and improve the quality of the generated high-resolution image. Another commonly used handcrafted prior is the smoothness prior, which assumes that the high-resolution hyperspectral image has a certain smoothness property. This prior is often used in regularization-based methods of hyperspectral super-resolution. The smoothness prior can be used to constrain the optimization problem and generate a high-resolution image that is spatially smooth. The choice of the appropriate handcrafted prior depends on the specific application and the characteristics of the input data. Handcrafted priors are often used in combination with optimization techniques, such as convex optimization or non-negative matrix factorization, to generate a high-resolution image. It is important to note that the use of handcrafted priors can improve the quality of the generated high-resolution image, but they require a prior knowledge of the hyperspectral data and may not be suitable for all applications.

Recently, existing fusion-based HSI SR methods have widely leveraged the handcrafted image priors in a latent HR-HS image for robustly solving the inverse optimization problem and usually rely on physical and statistical models to exploit the correlations between different bands of the hyperspectral image and the spatial structure of the scene. The investigated image priors play a key role in obtaining a plausible solution in the optimization problem. The popularly used image priors are mainly used to explore the hidden knowledge in spatial and spectral representation such as physical spectral mixing, sparsity, low-rank, and similarity [65]. Since the non-negative physical property of the materials in the scene plays a role in representing prior for constraining the spectral coefficient, Wycoff *et al.* [66] proposed non-negative matrix factorization (NMF) to boost performance. Yokoya et al. [49] proposed to decompose the latent HR-HS image into a non-negative end-member matrix and an abundance matrix called negative matrix factorization (NMF). Then, they exploited a coupled NMF version (CNMF) to fuse a pair of HR-MS and LR-HS images, whereas Lanaras et al. [45] used a proximal alternating linearized-minimization method to optimize the coupled spectral unmixing model for HSI SR. Subsequently, the sparsity reguralized decomposition was extensively investigated by imposing sparse constraints on the abundance matrix [67]. Akhtar *et al.* [68] proposed a sparse spatio-spectral representation via computing the sparse coefficients of all pixels in a local grid region based on the part of selected spectral atoms, and further explored a Bayesian dictionary learning and sparse coding algorithm for pursuing better performance. Dong *et al.* [46] investigated a non-negative structured sparse representation via imposing structure similarity of the latent HR-HS image as constraints on sparse coefficient estimation. Grohnfeldt et al. [69] employed a joint sparse representation for separately modeling the spatial structure (local patch) in each individual band image. Based on the inherent low-dimensionality of the spectral space and the 3D structure in a latent HR-HS image, tensor factorization and low-rank image priors were actively integrated for the HSI SR problem [70], and the feasibility of the reconstructed HR-HS image was demonstrated. Most recently, Han *et al.* [48] further combined the local spectral and global structure self-similarity constraints into a sparsity-promoted model and validated promising performance. All the above methods constructed the observation model for formulating the degradation process of the available LR-HS and HR-RGB images. They attempted to leverage the hand-crafted priors such as the negative physical property of the materials, spectral mixing characteristic, sparsity in representation, and similarity structure to pursue performance boosting. However, it remains to be an art to discover a proper prior for a specific scene.

2.2 Deep Learning-based Super-Resolution

Based on the successful application of deep convolutional neural network in nature RGB image super-resolution, deep learning was also investigated for fusion-based HSI SR tasks. Instead of exploiting the hand-crafted image priors, in this method, the inherent image priors hidden in a latent HR-HS image are automatically learned, and a superior reconstruction performance can be achieved. The current fusion-based HSI SR employing deep learning is mainly divided into the fully supervised learning method and the unsupervised learning method. Meanwhile, resolution enhancement CNN of hyperspectral images depends on the characteristics of the image and the desired result, which enables the generation of high spatial and high spectral resolution images from any of the following observational data: (1) low resolution hyperspectral images, such as LR HS images; (2) high resolution multispectral images, such as HR RGB images; and (3) a combination of low resolution hyperspectral images and high resolution multispectral images. Depending on the type of observational data, resolution enhancement CNN frameworks for hyperspectral images can be divided into three categories: spatial-CNN, spectral-CNN, and fusion-CNN.

Motivated by the tremendous success of the DCNN on different vision tasks, DCNN based methods have been proposed for the HSI SR task, in which handcrafted priors exploring is no longer required. To train the HS image prediction model, it is necessary to previously collect the training triplets, including the LR-HS, HR-RGB images, and corresponding ground-truth (label), i.e., the HR-HS images in the fully supervised method, and provided an elaborate design for fusing multiple modalities of observations with different spatial and spectral structures. Han et al. [71] conducted an initial investigation by directly inputting the fused data of an HR-RGB image and an up-sampled LR-HS image to a simple 3-layer CNN, and used a more complex CNN with a residual structure to achieve better performance [72]. Palsson et al. [73] explored a 3D CNN-based MS/HS fusion scheme, and integrated the principal component analysis (PCA) to reduce the computational cost. Dian et al. [57] proposed a multi-stage method, and used a simple optimization strategy for initial HS reconstruction and final refinement, while adopting a 20-layer CNN for learning a latent HR-HS image from the initial one. More recently, Wang et al. [74] exploited an efficient hyperspectral image fusion network by iteratively integrating the representation relations between the target and observations into the deep-learning network to achieve superior performance. Han *et al.* [55] further investigated a multi-level and multi-scale spatial and spectral fusion network for effectively merging the available LR-HS and HR-RGB images with a massive difference in spatial structure. Xie *et al.* [75] investigated a MS/HS fusion network via leveraging the observation models of low-resolution images and the spectral low-rankness knowledge of HR-HS image, and exploited a proximal gradient strategy for solving the proposed MS/HS fusion framework. Moreover, Zhu *et al.* [76] explored a lightweight deep neural network-based framework, namely progressive zero-centric residual network (PZRes-Net), to pursue the efficiency and effectiveness of the HS image reconstruction problem. Despite the high performance gain, these methods fail to generalize well among different datasets, and they need to separately train the reconstruction models regarding the datasets under investigation, even for small imaging condition changes. Although the reconstruction performance has remarkably progressed, all the above DCNN based methods are required to be trained with a large number of training samples previously prepared, which are the set of triplets consisting of not only the LR-HS and HR-RGB images but also the corresponding HR-HS images as labels.

Deep unsupervised learning-based methods of hyperspectral image super-resolution are an emerging approach that aims to learn the underlying distribution of the low-resolution and high-resolution hyperspectral images without requiring paired training data. Unlike supervised methods, unsupervised methods do not require a large dataset of paired low-resolution and high-resolution hyperspectral images for training. Instead, they use self-supervised learning or unsupervised learning techniques to extract information from the low-resolution image and generate the high-resolution image.

To deal with the generalization limitation in the fully supervised learning method, the unsupervised neural network was proposed as a good solution to the HSI SR problem. It is well known that the corresponding training triplets, especially the HR-HS images, are extremely hard to be collected in real applications. Thus, the quality and amount of the collected training triplets generally become the bottleneck of the DCNN based methods. Most recently, Qu et al. [77] attempted to solve the HSI super-resolution problem in an unsupervised way and designed an encoder-decoder architecture for exploiting the approximate low-rank prior structure of the spectral model in the latent HR-HS image. This unsupervised framework did not require any training samples in an HSI dataset and could restore the HR-HS image using a CNN-based end-to-end network. However, this method needed to be carefully optimized step-by-step in an alternating way, and the HS image recovery performance was still not enough. Based on the deep image prior (DIP) and the fact that the convolutional neural network itself can capture a large number of low-level image statistics to achieve a well-reconstructed natural image, Sidorov et al. [59] extended the DIP concept for automatically learning the underlying priors for HS images (DHPs) and applied it to the spatial resolution enhancement of an hyperspectral image. However, the DHPs can only leverage the observed LR-HS image for network training and cannot efficiently learn both the spatial structure and spectral attribute image priors for reconstructing a latent HR-HS image. Furthermore, Zhang et al. [78] leveraged the generated training triplets (the LR-HS, HR-RGB, and HR-HS images) with different degradation models to learn a common deep model for predicting an initial HR-HS image, and then exploited unsupervised adaptation learning for fine-turning the initial estimation and automatically learning the degradation operations of the under-studying observations. Although remarkable performance gain has been achieved with different degradation models compared with most state-of-the-art methods, the performance of the fine-turning HR-HS image in the adaptation learning is greatly affected by the initially estimation in the common model, and is easy to fall into a local minimum solution. In addition, Nie et al. [79] proposed two steps of learning method, where the spatial and spectral degradation models were first predicted via modeling the relation between the HR-HS image and the observations: LR-HS and HR-RGB images, and then the latent HR-HS image is reconstructed with the previously estimated degradation models. Liu et al. [80] proposed an unsupervised multispectral and hyperspectral image fusion (UnMHF) network using the observations of the under-studying scene only, which estimates the latent HR-HS image with the learned encoder-decoder-based generative network from a noise input and can only be adopted to the observed

LR-HS and HR-RGB image with known spatial down-sampling operation and camera spectral function (CSF). Later, Uezato et al. [81] exploited the similar method for unsupervised image pair fusion, dubbed as guided deep decoder (GDD) network for the known spatial and spectral degradation operation only. Thus, the UnMHF [80] and GDD [81] can be categorized into the non-blind paradigm, and lack of generalization in real scenario. Zhang et al. [82] proposed two-steps of learning methods via modeling the common priors of the HR-HS image in a supervised way and then adapted to the under-studying scene for modeling its specific prior in an unsupervised manner. Although the unsupervised fine-tuning for a specific target scene is possible to be conducted, it still required the common model learning in a fully-supervised manner with large amount of prepared training triplets, and the adaptation performance greatly depended on the pre-estimated HR-HS image with the learned common model. In addition, the unsupervised adaptation is capable of learning the spatial degradation operation of the observed LR-HS image but can only deal with the observed HR-HS image with known CSF, and thus it would be categorized as semi-blind paradigm for possibly learn the spatial degradation operations only in the observed LR-HS image. Moreover, Fu et al. [83] aimed to select/learn an optimal CSF and then designed a 3-band sensor system capturing an image for a scene, which is capable of leading to a best reconstruction performance of a HR-HS. Given the captured image with this specially designed sensor, they exploited an unsupervised hyperspectral image super-resolution method using the designed loss function formulated by the observed LR-HS and HR-RGB images only, and recovered the latent HR-HS image in a non-blind way with the known spatial degradation operation and CSF. Further, since the unsupervised adaptation subnet in [82] and the method [83] utilizes the under-studying observed images only instead of the requirement of additional training samples for guiding the network training, they can also be dubbed as self-supervised learning strategy. However, these learning methods based on the under-studying observed images only are easy to drop into a local solution, and the final prediction heavily depends on the initial input of the network. Though, feasibility and potential of the HR-HS image super resolution with unsupervised strategy is verified, current methods usually explore different steps of learning for obtaining acceptable performance for this challenge unsupervised learning.

Deep unsupervised learning-based methods of hyperspectral image super-resolution are still a relatively new approach, and there is ongoing research in this area. One of the main advantages of unsupervised learning-based methods is that they do not require a large amount of labeled data, which can be difficult to obtain in some applications. However, they can be more challenging to train compared to supervised methods and may require additional regularization techniques to avoid overfitting.

2.3 Related Work

2.3.1 Deep Image Prior

Deep Image Prior (DIP) [58] is a recently proposed deep learning approach for image restoration tasks, including image denoising, inpainting, and super-resolution. The key idea behind DIP is to use a deep neural network as a prior for image restoration, rather than relying on handcrafted priors or models. DIP works by training a deep convolutional neural network (CNN) to map a random noise vector to an output image. The network is trained in an unsupervised manner, meaning that it does not require a dataset of paired low- and high-resolution images for training.

Instead, the network is trained to minimize the difference between the output image and a corrupted input image, without any explicit regularization or constraint on the network weights. During the testing phase, DIP is used to restore a corrupted image by initializing the network with a random noise vector and then optimizing the network weights to minimize the difference between the restored image and the corrupted image. This optimization process is typically done using a gradient descent algorithm. The key advantage of DIP is that it can achieve state-of-the-art performance on image restoration tasks, without requiring any training data or explicit prior knowledge. This makes DIP a powerful tool for image restoration in scenarios where training data is scarce or not available. However, the main limitation of DIP is that it can be computationally expensive and may require significant computational resources to run on large images or high-resolution hyperspectral data. In addition, DIP may not always provide the same level of interpretability or control as traditional handcrafted prior-based or model-based methods, as the underlying network architecture and optimization process can be difficult to interpret.

2.3.2 Internal Learning

In deep learning, external data typically refers to data that is not part of the primary dataset being used to train a model. This can include additional datasets that are used to augment the training data or provide additional information to the model, as well as external resources such as pre-trained models, image databases, or text corpora. The use of external data can be especially beneficial in cases where the primary dataset is limited in size or scope, or where the model requires additional context or information to perform well on a particular task. For example, in image classification, a model might be trained on a small dataset of images, but additional data from other sources can be used to augment the training data and help the model learn more robust and generalizable features. External data can also be used to address issues such as class imbalance or bias in the primary dataset. For example, in medical imaging, external data from other hospitals or research studies can be used to balance the representation of different diseases or conditions in the training data. However, the use of external data in deep learning can also raise concerns around generalization and overfitting. It is important to carefully evaluate the quality and relevance of external data, and to ensure that it is used appropriately and ethically.

On the other hand, internal data in deep learning is used for training and evaluation of the model, and is typically stored in memory or on disk during the training process. This includes the input data (such as images, audio, or text) as well as the corresponding labels or targets. The use of internal data in deep learning is central to the process of training a model to make accurate predictions on new, unseen data. During training, the model is repeatedly exposed to batches of internal data, and adjusts its parameters to minimize the difference between its predicted output and the true target value. Internal data is typically partitioned into separate sets for training, validation, and testing. The training set is used to update the model's parameters, while the validation set is used to monitor the model's performance and adjust hyperparameters such as learning rate or regularization strength. The test set is used to evaluate the final performance of the model on new, unseen data. The quality and representativeness of the internal data is critical to the performance of a deep learning model. If the training data is too limited, noisy, or biased, the model may fail to generalize well to new data, or may overfit to the training set. It is therefore important to carefully curate and preprocess the internal data to ensure that it is representative of the target population, and to use appropriate techniques such as data augmentation or regularization to improve the model's robustness and generalization ability.

In conclusion, the main difference between internal data and external data is that internal data is used to directly train and evaluate the model, while external data is used to support or improve the training process. Internal data is typically carefully curated and preprocessed to ensure its quality and representativeness, while external data may be less controlled and may require additional processing or filtering to be useful for deep learning. In general, the quality and relevance of the external data can have a significant impact on the performance of the model, and its integration with the internal data and training process is a key consideration in the design of a successful deep learning system.

Based on internal data, internal learning can learn a system or model from its own internal representations or activities, rather than from external input or feedback. In other words, the system is able to generate its own training data and learn from it, rather than relying solely on external data. Internal learning is commonly used in the context of artificial neural networks, where the network is trained to learn a task based on a set of input-output pairs. However, internal learning can also be used to improve the network's performance and generalization ability by allowing the network to learn from its own internal representations, which can capture higher-level features and relationships in the data. One example of internal learning in neural networks is unsupervised learning, where the network is trained to identify patterns or structure in the data without any external labels or feedback. In this case, the network is able to learn from its own internal representations of the data, such as the activation patterns of hidden units or the clustering of data points in feature space. Internal learning can be applied to various types of data, including natural images. In the context of natural image processing, internal learning refers to the process of a model learning from its own internal representations or features of natural images, rather than relying solely on external input or feedback. This can be achieved using techniques such as autoencoders. In the case of autoencoders, the model is trained to encode natural images into a lower-dimensional representation, and then decode the representation back into an image that is as close as possible to the original. This process encourages the model to learn useful and meaningful features of natural images that can be used for tasks such as image classification or segmentation. Another example of internal learning for natural images is self-supervised learning, where a model is trained to predict certain properties or relationships in the data based on its own internal representations. For example, a model can be trained to predict the rotation or color of a natural image based on its internal features. This process encourages the model to learn useful and meaningful features of natural images that can be used for downstream tasks such as object detection or semantic segmentation. Internal learning can also be combined with external feedback, such as in the case of reinforcement learning for natural image processing. In this case, a model is trained to perform a specific task, such as image classification or object detection, based on a reward signal that is generated by the environment. The model can also learn from its own internal representations and features of natural images to improve its performance on the task.Overall, internal learning is a powerful approach to natural image processing that can enable models to learn useful and meaningful features and representations from natural images, without relying solely on external input or feedback. But, internal learning can also be challenging, as it requires careful design and tuning of the network architecture, training algorithm, and hyperparameters to ensure that the network is able to learn useful and meaningful representations from its own internal activities.

2.3.3 Self-Supervised Learning

Self-supervised learning is a type of machine learning where a model is trained to solve a task without the use of explicit supervision or labels. Instead, the model learns to extract useful features from the input data, which can then be used for a variety of downstream tasks. In self-supervised learning, the input data itself serves as the supervisory signal, and the goal of the model is to learn to predict some aspect of the data that is not provided explicitly as a label. This can be done in various ways, such as by removing part of the input and asking the model to reconstruct it, or by asking the model to predict the order of shuffled input sequences. Self-supervised learning is a promising approach for training deep neural networks on large amounts of unlabelled data, which is often readily available, in order to learn useful features that can be transferred to other tasks. By contrast, traditional supervised learning typically requires large amounts of labeled data, which can be expensive and time-consuming to obtain, and may not always be available. Self-supervised learning has been shown to be effective in a range of applications, including computer vision, natural language processing, and speech recognition. Some examples of self-supervised learning methods include contrastive predictive coding, denoising autoencoders. Self-supervised learning methods have been applied to hyperspectral image super-resolution, with the goal of learning to reconstruct high-resolution hyperspectral images from low-resolution inputs in an unsupervised manner.

Specifically, self-supervised learning is a general learning framework via resorting to surrogate (pretext) tasks that can be formulated using only un-annotated data. In general, the pretext task is usually designed to realize it relying on learning a proper image representation with large amount of handy images around the world. Self-supervised techniques have been used for various applications in a broad range of computer vision topics [84], and manifest state-of-the-art performance among the approaches for learning visual representation with the unsupervised images only. To date, the vision tasks generally adopted the self-supervised learning methods, in which similar samples with the desired data in target task are available for the self-supervised learning despite incomplete information and some side information such as mathematically transformed version. However, in the HSI SR scenario, there are scarce HR-HS images publicly released. It is tough to collect large amounts of HR-HS images even without the complete correspondence of the triplets for conducting the representation learning in a self-supervised way. Self-supervised learning methods that have been used for hyperspectral image super-resolution, for example, deep internal learning. Deep Internal Learning (DIL) is a self-supervised learning method that uses internal data as a supervisory signal. In DIL, the model is trained to predict a portion of the input data from another portion of the input data, without using any external labels. DIL has been shown to be effective for hyperspectral image super-resolution, as it can learn to extract useful features from the input data that can be used for image reconstruction.

Chapter 3

Unsupervised Learning of Hyperspectral Image Prior

In this chapter, we proposed a deep unsupervised image-specific generative framework of hyperspectral image super resolution for automatically generating a high-resolution HS image from its low-resolution HS and high-resolution RGB observations without any external sample. We incorporate the deep learned priors of the underlying structure in the latent HR-HS image with the mathematical model for formulating the degradation procedures of the observed LR-HS and HR-RGB observations, and introduce an unsupervised end-to-end deep prior learning network for robust HR-HS image recovery. Experiments on two benchmark datasets validated that the proposed method manifest very impressive performance, and even better than most state-of-the-art supervised learning approaches.

3.1 Problem formulation

Given the observed image pair (LR-HS image $\mathbf{X} \in \mathbf{R}^{w \times h \times L}$ and HR-RGB image $\mathbf{Y} \in \mathbf{R}^{W \times H \times 3}$), where w and h represent the width and height, the goal of HSI SR is to recover a HR-HS image $\mathbf{Z} \in \mathbf{R}^{W \times H \times L}$, where W and H represent the width and height of $\mathbf{Y}$ and $\mathbf{Z}$, and L is the number of spectral channels in the HR-HS image. Generally, the mathematical relation for formulating degradation operations between the observed images: $\mathbf{X}$, $\mathbf{Y}$ and the target HR-HS image $\mathbf{Z}$ can be expressed as follows:

$$\mathbf{X} = \mathbf{k}^{(\mathbf{Spa})} \otimes \mathbf{Z}^{(\mathbf{Spa})} \downarrow + \mathbf{n}, \mathbf{Y} = \mathbf{Z} * \mathbf{C}^{(\mathbf{Spec})} + \mathbf{n}, \tag{3.1}$$

where $\otimes$ denotes the 2D convolution operator, $\mathbf{k}^{(\mathbf{Spa})}$ represents the spatial 2D blur kernel, $(Spa) \downarrow$is the spatial decimation (down-sampling) operator, $\mathbf{C}^{(\mathbf{Spec})}$ is the spectral sensitivity function of the RGB camera (three 1D spectral filters) for converting the L spectral band to the RGB band, and $\mathbf{n}$ is the additive white Gaussian noise (AWGN) of the noise level **œ**. The mathematical expression of the above degradation models can be expressed in the following simplified matrix format:

$$\mathbf{X} = \mathbf{DBZ} + \mathbf{n}, \mathbf{Y} = \mathbf{ZC} + \mathbf{n}, \tag{3.2}$$

where $\mathbf{B}$ and $\mathbf{D}$ stand for the blurring matrix in the spatial domain and down-sampling matrix, respectively, for transforming $\mathbf{Z}$ to $\mathbf{X}$. $\mathbf{C}$ denotes the spectral sensitivity function (CSF) of an RGB sensor. Assuming that the degradation parameters $\mathbf{B}$, $\mathbf{D}$, and $\mathbf{C}$ (which can be obtained from the hardware design of the HS and RGB sensors) are known, a heuristic approach to intuitively minimize the following reconstruction errors using the observed **X** and **Y** for estimating **Z** is given as follows:

$$\mathbf{Z}^* = \arg\min_{\mathbf{Z}} \alpha\beta_1 ||\mathbf{X} - \mathbf{DBZ}||_F^2 + (1-\alpha)\beta_2 ||\mathbf{Y} - \mathbf{ZC}||_F^2, \tag{3.3}$$

where $|| \cdot ||_F$ stands for the Frobenium norm. Since the element numbers in the HR-RGB and LR-HS images are different, it generally needs to introduce the normalization weights such as $\beta_1 = 1/N_1$ and $\beta_2 = 1/N_2$, where N_1 and N_2 are the products of the pixel numbers and the spectral bands in LR-HS and HR-RGB images, respectively. Beside, we further exploit a hyperparameter α $(0 \leq \alpha \leq 1)$to adjust the contributions between these two reconstruction errors. Eq. 3.3 aims to obtain an optimal $\mathbf{Z}^*$ for minimizing the weighted reconstruction error of the observations. Using the assumed AWGN in Eq. 3.2, Eq. 3.3 is completely equivalent to maximize the likelihood of a latent HR-HS image given the observation of **X** and **Y**. It is known that in the HSI SR problem the total number of unknown variables in **Z** is much greater than the known variables in **X** and **Y**, and this constitutes a severely ill-posed problem. Recovering a robust HR-HS image based on the observations is an extremely difficult task. To overcome this problem, most existing methods explore various hand-crafted image priors to model the underlying structure of the HR-HS image, and then impose a regularization term on the reconstruction error minimization problem, which can be formulated as follows:

$$\mathbf{Z}^* = \arg\min_{\mathbf{Z}} \alpha\beta_1 ||\mathbf{X} - \mathbf{DBZ}||_F^2 + (1-\alpha)\beta_2 ||\mathbf{Y} - \mathbf{ZC}||_F^2 + \lambda\phi(\mathbf{Z}), \tag{3.4}$$

where $\phi(\mathbf{Z})$ is a term for modeling the underlying structure of **Z**, and λ represents a hyper-parameter, which balances the regularization term and reconstruction error distributions. By introducing the prior probability function $Pr(\mathbf{Z})$, where $\phi(\mathbf{Z}) = -\log(\mathbf{Pr}(\mathbf{Z}))$, Eq. 3.4 can be explained as the widely used maximum a posterior (MAP) framework. Although high performance gain can be achieved using various hand-crafted image prior in the HSI SR scenario, finding an appropriate image priors for a specific scene remains a challenging task. This work aims to use the powerful learning capability of deep-learning networks for automatically learning the underlying image priors in latent HR-HS images. Based on a DIP work, where the deep network architecture itself possesses a large number of low-level image statistics (image priors), a generative network for learning the spatial and spectral priors in a latent HR-HS image is employed. Then a reliable HR-HS image constrained by the learned priors using only the low-quality observations is reconstructed.

3.2 Motivation

This section introduces a deep unsupervised learning framework for the HR image fusion problem. We will firstly present the motivation from conventional supervised learning paradigm (common prior learning) to our unsupervised learning framework (specific prior learning), and then detail the proposed specific prior learning framework with un-supervision as well as the adopted network input and the learnable degradation module.

3.2.1 Common Prior Learning with Supervision

The recent deep learning-based HS image fusion methods manifest the DCNN scheme can effectively capture the intrinsic characteristics (common priors) of latent HS images in a fully-supervised learning manner using the previously prepared training samples (external dataset). Specifically, given N training triplets $(\mathbf{X_i}, \mathbf{Y_i}, \mathbf{Z_i})(\mathbf{i} = \mathbf{1}, \mathbf{2}, \cdots, \mathbf{N})$, the supervised deep learning methods aim to learn a common CNN model by minimizing the following loss function:

$$\theta^* = \arg\min_{\theta} \sum_{i}^{N} \|\mathbf{Z}_i - F_{CNN}(\mathbf{X}_i, \mathbf{Y}_i)\|^2 , \tag{3.5}$$

where F_{CNN} represents the transformation of the DCNN network with the to-be-learned parameters θ. It should be noted that these methods carry out the off-line training procedure for obtaining the optimal network parameters θ^* instead of directly searching the raw image space $\mathbf{Z}$, and can capture common priors hidden in the training samples by the powerful modeling capability of the DCNN. After learning the network parameters θ^*, the latent HR-HS image for any under-studying observations $(\mathbf{X}_t, \mathbf{Y}_t)$ can be easily reconstructed as: $\hat{\mathbf{Z}}_t = F_{CNN}^{\theta^*}(\mathbf{X}_t, \mathbf{Y}_t)$. Although the promising performance with these supervised deep learning methods has been achieved, it is mandatory to provide large-scale training triplets to learn a good model, including the LR-HS, HR-RGB, and HR-HS images which are especially hard to be collected in the HS image fusion scenario.

3.2.2 Specific Prior Learning with Un-supervision

In this section, the deep unsupervised fusion-learning framework for the HSI SR problem is introduced. Recent deep-learning-based HS reconstruction methods proved that a DCNN is capable of effectively capturing the underlying spatial and spectral structures (common prior information) of latent HS images and demonstrated promising performance. However, these methods are generally implemented in a fully supervised manner, and require large-scale training triplets containing LR-HS, HR-RGB, and HR-HS images, which are difficult to be specifically collected for obtaining the training labels (HR-HS images). Extensive research on natural image generation (DCGAN [85] and its variants) proved that high-definition and high-quality images with some defined characteristics and attributes can be successfully generated from a random noisy input without the supervision of high-quality ground-truth. This indicates that the inherent structure (priors) of a latent image with the defined characteristics can be captured by searching the neural network parameter space, starting from a random initial state. Moreover, DIPs [58] were exploited to model the more specific structures of an under-studying scene with the guidance of its degradation version only and successfully applied to several natural image restoration tasks such as image de-noising, impainting, and super-resolution. In this paper, this un-supervision paradigm is followed, aiming to learn the specific spatial and spectral structures (priors) of a latent HR-HS image with the guidance of its degraded observations (LR-HS and HR-RGB images). The conceptual scheme of the proposed image-specific generative network (ISGM) is illustrated in Figure 3.1. Specifically, a generative neural network G_θ (θ is the network parameter to be learned) is leveraged to model the underlying spatial and spectral structures of a latent HR-HS image $\mathbf{Z}$. By replacing $\mathbf{Z}$ with G_θ in Eq. 3.4 and removing the regularization term $\phi(\mathbf{Z})$ due to the automatically captured priors in the generative network, the fusion-based HSI SR model can be reformulated as follows:

$$\theta^* = \arg\min_{\theta} \alpha\beta_1 ||\mathbf{X} - \mathbf{DB}G_\theta(\mathbf{z_{in}})||_F^2 + (1-\alpha)\beta_2 ||\mathbf{Y} - G_\theta(\mathbf{z_{in}})\mathbf{C}||_F^2, \tag{3.6}$$

where z_{in} is the input of the generative neural network, and $G_\theta(z_{in})_i$ represents the $i-th$ element of the estimated HR-HS image. Instead of directly optimizing a raw HR-HS image, which is extremely large and not unique, Eq. 3.6 aims to search the parameter space of the generative neural network G_θ for leveraging the possessed priors in it. To solve the objective function using the searching parameter space of the generative neural network, the substantial architecture of the generative neural networkand a detailed description of the used input data are provided.

3.3 Architecture of the generative neural network

An arbitrary DCNN architecture can be adopted to serve as the generative neural network G_θ. Due to diverse information, such as salient structures, rich textures, and complex spectra in a latent HR-HS image, a generative neural network G_θ is required to provide sufficient modeling capabilities. Various generative neural networks, such as that in the adversarial learning scenario [Pix2pix and others], were already proposed and demonstrated their great potential to generate high-quality natural images [86]. In this work, an encoder-decoder architecture with its multi-level feature-learning property and simplification is leveraged, and the skip connections between the encoder and decoder paths are leveraged for feature reusing. A detailed generative neural network is illustrated in Figure 3.1 with different input data modalities. Both the encoder and decoder consist of 5 blocks, which can learn the representative features in different scales, and the outputs of all 5 blocks in the encoder side are transferred to the corresponding decoder with skip connection for reusing the extracted detailed features. Each block is composed of 3 convolutional layers, following the RELU activation function, where the max-pooling layer with a 2 × 2 kernel is used to reduce the feature map size between the encoder blocks, whereas the up-sampling layer is employed to doubly recover the feature map size between the decoder blocks. Finally, a convolutional output layer is adopted for estimating the latent HR-HS image. However, in the unsupervised learning setting, there is no ground-truth HR-HS image for guiding or evaluating the training states of the generative neural network. Then, the observed HR-RGB and LR-HS images are leveraged to construct the evaluation criterion described in Eq. 3.6.

Regarding the above generative network G_θ, any DCNN architecture is possibly employed to serve as our proposed framework. Since the latent HR-HS images usually contain large variety of contents such as salient structures, complicated spectra and rich textures, and thus the adopted generative network G_θ should have sufficient modeling capacity to provide reliable HR-HS recovery. Many generative architectures [87] have been investigated, and made significant success in generating high-quality natural images [88] by incorporating the advanced adversarial learning technique. While our unsupervised framework requires training specific-CNN model for each under-studying observation, and thus a shallow network is preferred to reduce the training time. Moreover, being widely known that the context exploitation in larger receptive field using deeper architecture can enhance the representation capability. Thus, a shallow network, which can explore the context in a larger receptive field, should be suitable to serve as our network architecture.

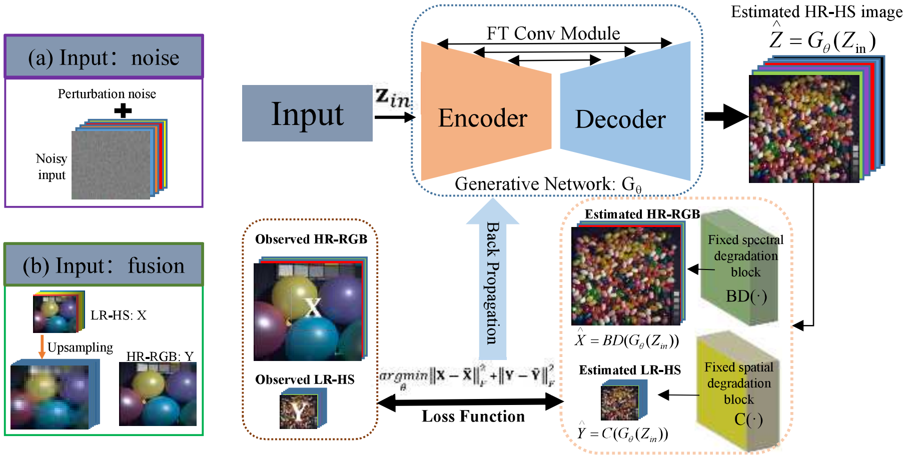


FIGURE 3.1: Conceptual diagram of the proposed image-specific generative network (ISGM) for unsupervised deep HSI SR.

It is well known that the encoder-decoder network has a shallow structure while enabling large-scale spatial context exploration due to the down-sampling operation between the adjacent scales, and then the encoder-decoder structure serves as the basis of our generative network. In order to learn the representative features for various situations, both the encoder and the decoder paths in our generative network are built of diverse scales of blocks. Using a convolution-based feature transfer module (FT Conv module), the outputs of the encoder blocks are transferred to the associated decoder blocks in order to reuse the derived detailed features. We specifically employ a point-wised convolution layer following the LeakyReLU and Batchnormalization layers to serve as the FT Conv module, which can make advantage in two ways compared with the simple skip connection: 1) reduce feature redundancy by only transferring the un-maintained features in the down-sampling stream to the decoder path; 2) increase the efficiency of the decoder path by reducing the channel number of the transfered feature map. In our experiment, since most information can be maintained in the down-sampling stream, we set the channel number of the transfered feature as 4 for lowering computational cost. In both encoder/decoder paths, the block at each scale includes 3 convolution/LeakyReLU pair layers, where the size of the feature map between the encoder blocks is cut in half by a max-pooling layer, and the size of the feature map between the decoder blocks is doubled by an up-sampling layer. Finally, the latent HR-HS image is predicted using a straightforward convolution-based reconstruction layer.

3.4 Input data to the generative neural network

There remain a difficulty in Eq. 3.6, which should be addressed for the HS image fusion problem, which is how to design the input of the generative network for capturing both low-level spatial statistics and the spectral correlation in the network training procedure. We classify the input data into two types, the first is an unconditional noisy input with a random perturbation added to check the robustness, corresponding to the image-specific generative network (ISGM (noise)) model; in particular, to contrast with the addition of random perturbation, we also perform experiments without random perturbation, i.e., the ISGM+ (noise) model. The second input data is a conditional fusion context of fused observations HR-RGB and

LR-HS, which corresponds to the ISGM (fusion) framework.

3.4.1 The Noise Input

The deep learning based methods such as DCGAN [85] and its variants verified that high-definition and high-quality images with a specific concept can be generated from a random noise, which means that to search the network parameter space from the initial random state can learn the inherent structure (prior) in the latent image of a specific concept. In addition, DIP [58] explored image prior possessing capability of network architecture for different restoration tasks of natural RGB images, and manifested impressive results. Therefore, a popular generative neural network can be trained to create a target image with the predefined characteristics, where randomly sampled noisy vectors based on a distribution function, such as Gaussian or uniform distribution, are usually used as inputs to ensure sufficient variety and diversity in the generated images. Motivated by this, we proposed a ISGM (noise) model. But in our HSI SR task, the corresponding HR-HS images of the observed degradation version (LR-HS and HR-RGB images) must be acquired. Therefore, an intuitive way is to adopt an initially sampled noisy vector z_{in}^{0} and fix it to search the optimal network parameter space for a specific HR-HS image. However, the unconditional fixed noisy input possibly leads to the generative neural network falling into a local minimum state. This results in an un-plausible estimation of the latent HR-HS image. Thus, a perturbation in the unconditional fixed initial input with a small randomly generated noisy vector at each training step is proposed to avoid a local minimum state. The input vector for the $i-th$ training step can be expressed as follows:

$$z_{in}^{i} = z_{in}^{0} + \beta \mathbf{n}_{i}, \tag{3.7}$$

where $\mathbf{n}_i$ is the randomly sampled noise vector at the $i-th$ training step, and β denotes the perturbation degree (a small scalar value). After learning the generative neural network G_θ with the perturbed input, a prediction using the fixed noisy vector $\mathbf{Z}^* = G_\theta(z_{in}^{0})$ as the final estimated HR-HS image is performed.

This image-specific generative network (ISGM (noise)) model employs noise vectors produced at random and sampled from a uniform distribution as input to provide low-level spatial statistics. But this research is less effective at identifying spectral and spatial correlations and is more challenging to optimize due to random noise vectors. We propose a solution to this issue in the next section. In the next part, we substitute observed LR-HS and HR-RGB images for entirely artificial noise.

3.4.2 The Fusion Context Input

The deep image prior (DIP) network [58] is designed to capture the low-level spatial statistics and use a randomly generated noise vector sampled from the uniform distribution as the input. However, with the randomly selected noise vector, the DIP is limited in capturing spectral and spatial correlation, and the optimization is more difficult. Meanwhile, the learning procedure is usually unstable without any provided knowledge about the spectral attribute and spatial structure. Hence, we improved the image-specific generative network (ISGM (noise)) model above by leveraging the observed LR-HS and HR-RGB images instead of the randomly generated noise as the network input.

Specifically, we proposed a ISGM (fusion) model, which also learns the network parameters exclusively using observed LR-HS and HR-RGB images. In the proposed DSSH framework, given the observed images: $\mathbf{X}$ which consists of the attribute of the hyper-spectra of the latent HR-HS image in spite of the low resolution in the spatial domain, and **Y** which possesses the high-resolution spatial structure despite the small number of spectral channels, we leverage the fused context of the observations for providing the insights of the specific spatial and spectral priors in the network learning. Specifically, we first transform the LR-HS image using an up-sampling layer to the same spatial size with the HR-RGB image and concatenate it with the HR-RGB image, which is expressed as

$$\mathbf{Z}_{\mathbf{in}}^{\mathbf{0}} = Stack(UP(\mathbf{X}), \mathbf{Y}). \tag{3.8}$$

Although we can directly use the simply fused context as the input of the generative network G_θ, it usually leads to converging to a local minimum. In this study, we add some perturbation to learn a more robust model for capturing the specific spatial and spectral priors and express it as

$$\boldsymbol{Z}_{in}^{i} = \mathbf{Z}_{in}^{\mathbf{0}} + \lambda \boldsymbol{\mu}, \tag{3.9}$$

where $\boldsymbol{\mu}$ denotes the randomly generated 3D tensor sampling from the uniform distribution and has the same size with the fused context Z_{in}^{0} while λ is a small scale value representing the perturbation degree. In all of our experiments, we initialize λ as 0.01 and half it in every 1000 steps of the training procedure. The perturbation is conducted in all training steps of the generative network G_θ.

Concerning the designed generative network G_θ, it is possible to employ any DCNN architecture for our proposed framework. The latent HR-HS images usually contain salient structures, rich textures, and complex spectra, which requires the generative network G_θ to have sufficient modeling capacity. There are multiple generative architectures such as in adversarial learning scenario [89] being proposed, and have manifested significant progress in generating high-quality natural images [90]. This study follows a simple encoder-decoder architecture with skip connections for integrating the features between symmetric blocks of encoder and decoder, which is shown in the backbone network part in Fig. 3.1 The output of the final convolutional layer is the latent HR-HS image. In our deep image-specific unsupervised learning setting, we do not have any information about the latent HR-HS image and cannot construct the criterion for evaluating the network state. Instead, we aim at resorting to the observed LR-HS and HR-RGB image for constructing the evaluation criterion.

3.4.3 The RGB Input

Despite the great potential for the real HSI SR scenario, the method above still lead to limited SR performance even for the well-registered input pair. This study aims to exploit an unsupervised strategy, and can be easily adapted to deal with the unregistered pair via incorporating the spatial transformation modules in the existing approaches. To this end, this work presents a new deep RGB-guided generating framework for unsupervised HSI SR by treating the observed HR-RGB image as the network input instead of the random noise. Specifically, we employ the observed HR-RGB image possessing high-resolution spatial structure as the network input to serve as the conditional guidance instead of a noisy input. Moreover, since our generative network mainly employs the 2D convolution operation, we adopt a specifically designed convolution layer to implement the spatial and spectral degradations, and thus obtain the estimated LR-HS and HR-RGB image from the output of the generative network to formulate the loss function via evaluating the reconstruction errors of the LR-HS and HR-RGB observations for the specific-network training, which requires no any external data in our proposed unsupervised framework. After training, the HR-RGB image with high-resolution spatial structure is fed into the specific generative model to predict the final HR-HS image. Experimental results on several public HS datasets demonstrate that the proposed method can obtain promising performance improvement compared to the state-of-the-art methods. We summarize the main advantages of this study as follows:

I. We elaborate the input to the generative network. We not only leverage the available observation as input but also as one term of reconstruction error for robust HR-HS image estimation.

II. We directly learn the specific model using the available observation without the requirement of any labeled training sample.

III. We devise simple specialized convolution layers to implement the degradation models, such as the modified depth and point convolution layers, which can be easily optimized together with the generative network.

Unlike the unsupervised learning networks that take the observed HR-RGB and LR-HS images as input pairs to the generative network, we resort to only a HR-RGB image as the input. In detail, given the predicted HR-HS image $\hat{\mathbf{Z}} = G_\theta(\cdot)$, where G_θ denotes the generative network and θ are its parameters, we specifically design a depth-wise convolution layer to implement the spatial degradation model: $f_{Spa}(\hat{\mathbf{Z}})$ and a point-wise convolution layer to carry out the spectral transformation $f_{Spe}(\hat{\mathbf{Z}})$, which are paralleled after the generative network. By simply fixing the weights of the specifically designed convolution layers using the blurring kernel in the spatial degradation matrix **DB** and the CSF matrix **C**, we can transform the output of the generative network to obtain the approximated versions of the HR-RGB and LR-HS images: **X** and **Y**. Concretely, we set the weights of the $k \times k$ kernel in each channel of the depth-wise convolution layer f_{Spa} with the 'False' bias term as the blurring kernel $\mathbf{B} \in \Re^{\mathbf{k} \times \mathbf{k}}$, and define the 'stride' parameter as the down-sampling factor $s = \frac{W}{w}$ to obtain a spatially degraded version $\hat{\mathbf{X}} \in \Re^{\mathbf{w} \times \mathbf{h} \times \mathbf{L}}$. Whilst the kernel weights $w_{Spe} \in \Re^{1 \times 1 \times L \times 3}$ of the point-wise convolution layer f_{Spe} with the 'False' bias term is fixed as the CSF matrix $\mathbf{C} \in \Re^{\mathbf{L} \times \mathbf{3}}$, which is decided according to the color camera for capturing the HR-RGB image, to produce a spectral-degraded version $\hat{\mathbf{Y}} \in \Re^{\mathbf{W} \times \mathbf{H} \times \mathbf{3}}$.

Then following Eq. 3.6, we minimize the prediction error of the LR-HS and HR-RGB observations to train the generative network. In Eq. 3.6, since the generative network G_θ has the powerful capability of automatically learning and modeling the priors in the latent HR-HS image, we do not explicitly impose any regularization term for prior modeling. The proposed deep RGB-guided generative framework is shown in Fig. 3.2, which can be trained with LR-HS and HR-RGB observations without requiring any external data. Next, we will instantiate the architecture of the generative network and the network input.

Most generative neural networks are used to generate a target image with the predefined characteristics from a noisy vector, which is often randomly sampled from a probability distribution, such as uniform or Gaussian distribution. As validated in the recent research, the randomly sampled noisy input usually enables that the generated images have enough variety and diversity. Our HSI SR task aims to learn a corresponding HR-HS image with the observed LR-HS and HR-RGB images. Simply taking a noise as the input cannot make full use of the available observations. Thus, we attempt to treat the available observation to serve as the conditional

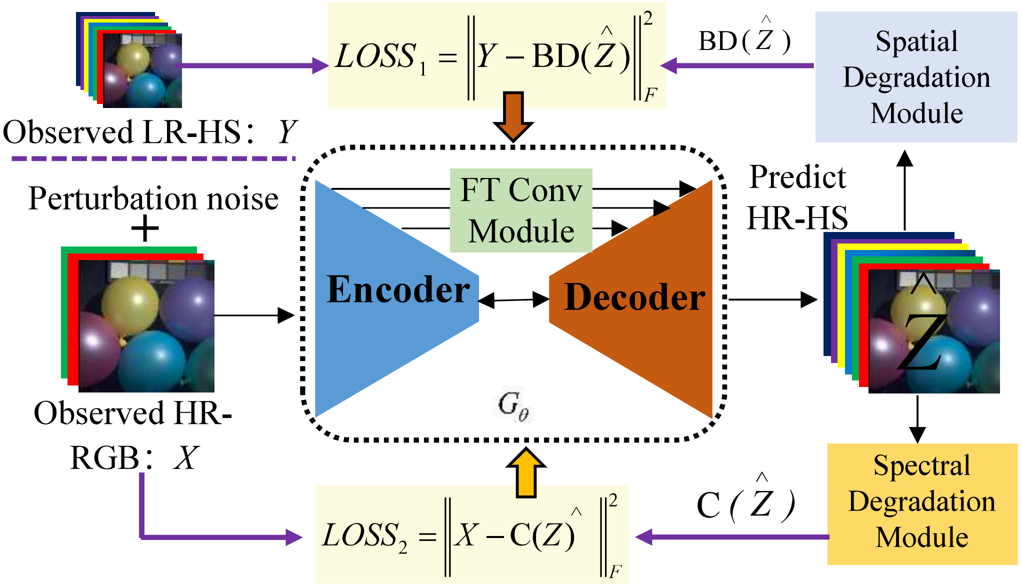


FIGURE 3.2: Proposed framework of deep RGB-guided image-specific generative network (ISGM (RGB)). FT Conv Module denotes feature transfer convolution module. Spatial Degradation module is implemented by a depth-wise convolution layer while Spectral Degradation module is conducted by a point-wise convolution layer.

guidance for our generative network. It is known that the observed HR-RGB image possesses a high spatial resolution structure, and is prospected to guide the 2D convolution-based generative network to learn a more reliable HR-HS image. It is also feasible to use the observed LR-HS image as the conditional input of the network. However, a low-resolution spatial structure would cause the network training procedure into some local minimum position, and thus yield adverse effects on the predicted results. What is more, the expanding factor in the spectral domain such as 10 for estimating 31 spectral bands from the RGB is usually much smaller than that of the spatial domain (a total 64 (8 × 8) for an up-sampling factor 8), and thus we adopt the observed HR-RGB image as the network input for conditional guidance, which is expressed as $\mathbf{Z}^* = G_\theta(\mathbf{Y})$.

In summary, the generative network for our HSI SR task has the following potential inputs with their respective advantages.
1. The randomly sampled noisy vector, which can maintain the diversity of the generated images;
2. The combined LR-HS and HR-RGB images: $f_{Concat}(\mathbf{X}\uparrow, \mathbf{Y})$ (conditional input) for leveraging both HR spatial structure in $\mathbf{Y}$ and spectral attribute in $\mathbf{X}$, where the great blurring in the LR-HS image would yield a baneful effect on the robust learning of the plausible spatial structure;
3. The HR-RGB image: $\mathbf{Y}$ (conditional input) for leveraging the HR spatial structure to recover the plausible HR-HS image;
4. The conditional inputs with small perturbation such as noise and dropout operation in the training phase to avoid dropping to a local minimum.
As mentioned above, our image-specific generative network (ISGM (RGB)) model employs the HR-RGB image: $\mathbf{Y}$ to guide the learning of generative network, and will provide a comprehensive comparison with different inputs.

3.5 Loss Function

We expect that the network output: $G_\theta(\cdot)$ (θ: network parameters) should approach the required HR-HS image: $\mathbf{Z}$. The goal of this work is to search the network parameter space to pursue a set of optimal parameters for satisfying the above criteria. However, due to the unknown $\mathbf{Z}$, it is impossible to construct quantitative criteria directly using $\mathbf{Z}$ for this task.

With the availability of the LR-HS and HR-RGB images, we turn to $\mathbf{X}$ and $\mathbf{Y}$ to formulate quantitative criteria (loss function) for network learning. Since, as in Eq. 3.2, the LR-HS image $\mathbf{X}$ is a blurred down-sampled version of $\mathbf{Z}$, and HR-RGB image $\mathbf{Y}$ is a transformed version of $\mathbf{Z}$ in channel direction using CSF: $\mathbf{C}$, we implement the two operations as two convolutional layers with pre-defined weights (non-trainable) after the output layer of the baseline hourglass-like network. The convolutional layer for blurring/down-sampling operator has the kernel size and stride according to the spatial expanding factor W/w and the kernel weights are pre-calculated according to Lanczos2 filter. The output of this layer is denoted as $f_{Spa}(G_\theta(\mathbf{n}))$, which has the same size and should be approximated to $\mathbf{X}$. Thus according to $\mathbf{X}$, the first loss function is formulated as:

$$L_1(\mathbf{n},\mathbf{X}) = \left\|\mathbf{X} - f_{Spa}(G_\theta(\mathbf{n}))\right\|_F^2 \tag{3.10}$$

While the spectral transformation operation (from $\mathbf{Z}$ to $\mathbf{Y}$) is implemented as the convolutional layer with 1×1 kernel size, input and output channels: L and 3, where the kernel weight is fixed as the CSF: $\mathbf{C}$ according to the used RGB camera. Then the output of this layer should be an optimal approximation of $\mathbf{Y}$. Denoting it as $f_{Spe}(G_\theta(\mathbf{n}))$, the second loss function is formulated as:

$$L_2(\mathbf{n},\mathbf{Y}) = \left\|\mathbf{Y} - f_{Spe}(G_\theta(\mathbf{n}))\right\|_F^2 \tag{3.11}$$

Via combining the L_1 and L_2 loss functions, we finally minimize the following total loss for searching a set of network parameter from the initialed random state:

$$L(\mathbf{n},\mathbf{X},\mathbf{Y}) = \arg\min_\theta L_1(\mathbf{n},\mathbf{X}) + L_2(\mathbf{n},\mathbf{Y}) \tag{3.12}$$

From Eq. (3.12), it can be seen the network is learned with the available observations only without any additional training samples. After completing training, the baseline network output: $G_\theta(\mathbf{n})$ is our required HR-HS image.

3.6 Experiment Results

In this section, we will conduct extensive compared experiments and perform ablation studies to demonstrate the effectiveness of our proposed deep unsupervised HS image reconstruction method.

3.6.1 Experimental Setting

Datasets

In this study, we conduct experiments using two benchmark HS image datasets consisting of CAVE [91] and Harvard [92]. The CAVE dataset contains 32 HS images captured indoors from real-world materials and objects. The HS images have the

(a) CAVE dataset

(b) Harvard dataset

FIGURE 3.3: Example images in the CAVE and Harvard datasets.

same spatial resolution, e.g., 512*512, and each HS image consists of 31 successive spectral bands ranging from 400 nm to 700 nm with a 10 nm interval. Harvard dataset has 50 HS images collected under daylight illumination, both outdoors and indoors. The HS images have spatial resolution 1392 × 1040, and each image contains 31 spectral bands captured from 420 nm to 720 nm with a 10 nm interval. For both datasets, in general case, we generate the corresponding RGB image via adopting the spectral response function of the Nikon D700 camera [93] for each HS image while obtaining the LR-HS images via Bicubic down-sampling the HS images. Since our proposed method is a deep unsupervised framework without requiring any training samples, we can conduct experiments on all images in both datasets to compare with the traditional optimization-based methods. We illustrate some example images of the CAVE and Harvard datasets in Figure 3.3.

Evaluation Metrics

To objectively evaluate the performance of the proposed method with different state-of-the-art methods, we employ five commonly used metrics, consisting of the root mean square error (RMSE), peak signal to noise ratio (PSNR), structural similarity index (SSIM), spectral angle mapper (SAM), and relative dimensional global errors (ERGAS). The RMSE, PSNR, and ERGAS measure the numerical difference between the recovered HR-HS image and the ground-truth image to evaluate the spatial fidelity. Simultaneously, the SAM provides the average spectral angle of the two spectral vectors from the same spatial positions of the recovered and ground-truth HS images for indicating the spectral fidelity. Further, the SSIM is used to measure the spatial structure similarity in two images. Generally, a smaller RMSE, ERGAR, or SAM and a larger PSNR or SSIM mean better performance. In all experiments, we first set the hyper-parameter α as 0.5 in the loss function of Eq. 3.6 to provide the compared results, and then adjust α from 0 to 1.0 with the interval 0.2 to evaluate the reconstruction performance in the ablation study.

Implementation Details

The proposed method was implemented in Pytorch. The input noise was initially set to the same dimension as the to-be-estimated HR-HS image. To train the generative network, the Adam optimizer [94] with the simple L_2 norm-based loss was adopted. The learning rate was initialized to 1e-3 and reduced by 0.7 every 1000 steps. Moreover, the perturbation degree β was initially set to 0.05 and reduced by 0.5 every 1000 steps. The optimization process was terminated after 12,000 steps and fixed for all images with different upscale factors from different datasets. All experiments were conducted in the learning environment with Tesla K80 GPU. In our experiment, the learning time for an image with size of 512 × 512 is about 20 min.

3.6.2 Performance Evaluation

In this section, we provide the compared results using our proposed method with state-of-the-art approaches, including traditional optimization-based and deep learning-based methods. Further, we evaluate the compared results of the proposed methods with different experimental settings, including the input of the generative network and the different values of the hyper-parameter.

Comparison with traditional optimization-based methods

Table 3.1 shows the compared quantitative evaluation of the super-resolved HS images generated by the ISGM (noise) model for an up-scale factors of 32 of the CAVE and Harvard datasets with different paradigms, including the traditional methods with manually engineered image priors (GOMP [95], MF [50], SNNMF [66]). Since an HR-HS image in the proposed method can be predicted by the generative neural network using the initial fixed noise w/o perturbation, two predictions with different inputs (denoted as ISGM (noise) and $ISGM + (noise)$) can be achieved. Although the proposed ISGM (noise) method can be robustly learned with the provided hyper-parameters mentioned in the previous section, it is not necessary to adjust these parameters for different datasets and up-scale factors. It should also be noted that it is possible to improve its performance to conduct hyper-parameter turning for different datasets and up-scale factors. The proposed method exhibits comparable results (without integrating any prior knowledge) with the SNNMF [66] method by leveraging the manually engineered image priors, which evolved and proved to be efficient by conducting a long-time research effort. Additionally, the proposed ISGM (noise) method aims to generate latent HR-HS images from the noise input and does not leverage any existing rich spectral information and spatial structure in the observed LR-HS and HR-RGB images for regularizing the generative neutral network learning. Thus, the HSI SR performance is expected to be improved by imposing some constraints on the network training using the observations. This is left for future investigation. A comparison between the results obtained using the tradational mathematical model-based methods in the CAVE dataset and the Harvard dataset (with up-scale factors of 8 and 16) is prensented in Table 3.2. In these tables, it can be observed that the proposed method achieves better or comparable results with those obtained using traditional optimization-based methods. In Tables 3.1 and 3.2, the up-arrow indicates better result with larger value while the down-arrow denotes better result with smaller value.

Meanwhile, we provide the compared results using our proposed ISGM (fusion) method with state-of-the-art approaches with traditional optimization-based methods. There are many optimization-based methods recently proposed for the HS image fusion, including Generalization of Simultaneous Orthogonal Matching Pursuit (G-SOMP+) method [96], Sparse Non-negative Matrix Factorization (SNNMF) method [97], Bayesian sparse representation (BSR) method [98], Couple Spectral

TABLE 3.1: Comparison results between state-of-the-art traditional optimization-based methods methods (GOMP [95], MF [50], and SNNMF [66]) and ISGM (noise) method in the CAVE and Harvard Datasets with the Expanding Factor: 32.

Spatial Expanding Factor = 32										
Dataset	Cave					Harvard				
Method	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↑	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↑
GOMP	6.47	32.48	-	14.19	0.77	4.08	38.02	-	4.79	0.41
MF	3.03	39.37	-	6.12	0.40	1.96	43.19	-	2.93	0.23
SNNMF	3.26	38.73	-	6.50	0.44	2.20	42.03	-	3.17	0.26
ISGM (noise)	3.47	38.17	0.951	8.31	0.46	2.82	40.12	0.957	3.96	0.43
ISGM+ (noise)	3.34	38.47	0.955	8.12	0.44	2.62	40.75	0.963	3.91	0.42

Unmixing (CSU) method [93], and Non-Negative Structured Sparse Representation (NSSR) method [46]. The traditional optimization-based methods generally employed various hand-crafted priors for reconstructing robust HS images. The degradation operations (spatial blurring/down-sampling and spectral transformation) should be known in all the methods. Our proposed unsupervised network attempts to learn the specific priors of the latent HR-HS image automatically. For a fair comparison, we first approximate the Bicubic degradation using Lanczos kernel for initializing the weight of the spatial degradation block and initializing the spectral transform block with the CSF of Nikon D700 camera and do not conduct learning for these block. We conducted experiments for the spatial expanding factors 8 and 16 for performance evaluation, and the compared results on both CAVE and Harvard datasets are shown in Table 3.2.

From Table 3.2, we can observe that our proposed ISGM (fusion) method achieves comparable performance with the NSSR method [46] and better results than all other methods for the CAVE dataset while manifests the best performance than all optimization-based methods for the Harvard dataset. Especially, in comparison with the best optimization-based CSU [93] method, our method can lift the PSNR 1.41dB and 1.66dB, respectively for the upscale factors 8 and 16 in the Harvard dataset. Moreover, from the compared results in Table 3.2, it can be seen that different optimization-based methods illustrate unstable performance trends. For example, the CSU method [93] shows better performance than the NSSR method [46] in the Harvard dataset while the NSSR method [46] significantly outperforms the CSU method in the CAVE dataset. The performance instability of the optimization-based methods may be caused by the adopted hand-crafted priors where proper priors should be designed according to the content and characterization of the under-studying images in different datasets. Although our proposed framework only leverages the degraded observations: the LR-HS and HR-RGB images without the requirement of an external dataset as in the optimization-based methods, it is able of learning the specific prior of an under-studying image using the powerful modeling capability of a deep generative network. Therefore, stable and plausible reconstruction results on both datasets have been achieved as shown in Table 3.2.

TABLE 3.2: Comparison results between state-of-the-art traditional optimization-based methods methods and ISGM (noise, fusion, RGB) and ISGM+ (noise) method in the CAVE and Harvard Datasets with the Expanding Factor: 8 and 16.

Spatial Expanding Factor = 8										
Dataset	CAVE					Harvard				
Method	RMSE↓	PSNR↑	SSIM ↑	SAM↓	ERGAS↓	RMSE↓	PSNR↑	SSIM ↑	SAM↓	ERGAS↓
GOMP	5.69	33.64	-	11.86	2.99	3.79	38.89	-	4.00	1.65
SNNMF	1.89	43.53	-	3.42	1.03	1.79	43.86	-	2.63	0.85
MF	2.34	41.83	-	3.88	1.26	1.83	43.74	-	2.66	0.87
BSR	1.75	44.15	-	3.31	0.97	1.71	44.51	-	2.51	0.84
CSU	2.56	40.74	0.985	5.44	1.45	1.40	46.86	0.993	1.77	0.77
NSSR	1.45	45.72	0.992	2.98	0.80	1.56	45.03	0.993	2.48	0.84
ISGM (noise)	2.08	42.50	0.975	5.36	1.16	2.38	42.16	0.965	2.35	1.09
ISGM+ (noise)	1.96	42.98	0.977	5.22	1.10	2.12	43.23	0.971	2.30	1.01
ISGM (fusion)	1.44	45.61	0.992	3.27	0.79	1.17	48.27	0.993	1.75	0.77
ISGM (RGB)	1.35	46.20	0.992	3.05	0.77	1.07	49.17	0.994	1.59	0.72
Spatial Expanding Factor = 16										
GOMP	6.08	32.96	-	12.60	1.43	3.85	38.56	-	4.16	0.77
SNNMF	2.45	42.21	-	4.61	0.66	1.93	43.31	-	2.85	0.45
MF	2.71	40.43	-	4.82	0.73	1.94	43.30	-	2.85	0.47
BSR	2.36	41.57	-	4.57	0.58	1.93	43.56	-	2.74	0.42
CSU	2.87	39.83	0.983	5.65	0.79	1.60	45.50	0.992	1.95	0.44
NSSR	1.78	44.01	0.990	3.59	0.49	1.65	44.51	0.993	2.48	0.41
ISGM (noise)	2.61	40.71	0.967	6.62	0.70	2.81	40.77	0.953	3.01	0.75
ISGM+ (noise)	2.50	41.03	0.969	6.43	0.67	2.56	41.66	0.959	2.95	0.71
ISGM (fusion)	1.76	43.84	0.999	3.76	0.49	1.32	47.16	0.992	1.99	0.47
ISGM (RGB)	1.71	44.15	0.990	3.63	0.48	1.28	47.37	0.992	1.92	0.49

Comparison with deep learning-based methods

Table 3.3 shows the compared quantitative evaluation of the super-resolved HS images generated by the ISGM (noise) model for an up-scale factors of 32 of the CAVE and Harvard datasets with different paradigms, including the unsupervised methods with learned priors (DHIP [59], uSDN [77]), and fully supervised deep-learning methods (PNN [99], 3D-CNN [100], and CMHF-net [75]). In Table 3.3, it can be observed that the proposed method significantly outperforms the SoTA methods, belonging to the same unsupervised paradigm with the learned image priors (DHIP [59] and uSDN [77]). For the results obtained using the uSDN [77] method, the released code (`https://github.com/mbaddeley/usdn`, accessed on October 18, 2020),

TABLE 3.3: Comparison results between unsupervised methods with learned image priors (DHIP [59] and uSDN [77]), and fully supervised deep-learning methods (PNN [99], 3D-CNN [100], and CMHF-net [75]) and ISGM (noise) method in the CAVE and Harvard Datasets with the Expanding Factor: 32.

Spatial Expanding Factor = 32											
Dataset		CAVE					Harvard				
Method		RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↑	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↑
Supervised	PNN	6.10	32.42	0.962	14.73	1.35	-	-	-	-	-
	3D-CNN	4.63	34.82	0.937	8.96	1.09	-	-	-	-	-
	CMHF-net	3.51	37.23	0.962	7.30	0.82	-	-	-	-	-
Unsupervised	uSDN	4.22	35.79	0.922	15.75	0.52	3.33	37.75	0.973	6.09	0.52
	DHIP	16.01	24.73	0.745	13.08	2.15	13.25	26.23	0.719	5.68	1.41
	ISGM (noise)	3.47	38.17	0.951	8.31	0.46	2.82	40.12	0.957	3.96	0.43
	ISGM+ (noise)	3.34	38.47	0.955	8.12	0.44	2.62	40.75	0.963	3.91	0.42

was re-run with default hyper-parameters, and only 8 samples out of 32 images provided correct super-resolved results. Thus, the average performance using only 8 correct samples is presented in Table 3.3. Moreover, the proposed method exhibits better performance than the fully supervised deep-learning methods and most traditional methods. One possible reason for this performance is the small number of samples used for training the HSI SR model, which is a challenge issue in the HSI SR scenario. A comparison between the results obtained using the unsupervised SoTA methods in the CAVE dataset and Harvard dataset is prensented (with up-scale factors of 8 and 16) in Tables 3.4 respectively. In these tables, it can be observed that the proposed method achieves better or comparable results with those obtained using SoTA methods. Furthermore, it must be clarified that the uSDN method cannot produce accurate results for the same samples, especially for large up-scale factors, and the average evaluations computed only with the correct outputs are presented in Tables 3.3 and 3.4.

Recently, deep learning-based methods have been extensively investigated for HS image fusion, most of which are implemented in a completely supervised and non-blind way. A few works attempted to employ an unsupervised non-blind strategy for the HS image fusion scenario, such as the unsupervised sparse Dirichlet-net (uSDN) [77], the deep hyperspectral image prior (DHP) [101], and the GDD method [81]. Our method falls in the unsupervised direction for HS image fusion. In this part, we provide the comparison with both supervised and un-supervised deep learning-based methods including SSF-Net [56], ResNet [102], DHSIS [57], uSDN [77], and DHP [101]. Since the supervised deep learning-based methods need training samples for model learning, we follow the experimental setting as in [56] and only give the compared results on 12 test images of the CAVE dataset and 10 test images of the Harvard dataset. The compared results on the CAVE and Harvard datasets are shown in Table 3.4 for both spatial expanding factor: 8 and 16, respectively. From Table 3.4, we can see that our proposed method can greatly outperform the deep unsupervised learning-based methods as well as manifests better performance than the supervised methods. In detail, our proposed method improves the PSNR values with 4.58dB/6.03dB and 3.94dB/6.80dB to the best unsupervised deep learning method for the upscale factor 8/16 on both CAVE and Harvard datasets.

TABLE 3.4: Comparison with the deep non-blind learning-based methods including the supervised approaches: SSFNet [56], ResNet [102], DHSIS [57] and the un-supervised approaches: uSDN [77], DHP [101], GDD [81] on the CAVE and Harvard datasets for both spatial expanding factors: 8 and 16.

Spatial Expanding Factor = 8											
Dataset		CAVE					Harvard				
Method		RMSE↓	PSNR↑	SSIM ↑	SAM↓	ERGAS↓	RMSE↓	PSNR↑	SSIM ↑	SAM↓	ERGAS↓
Supervised	SSFNet	1.89	44.41	0.991	3.31	0.89	2.18	41.93	0.991	4.38	0.98
	ResNet	1.47	45.90	0.993	2.82	0.79	1.65	44.71	0.984	2.21	1.09
	DHSIS	1.46	45.59	0.990	3.91	0.73	1.37	46.02	0.981	3.54	1.17
Unsupervised	uSDN	4.37	35.99	0.914	5.39	0.66	2.42	42.11	0.987	3.88	1.08
	DHP	7.60	31.40	0.871	8.25	4.20	7.94	30.86	0.803	3.53	3.15
	GDD	1.68	44.22	0.987	3.81	0.96	1.30	47.02	0.990	1.94	0.90
	ISGM (noise)	2.08	42.50	0.975	5.36	1.16	2.38	42.16	0.965	2.35	1.09
	ISGM+ (noise)	1.96	42.98	0.977	5.22	1.10	2.12	43.23	0.971	2.30	1.01
	ISGM (fusion)	1.44	45.61	0.992	3.27	0.79	1.17	48.27	0.993	1.75	0.77
	ISGM (RGB)	1.35	46.20	0.992	3.05	0.77	1.07	49.17	0.994	1.59	0.72
Spatial Expanding Factor = 16											
Supervised	SSFNet	2.18	41.93	0.991	4.38	0.98	1.94	43.56	0.980	3.14	0.98
	ResNet	1.93	43.57	0.991	3.58	0.51	1.83	44.05	0.984	2.37	0.59
	DHSIS	2.36	41.63	0.987	4.30	0.49	1.87	43.49	0.983	2.88	0.54
Unsupervised	uSDN	3.60	37.08	0.969	6.19	0.41	9.31	39.39	0.931	4.65	1.72
	DHP	11.31	27.76	0.805	10.66	3.09	10.38	38.44	0.754	4.57	2.08
	GDD	2.12	42.24	0.983	4.41	0.61	1.66	44.64	0.986	2.50	0.64
	ISGM (noise)	2.61	40.71	0.967	6.62	0.70	2.81	40.77	0.953	3.01	0.75
	ISGM+ (noise)	2.50	41.03	0.969	6.43	0.67	2.56	41.66	0.959	2.95	0.71
	ISGM (fusion)	1.76	43.84	0.999	3.76	0.49	1.32	47.16	0.992	1.99	0.47
	ISGM (RGB)	1.71	44.15	0.990	3.63	0.48	1.28	47.37	0.992	1.92	0.49

Moreover, compared with the best supervised methods, the improvements of the PSNR values are 0.36dB/2.12dB and 0.02dB/3.20dB, respectively, which demonstrates a significant margin profit of our method over the deep learning SoTA approaches for the large upscale factor. As we know that there usually have no sufficient training samples in supervised deep learning-based HS image reconstruction tasks compared with the gray/RGB image super-resolution problem. For example, there are only 32 HS images in the CAVE dataset and 50 images in the Harvard dataset, which are far less than the training sample numbers such as thousands of training samples in the RGB image super-resolution task. Although the network optimization is separately carried out for each dataset, the learned CNN model using the training sample still faces difficulty for well being adapted to the test sample. While our method can not only make full use of the powerful modeling capability of the deep network but also effectively learn the specific prior for an under-studying scene. Thus, the great performance gain with our proposed method has been obtained even compared with the supervised deep learning-based approaches.

TABLE 3.5: Ablation study of the evaluation results using our proposed method: ISGM+ (noise) with different weights α values of 0.3, 0.5 and 0.7 in the CAVE and Harvard datasets for both spatial expanding factors: 8, 16 and 32.

Dataset		CAVE			Harvard		
Spatial Expanding Factor	α	PSNR↑	SAM↓	ERGAS↓	PSNR↑	SAM↓	ERGAS↓
8	0.3	42.19	5.09	0.95	43.07	2.16	0.93
	0.5	42.91	4.40	0.86	41.68	2.19	1.06
	0.7	42.16	4.75	0.92	41.85	2.18	1.09
16	0.3	40.75	5.71	0.54	40.95	2.90	0.66
	0.5	40.75	5.87	0.55	40.79	2.70	0.62
	0.7	40.42	5.64	0.58	41.90	2.48	0.52
32	0.3	38.87	7.07	0.33	39.46	3.85	0.44
	0.5	38.03	7.26	0.37	40.02	3.51	0.39
	0.7	39.11	6.62	0.33	39.07	3.38	0.39

Ablation study

In order to evaluate the effect of different data terms on the loss function in ISGM (noise) model, we set the hyper-parameter α value as 0.3, 0.5 and 0.7, respectively, and provide the compared result in Table 3.5. Table 3.5 manifests the quantitative metrics of PSNR, SAM and ERGAS with our ISGM+ (noise) method, which illustrates that there are no large impact on the super-resolution performance via finetuning the hyper-parameter α. We are going to investigate in detail this phenomenon in the future.

In our proposed ISGM (fusion) method, we adjust the hyper-parameter α from 0 to 1.0 to verify the effectiveness of the loss terms in Eq. 3.6. The compared results on the CAVE dataset for the upscale factor 8 are shown in Table 3.6, which manifests the best performance can be achieved with $\alpha = 0.4$ and the second best one is obtained with $\alpha = 0.5$. When we set α as 0 or 1.0, which means only one loss term has been used while completely ignoring the other one, the reconstruction performance is significantly degraded especially with $\alpha = 0$. However, with α changing from 0.2 to 0.8, the reconstruction performance remains very stable, which means the plausible and robust HR-HS image can be achieved as long as the incorporated loss is adopted without large effect by the weight value.

We validate the performance effect of ISGM (RGB) by varying the block (scale) numbers in the generative network, the used loss terms and the employed network inputs. As we mentioned above, we employed an encoder-decoder architecture to serve as our specific CNN model, where both encoder and decoder paths contain multiple blocks for extracting multi-scale contexts in different receptive fields. To verify the effect of the used multi-scale, we vary the block numbers from 2 to 5, and we carry out the HR-HS image learning experiments. The compared results are shown in Table 3.7, which manifests the larger block number enables the performance improvement while the generative network even with two blocks only can also achieve very promising results. Moreover, as described in Eq. 3.6, we adopted

TABLE 3.6: Ablation study of our proposed ISGM (fusion) method with different α in the loss function (Eq. 3.6) in the CAVE dataset for upscale factor: 8.

Spatial Expanding Factor = 8					
Dataset	CAVE				
alpha	RMSE↓	PSNR↑	SSIM ↑	SAM↓	ERGAS↓
0.0	25.98	19.97	0.631	40.02	12.50
0.2	1.52	44.99	0.990	3.24	0.67
0.4	1.45	45.45	0.991	3.16	0.63
0.5	1.46	45.35	0.991	3.13	0.64
0.6	1.49	42.26	0.991	3.15	0.66
0.8	1.47	45.20	0.991	3.13	0.66
1.0	3.33	38.36	0.961	4.73	1.51

TABLE 3.7: Ablation studies of different block numbers of the generative network and loss terms on CAVE dataset with the up-scale factor 8.

Block number	Loss	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
2	Both	1.45	45.49	0.992	3.47	0.81
3		1.42	45.69	0.992	3.28	0.81
4		1.38	46.05	0.993	3.13	0.77
5	Loss1	26.27	19.85	0.601	43.53	16.19
	Loss2	3.30	38.57	0.972	3.68	1.88
	Both	1.35	46.20	0.992	3.05	0.77
	Skip w/o FT Conv	1.54	44.97	0.992	3.29	0.91

the reconstruction errors of both observed HR-RGB and LR-HS images as the loss function (denoted 'both' loss), and it is also possible to employ one loss term for our network training, denoted as loss1 and loss2, respectively. The compared results with different loss terms are also given in Table 3.7, which demonstrates two terms of loss achieve much better performance. At the same time, we also verify the effectiveness of the FT Conv module, and provide the compared results with the FT Conv module or the simple skip connection in Table 3.7, which manifests that the FT Conv module can improve the PSNR 1.23dB.

Finally, we verify the effectiveness of our proposed ISGM (RGB) method with the RGB-guided input. As mentioned above, the potential inputs to our generative network have several choices such as the noise, the combined data of two observations (denoted as 'Fused'), the HR-RGB observation with HR spatial structure and the possible condition (HR-RGB or LR-HS) with small perturbation including noise and dropout. Our method takes the HR-RGB image as the conditional guidance to the generative network since the HR spatial structure would benefit the plausible estimation of the absent spatial information in the LR-HS observation. The comprehensive comparisons with different network inputs are manifested in Table 3.8, and the conditional guidance with the HR-RGB image gives the best super-resolving performance.

TABLE 3.8: Ablation studies of different network inputs on CAVE dataset with the up-scale factor 8.

Block number:5 Loss: both					
Input	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
Noise	2.10	42.53	0.978	5.30	1.12
Fused	1.46	45.47	0.992	3.27	0.81
Fused+ Perturbation	1.44	45.61	0.992	3.72	0.80
RGB+Perturbation	1.35	46.20	0.992	3.05	0.77
RGB+dropout	1.36	46.13	0.993	3.06	0.76
RGB	1.51	45.11	0.989	3.68	0.86

Perceptual Quality

To visualize the reconstruction results of the ISGM (noise) model, two representative images from the CAVE and Harvard datasets, respectively, with different upscale factors (8, 16, and 32) are shown in Figures 3.4–3.9 using different deep unsupervised methods (DHIP [59], uSDN [77], the proposed (ISGM (noise) method, and the ISGM+ (noise) method with a perturbation term). In all figures, the first row shows the original HR image and the super-resolved results with spatial and spectral fidelity indexes (PSNR Sam for the recovered images using different methods), whereas the second row shows the difference images between the recovered HR-HS image and the ground-truth HR-HS image. In these figures, it can be observed that the proposed ISGM (noise) method is capable of recovering the HR-HS image with a smaller difference to the ground-truth HR-HS image and more reliable spatial/spectral indexes for most cases, excluding the results of uSDN in Figure 3.9. As mentioned in section above, although the uSDN method is capable of achieving impressive performance for some specific images using the default hyper-parameters, it cannot provide accurate results for most of the images, especially for those with large upscale factors. Despite the better results obtained for some images using the uSDN method, accurate recovered results cannot be achieved for 20 images out of the total 50 images. Moreover, the proposed ISGM+ (noise) method is capable of further improving the performance of the ISGM-version method for all images with different upscale factors.

Finally we provide the visual comparison with the traditional optimization-based method: CSU [93] and NSSR [46], the supervised deep learning-based methods: DHSIS [57], and the un-supervised deep learning-based methods: uSDN [77], DHP [101]. Figure 3.10 and 3.11 show the compared results (spectral band 16: 550nm and band 31: 700 nm) of one representative image from the CAVE dataset and the Harvard dataset, respectively. From Fig. 3.10 and 3.11, we can see that our proposed method manifests a more plausible visualization, especially in the difference between the recovered HR-HS image and the ground-truth image, which verifies the higher spatial accuracy of our proposed method in a specific band. In addition, we further attempt to provide a global evaluation of all spectral bands in the recovered HS image instead of a specific band. We calculate SAM values of all pixels to measure the spectral distortion and express them as the angle degree with value range [0; 180]. The obtained SAM values of all pixels can be considered as the intensities of a SAM image and visualized in Fig. 3.12. Small magnitudes in the visualized SAM

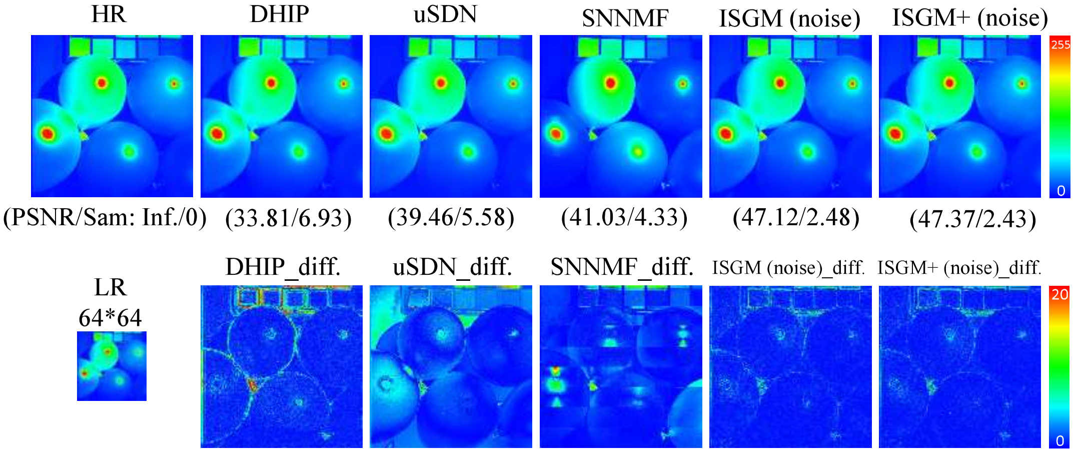


FIGURE 3.4: Recovered HR-HS image of the 'Balloon' sample in the CAVE dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 8.

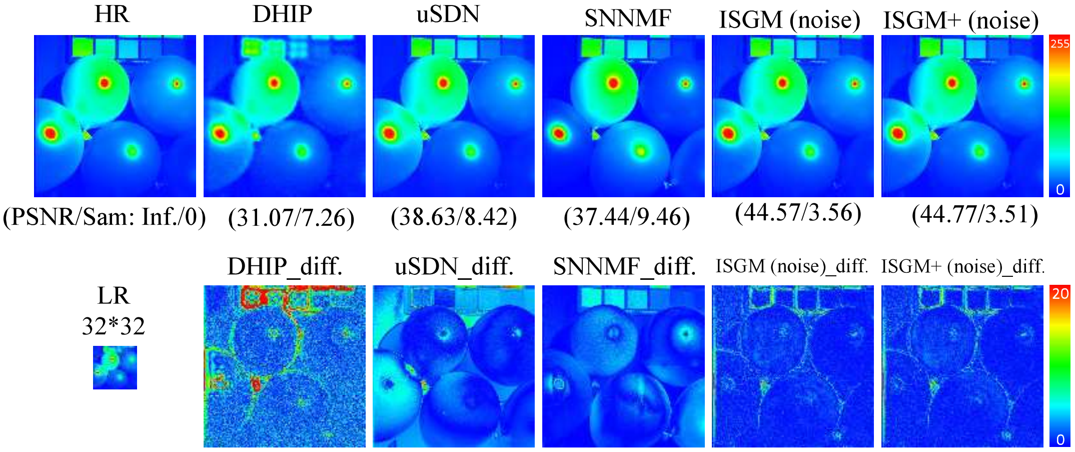


FIGURE 3.5: Recovered HR-HS image of the 'Balloon' sample in the CAVE dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 16.

images mean the small angle degrees and small spectral distortion. From Fig. 3.12, it can be observed that the SAM images of our proposed method show much smaller values for most pixels than the state-of-the-art methods.

3.7 Conclusion

In order to address the super-resolution issue for hyperspectral images, we provide a unsupervised deep hyperspectral image super-resolution framework with two different input modalities. A deep convolutional neural network is used to automatically learn the spatial and spectral features of latent HR-HS images from perturbed noisy input data and the fusion context that naturally collects a significant quantity of low-level image statistics. A special depth-wise convolution layer is designed to achieve degenerate transformations between observations and desired targets,

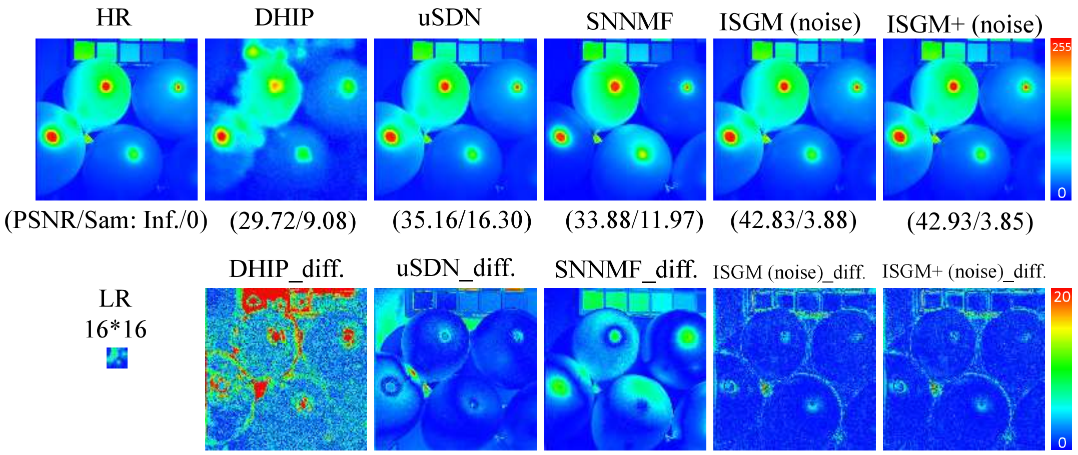


FIGURE 3.6: Recovered HR-HS image of the ′Balloon′ sample in the CAVE dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 32.

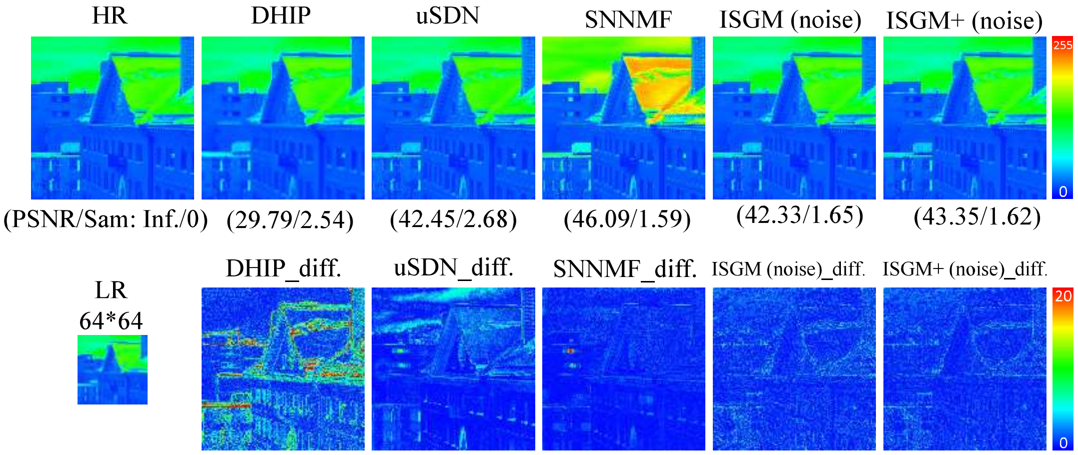


FIGURE 3.7: Recovered HR-HS image of the ′img1′ sample in the Harvard dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 8.

and this generates a universally learnable module that only uses low-quality observations. Without requiring training samples, the proposed an unsupervised deep learning framework can efficiently take advantage of the HR spatial structure of HR-RGB images and the detailed spectral characteristics of LR-HS images to deliver more accurate HS image reconstruction. We simply train the network parameters using the observed LR-HS and HR-RGB images and a generative network structure to reconstruct the underlying HR-HS images. Extensive research using the CAVE and Harvard datasets demonstrate the promising results in quantitative evaluation.

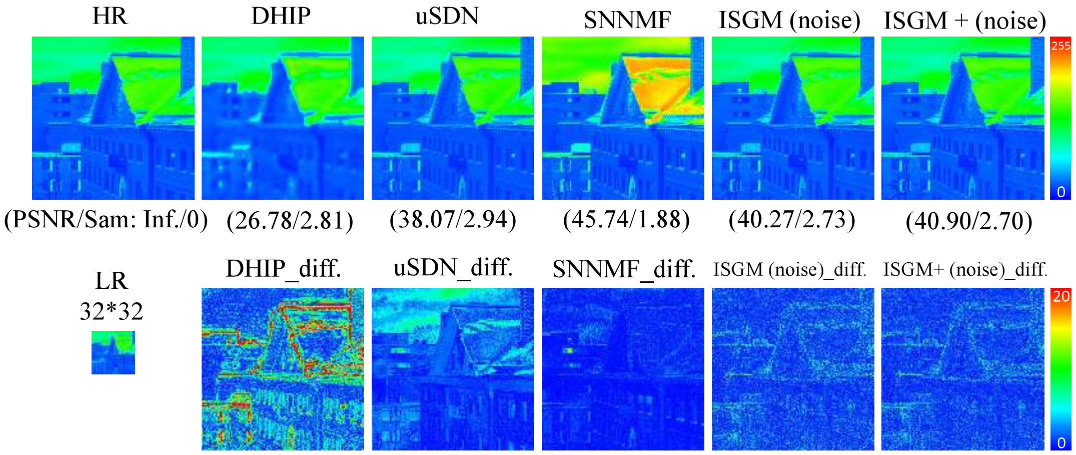


FIGURE 3.8: Recovered HR-HS image of the ′img1′ sample in the Harvard dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 16.

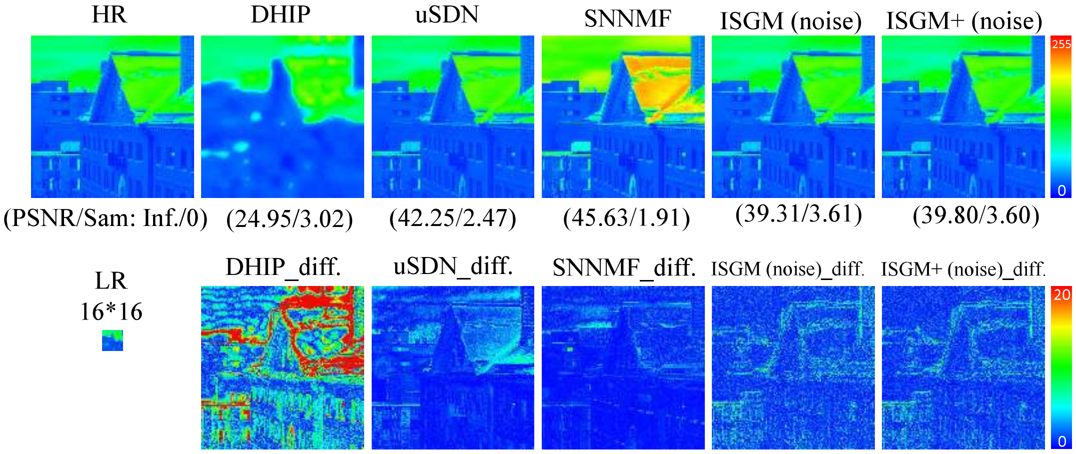
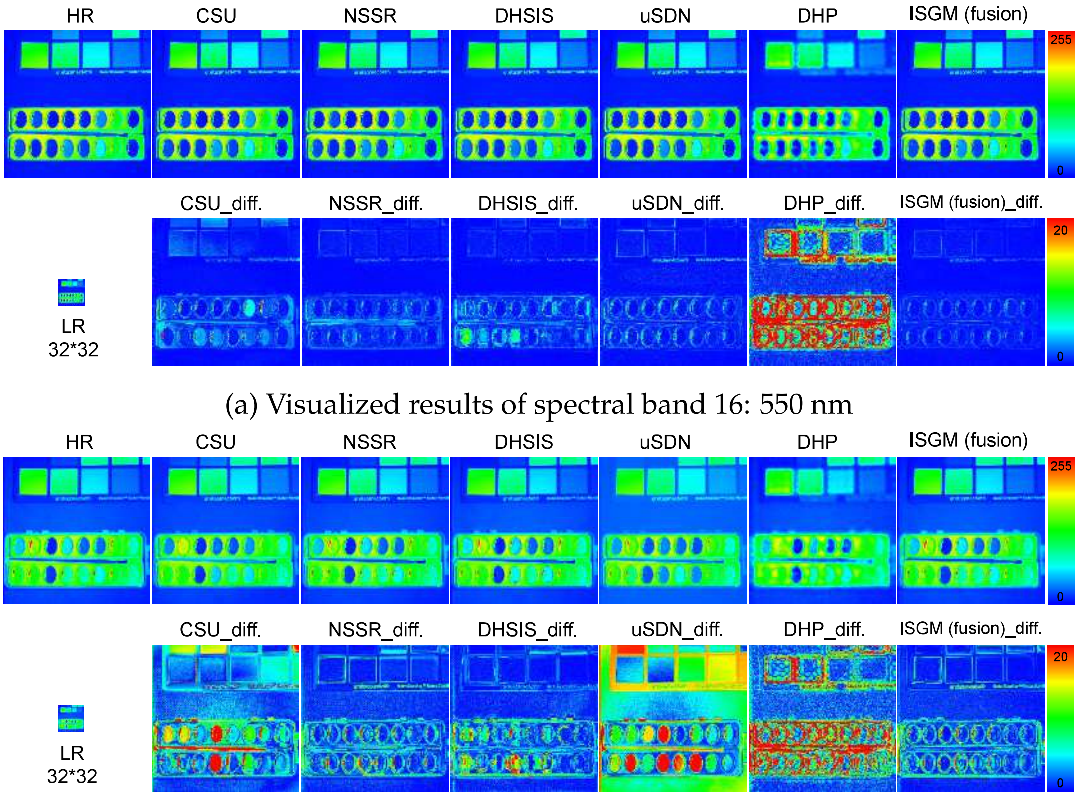


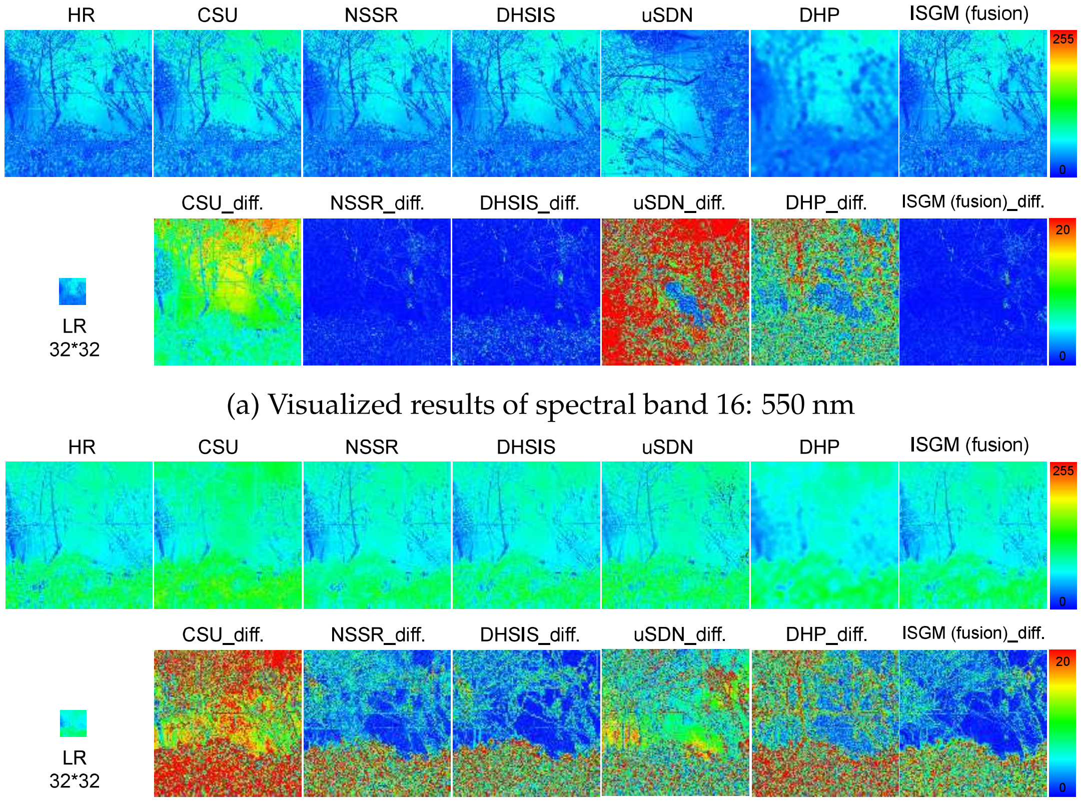
FIGURE 3.9: Recovered HR-HS image of the ′img1′ sample in the Harvard dataset using the DHIP [59], uSDN [77], SNNMF [66], and the proposed methods and corresponding difference images between the ground-truth and recovered images with an upscale factor of 32.



(a) Visualized results of spectral band 16: 550 nm

(b) Visualized results of spectral band 31: 700 nm

FIGURE 3.10: Visual compared results of one representative image: paints from the CAVE dataset with the traditional optimization-based method: CSU [93] and NSSR [46], the supervised deep learning-based methods: DHSIS [57], and the un-supervised deep learning-based methods: uSDN [77], DHP [101] for spatial expanding factor: 16. (a) Visualized results of spectral band 16: 550 nm. (b) Visualized results of spectral band 31: 700 nm.



(a) Visualized results of spectral band 16: 550 nm

(b) Visualized results of spectral band 31: 700 nm

FIGURE 3.11: Visual compared results of one representative image: imgb4 from the Harvard dataset with the traditional optimization-based method:CSU [93] and NSSR [46], the supervised deep learning-based methods: DHSIS [57], and the un-supervised deep learning-based methods: uSDN [77], DHP [101] for spatial expanding factor: 16. (a) Visualized results of spectral band 16: 550 nm. (b) Visualized results of spectral band 31: 700 nm.

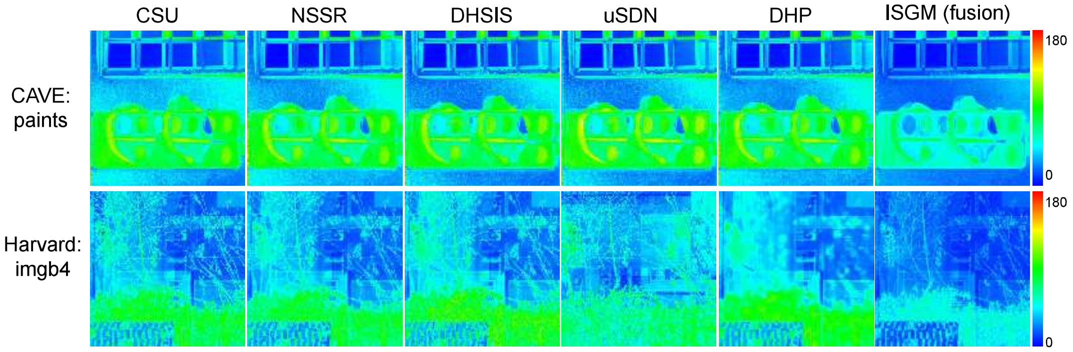


FIGURE 3.12: Sam results (spectral band 16: 550nm) of one representative image: paints from the CAVE dataset and imgb4 from the Harvard dataset with the traditional optimization-based method: CSU [93] and NSSR [46], the supervised deep learning-based methods: DHSIS [57], and the un-supervised deep learning-based methods: uSDN [77], DHP [101] for spatial expanding factor: 16.

Chapter 4

Unsupervised Blind Learning of Hyperspectral Image Super-Resolution

4.1 Traditional Non-Blind Spatial/Spectral Degradation

From the predicted HR-HS images of the generative neural network, it is possible to use mathematical degradation operations to approximate the LR-HS and HR-RGB images and then formulate quantitative criteria (loss function) for network learning in Eq. 3.6. However, mathematical operations outside the generative neural network may result in difficulties in the training procedure. In this work, without any loss of generalization, two parallel special convolutional blocks are leveraged to implement the traditional non-Blind spatial/spectral degradation model following the generative neural network and construct an end-to-end learnable framework. Specifically, the vanilla convolution layer is modified to be adapted for approximating the blurring/down-sampling operations and spectral transformation. Since, in a real scenario, each spectral band undergoes the same blurring/down-sampling transformations, the same kernel is defined for different spectral bands (channels) in a depth-wise convolutional layer with the stride parameter as the spatial expanding factor, and the bias term is set to 'False'. The formula for the blurring/dawn-sampling transformation can be expressed as follows:

$$\hat{\mathbf{X}} = f_{\mathbf{DB}}(G_\theta(z_{in}^0)), \tag{4.1}$$

where $f_{\mathbf{DB}}()$ represents the transformation of an especially designed depth-wise convolutional layer.

For implementing the spectral transformation from the generated HR-HS image $\hat{\mathbf{Z}}$ to the approximation of $\mathbf{X}$, a point-wise convolutional layer with 3 output channels is applied, and the bias term as set to 'False'. The formula for this spectral transformation can be expressed as follows:

$$\hat{\mathbf{Y}} = f_{\mathbf{C}}(G_\theta(z_{in}^0)), \tag{4.2}$$

where $f_{\mathbf{C}}()$ represents the spectral transformation with the point-wise convolutional layer. Using these two parallel blocks, an end-to-end learnable framework can be realized. For the known spatial blurring/down-sampling degradation, the kernel weight of the especially designed depth-wise convolutional layer with the known ones is initialized, and the trainable parameter is set to 'False'. Similarly, kernel weights are set for the point-wise convolutional layer as the known CSF (spectral transformation matrix) of the RGB camera. Therefore, the proposed image-specific generative model (ISGM) framework is considerably flexible and can be easily adapted to various degradation models. Moreover, it also has the prospect of automatically learning the transformation parameters in the embedded convolutional blocks for unknown degradations, which is left for future research. By replacing the spatial and spectral degradation operations with the designed convolutional blocks, the loss function for training the proposed deep unsupervised fusion learning network can be rewritten as follows:

$$\theta^* = \arg\min_{\theta} \alpha\beta_1 ||\mathbf{X} - f_{\mathbf{DB}}(G_\theta(\mathbf{z_{in}}))||_F^2 + (1-\alpha)\beta_2 ||\mathbf{Y} - f_{\mathbf{C}}(G_\theta(\mathbf{z_{in}}))||_F^2, \tag{4.3}$$

The optimization of Eq. 4.3 for obtaining the optimal parameter set of the generative neural network can be considered to be a kind of "zero shot" learning [103]. During the training procedure, only the low-quality image pairs (that is, the observed LR-HS and HR-RGB images) are used without the corresponding label (the HR-HS images) and thus the proposed method is completely unsupervised of being generalized for any real observations. The detail implementation for the proposed image-specific generative model (ISGM) is summarized in Algorithm 1.

Algorithm 1 Algorithm of the proposed deep unsupervised fusion learning method.

```
Require: The observed LR-HS image $\mathbf{X}$ and HR-RGB image $\mathbf{Y}$
Ensure: Latent HR-HS image $\mathbf{Z}$
Sample $\mathbf{z}_{in}^0$ from uniform distribution with seed 0
for $i = 0$ to max. iter. $(I)$ do
    Sample $\mathbf{n}_{(0,1)}^i$ from uniform distribution
    Perturb $\mathbf{z}_{in}^0$ with $\mathbf{n}_{(0,1)}^i$: $\mathbf{z}_{in}^i = \mathbf{z}_{in}^0 + \beta\mathbf{n}_{(0,1)}^i$
    $\hat{\mathbf{Z}} = G_\theta(\mathbf{z}_{in}^i, \theta^{i-1})$
    $\hat{\mathbf{X}} = f_{DB}(\hat{\mathbf{Z}})$
    $\hat{Y} = f_C(\hat{\mathbf{Z}})$
    Loss function: $\alpha\beta_1||\mathbf{X} - \hat{\mathbf{X}}||_F^2 + (1-\alpha)\beta_2||\mathbf{Y} - \hat{\mathbf{Y}}||_F^2$
    Compute the gradients regarding $G_\theta$
    Update $\theta$ using the ADAM algorithm [94] as $\theta^i$
end for
$\mathbf{Z} = G_\theta(\mathbf{z}_{in}^0)$
```

4.2 Learnable Blind Method

The degradation operations (spatial blurring/down-sampling and spectral transformation) should be known in all the methods. But is is unrealistic in a real scenario. How to implement the degradation operations (Blurring, down-sampling, and spectral transformation) following the generative network for developing an end-to-end learning framework is another problem.

Our proposed ISGM (blind) method attempts to learn the specific priors of the latent HR-HS image automatically to provide reconstruction results with unknown degradation models. In the following, we introduce details of our proposed method to the above problems. We leverage the observed LR-HS and HR-RGB images instead of the randomly generated noise as the network input. Simultaneously, we employ two specific convolutional layers to approximate the degradation operations, which can be implemented as both learn-able or fixed degradation models for different real problem settings. Next, we will substantiate the adopted input data to our self-supervised network and the implementation of the learn-able degradation module.

4.2.1 Non-blind

Non-blind degradation refers to a scenario where the degradation model and its parameters are known. Hyperspectral Image Super-Resolution (HSI-SR) refers to the task of increasing the spatial resolution of a hyperspectral image. In the case of non-blind degradation in HSI-SR, we have prior knowledge of the degradation model and its parameters, which can help us design effective algorithms for image super-resolution. The degradation model typically includes the blurring kernel and the downsampling operator used to generate the low-resolution image from the high-resolution image. One popular approach for solving HSI-SR problems is through the use of deep learning-based methods, where convolutional neural networks (CNNs) are trained on pairs of low-resolution and high-resolution hyperspectral images. These networks learn to map the low-resolution image to the high-resolution image, effectively removing the effects of the degradation model. To train these networks, it is important to have a large dataset of paired low-resolution and high-resolution hyperspectral images that are degraded using the same degradation model. Once trained, these networks can be used to perform super-resolution on new low-resolution hyperspectral images with the same degradation model.

In non-blind degradation, the degradation model and its parameters are known in advance, which means that the process of restoring the original image is simpler than in the case of blind degradation, where the degradation model and its parameters are unknown. Knowing the degradation model and its parameters can be very helpful when it comes to developing image restoration algorithms because it allows us to design methods that can effectively compensate for the specific types of degradation that have occurred. This knowledge can also be used to develop efficient algorithms for super-resolution, denoising, deblurring, and other image restoration tasks. In contrast, in the case of blind degradation, the restoration process is more challenging since the degradation model is unknown, and the task becomes one of estimating the model parameters as well as the original image. This makes the problem much more difficult and requires more advanced algorithms to solve.

4.2.2 (Semi-)Blind

From the predicted HR-HS image of the generative network, we can employ the degradation operations to obtain the approximated LR-HS and HR-RGB images for constructing the evaluation criterion of network training. However, simply using the mathematical operations to approximate the degradation model would lead to this part outside the network and cannot integrate into an end-to-end learning framework. In this study, we leverage two parallel blocks (seen as Fig. 4.1) following the generative backbone to approximate the degradation models as a whole learnable framework. Specifically, we modify a conventional depth-wise convolution layer to adapt to the blurring and down-sampling transformation. Since the same blurring and down-sampling transformation is conducted in each spectral band in a real scenario, we impose the same kernel on different spectral bands in the depth-wise convolution layer and set stride as a spatial expanding factor and the

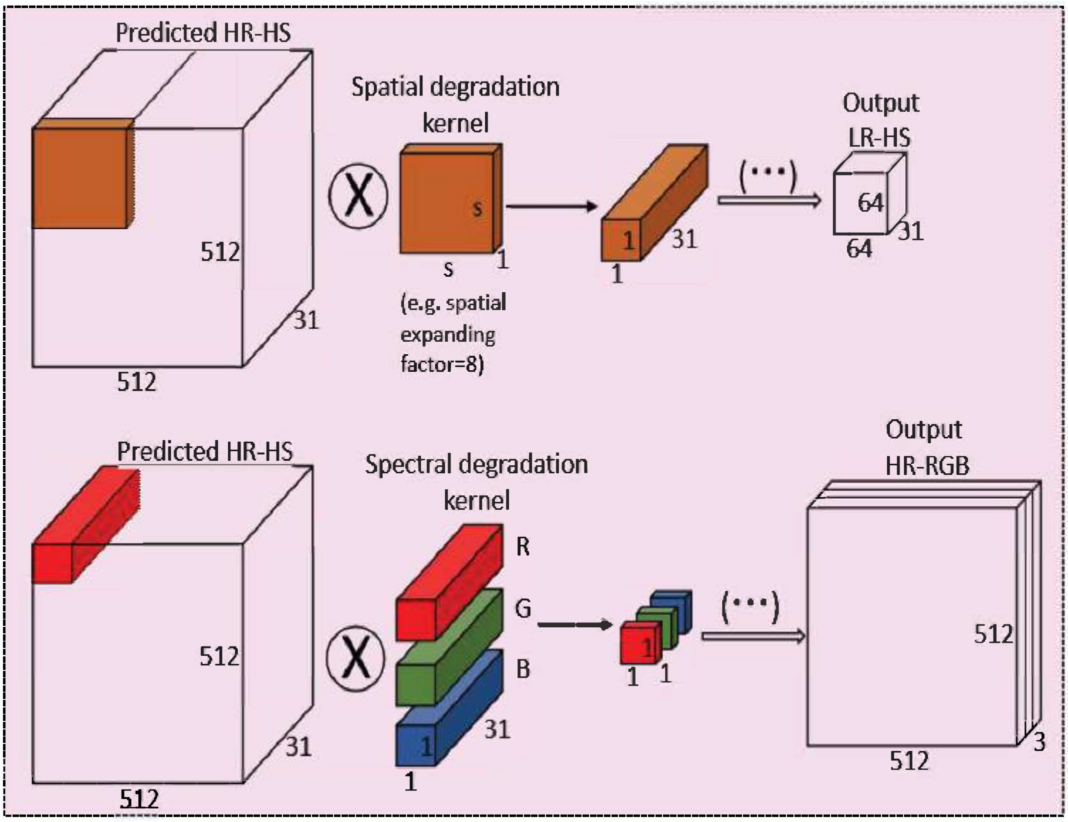


FIGURE 4.1: Concept of two parallel blocks for the degradation

bias term as False. The formulation for the blurring and down-sampling transformation is expressed as

$$\bar{\mathbf{X}} = f_{SDW}(G_\theta(\mathbf{Z}_{\{\mathbf{XY}\}})), \tag{4.4}$$

where $f_{SDW}(\cdot)$ denotes the operation in the specific depth-wise convolution layer. In detail, let denote the same kernel in the depth-wised convolution layer, which is used to separately convolute with all channels in the generated HR-HS image $G_\theta(\mathbf{Z}_{\{\mathbf{XY}\}})$, as $\mathbf{k}_{SDW} \in \mathbf{R}^{1\times 1\times s\times s}$, the transformation of $f_{SDW}(G_\theta(\mathbf{I}_{\{\mathbf{XY}\}}))$ with the spatial expanding factor as stride and the False bias would be boiled down to the conventional mathematical 2D convolution and nearest down-sampling operators via reducing the first- and second- modes of $\mathbf{k}_{SDW}$ to $\mathbf{k} \in \mathbf{R}^{s\times s}$:

$$\bar{\mathbf{X}} = \mathbf{k} \otimes G_\theta(\mathbf{Z}_{\{\mathbf{XY}\}})^{(\mathbf{Spa})} \downarrow . \tag{4.5}$$

where the weight parameters of $\mathbf{k}_{SDW}$ can be predefined by setting as the known blurring kernel of the real degradation procedure or automatically be learned when the blurring procedure is unknown. Thus, we can simply implement the f_{SDW} using the specific depth-wise convolution layer, and then easily get the approximated LR-HS image from the generated HR-HS image with G_θ.

Furthermore, we employ a conventional convolution layer with kernel size $1 * 1$ and the output channels 3 to implement the spectral transformation. Similarly, we set stride as 1 and the bias term as False, which is formulated as

$$\bar{\mathbf{Y}} = f_{SC}(G_\theta(\mathbf{Z}_{\{\mathbf{XY}\}})), \tag{4.6}$$

where $f_{SC}(\cdot)$ denotes the operation in the spectral convolution layer. In $f_{SC}(\cdot)$, the convolution kernel $\mathbf{k}_{SC} \in \mathbf{R}^{L\times3\times1\times1}$ is used to convert the detailed spectra of the generated HR-HS image $G_\theta(\mathbf{Z}_{\{\mathbf{XY}\}})$ to the degraded RGB image. Moreover, by reducing the third- and fourth- modes of $\mathbf{k}_{SC}$, the learnable kernel has the same dimensionality as the spectral sensitivity function $\mathbf{C}^{(Spec)}$ of an RGB sensor, and thus can be adopted to approximate the $\mathbf{C}^{(Spec)}$ in our overall network. With the above design, these two blocks can be parallelly implemented in our end-to-end learnable framework. If the blurring kernel of the spatial degradation is known, we simply initialize the weights of layer as the known kernel and set trainable as False in the learning procedure. Similarly, we also set the weights of the $1*1$ kernel in $f_{SC}(\cdot)$ as the known CSF of the RGB camera or automatically learn it in the network training procedure. Thus, the investigated learnable framework is very flexible and easily adapted to different real settings. Via substituting the degradation operation with our designed convolution blocks, the loss function for training our deep self-supervised network can be rewritten as

$$\begin{aligned}(\theta^*, \theta^*_{SDW}, \theta^*_{SC}) = \arg\min_{\theta} \alpha \left\|\mathbf{X} - f_{SDW}(G_\theta(\mathbf{Z}_{\{\mathbf{XY}\}}))\right\|^2 \\ + (1-\alpha)\left\|\mathbf{Y} - f_{SC}(G_\theta(\mathbf{Z}_{\{\mathbf{XY}\}}))\right\|^2 \\ s.t. 0 \le G_\theta(\mathbf{Z}_{\{\mathbf{XY}\}})_i \le 1 \forall i.\end{aligned} \tag{4.7}$$

In Eq. 4.7, it can be seen that instead of optimizing directly on the latent HR-HS image, we learn the parameters of the generative network for well reconstructing the target. The optimization process of our network can be explained as a kind of "zero-shot" self-supervised learning [104], where the generative network G_θ is trained using a test image pair (i.e., the observed LR-HS and HR-RGB images) only and no ground-truth HR-HS image is available. Thus, we dub our method as a self-supervised learning framework for HS image fusion.

4.3 Experiment Results

4.3.1 Comparison with (Semi-)Blind Methods

Our proposed ISGM(blind) method is exploited in a unified framework, which is capable of reconstructing the HR-HS image from the observations not only with the known spatial and spectral degradation operations but also with the unknown spatial or spectral degradation operations or both unknown. Thus our proposed method can be implemented in a complete blind setting (both unknown spatial down-sampling kernel for LR-HS image and the unknown CSF for HR-RGB image). The compared results using our proposed method with semi-blind and complete-blind settings, the state-of-the-art unsupervised semi-blind methods: UAL method [82] for spatial blind only, and the spatial blind implementation of NSSR [46] via setting the incorrect spatial kernel, have been given in Table 4.1. Since the UAL method [82] used the specified kernels for verifying the effectiveness to be adopted to various spatial kernels, we utilized Bicubic down-sampling without the lack of generalization, the average down-sampling [46], and the specified spatial kernels: K1 and K2 in [82], to create the LR-HS image, and then manifest the compared HS image reconstruction performance via automatically learning the down-sampling kernels in our proposed ISGM(blind) method for a fair comparison. From Table 4.1, we can see that our proposed method outperforms most SoTA methods in the same experimental setting, and has great potential to be adopted to any real scenario. With the same experimental setting, our proposed method improves the PSNR values from

TABLE 4.1: Comparison with the (semi-)blind methods including the unsupervised semi-blind approaches: UAL [82] and the semi-blind implementation of NSSR [46] on the CAVE and Harvard datasets for both spatial expanding factors: 8 and 32.

Method	Real Down-sampling Kernel	CAVE					Harvard				
		RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS ↓	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
Spatial Expanding Factor = 8											
NSSR (Bic) [46]	Bicubic	3.41	38.03	0.968	5.35	1.52	2.76	39.77	0.981	2.00	1.30
NSSR (Ave) [46]	Average	2.76	39.77	0.981	2.00	1.30	3.27	38.55	0.972	5.17	1.78
UAL [82]	K1	1.85	43.23	0.986	6.72	-	2.08	42.38	0.982	2.67	-
	K2	2.01	42.72	0.986	6.78	-	-	-	-	-	-
ISGM (Spatial blind)	K1	1.47	45.14	0.990	3.54	0.66	1.15	47.59	0.994	1.70	0.78
	K2	1.56	44.71	0.989	3.64	0.69	1.12	47.75	0.994	1.70	0.79
	Bicubic	1.70	44.05	0.988	3.70	0.75	1.33	46.28	0.992	1.95	0.93
ISGM (Spectral blind)	Bicubic	1.64	44.36	0.989	3.66	0.72	1.28	46.67	0.992	1.86	0.89
ISGM (Complete blind)	Bicubic	1.68	44.10	0.988	3.72	0.74	1.32	46.44	0.992	1.91	0.91
Spatial Expanding Factor = 32											
UAL [82]	K1	2.66	40.43	0.983	7.62	-	2.14	41.82	0.979	3.30	-
ISGM (Spatial blind)	K1	2.96	39.21	0.972	6.68	0.32	1.87	43.22	0.987	2.91	0.35
	Bicubic	2.84	39.42	0.973	6.56	0.32	1.96	42.71	0.986	2.92	0.36
ISGM (Spectral blind)	Bicubic	2.75	39.80	0.974	6.26	0.31	1.92	43.00	0.986	2.87	0.37
ISGM (Complete blind)	Bicubic	4.17	36.75	0.972	6.31	0.39	3.71	37.46	0.984	2.95	0.43

43.23dB/42.73dB to 45.14dB/44.71dB with the degradation kernels K1/K2 [82] for the upscale factor 8 on the CAVE dataset while from 42.38dB to 47.59dB with the kernel K1 on the Harvard dataset. Finally, we also added the compared results of the traditional optimization-based NSSR with incorrect spatial kernel in Table 4.1. In these additional experiments, we created the LR-HS image by down-sampling the original HR-HS image with the average kernel and Bicubic operation, respectively. Since the NSSR method cannot automatically learn the spatial degradation operation, we assumed the spatial down-sampling kernel as Gaussian kernel without the lack of generalization under the unknown spatial degradation assumption, and conducted the NSSR method to recover the HR-HS image, denoted as NSSR (Ave) and NSSR (Bic), respectively. From Table 4.1, it is obvious that our proposed method significantly improve the performance compared with NSSR [46] under the blind-experimental settings.

4.3.2 Comparison with SoTA Methods

For comparing our ISGM (blind) method, we first carried out our tests for the spatial scale factors 8 and 16, and we contrasted our approach with the most recent techniques, such as those based on mathematical optimization: GOMP [95], MF [50], SNNMF [66], CSU [45], NSSR [46], supervised deep learning-based methods: SCT-SDCNN (w/o the HR-RGB image as input) [105], SSFNet [48], DHSIS [57], ResNet

TABLE 4.2: Compared results with the SoTA methods including mathematical optimization-based and deep learning-based methods on both CAVE and Harvard datasets with the up-scale factor 16.

Dataset		CAVE					Harvard				
Method		RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
Mathematical optimization	GOMP [95]	6.08	32.96	-	12.60	1.43	3.83	38.56	-	4.16	0.77
	MF [50]	2.71	40.43	-	4.82	0.73	1.94	43.30	-	2.85	0.47
	SNNMF [66]	2.45	42.21	-	4.61	0.66	1.93	43.31	-	2.85	0.45
	CSU [45]	2.87	39.83	0.983	5.65	0.79	1.60	45.50	0.992	1.95	0.44
	NSSR [46]	1.78	44.01	0.990	3.59	0.49	1.65	44.51	0.993	2.48	0.41
Deep learning	SSFNet [48]	2.18	41.93	0.991	4.38	0.98	1.94	43.56	0.980	3.14	0.98
	DHSIS [57]	2.36	41.63	0.987	4.30	0.49	1.87	43.49	0.983	2.88	0.54
	MHF-net [75]	-	44.51	0.992	4.00	0.38	-	46.23	0.987	3.09	0.54
	LTTR [107]	-	42.48	0.987	4.25	0.47	-	45.82	0.986	3.11	0.65
	CNN-FUS [106]	-	40.37	0.979	5.85	0.59	-	43.47	0.966	5.41	0.92
	ResNet [102]	1.93	43.57	0.991	3.58	0.51	1.83	44.05	0.984	2.37	0.59
	MoG-DCN [108]	-	46.84	0.995	2.62	0.31	-	46.43	0.987	2.93	0.53
	uSDN [77]	3.60	37.08	0.969	6.19	1.35	9.31	39.39	0.931	4.65	1.72
	DUFL [109]	2.61	40.71	0.967	6.62	0.70	2.81	40.77	0.953	3.01	0.75
	ISGM (blind)	1.71	44.15	0.990	3.63	0.48	1.28	47.37	0.992	1.92	0.49

[102], CNN-FUS [106], LTTR [107], MHF-net [75], MoG-DCN [108], and unsupervised deep learning-based methods: uSDN [77], DUFL [109]. Table 4.2 compares the outcomes of the upscaling factor 16 for the Harvard and CAVE datasets. The supplemental material displays the comparison outcomes of the upscaling factor 8. According to Table 4.2, our method can perform much better than the majority of SoTA methods across all evaluation metrics. More particular, our approach outperforms both unsupervised optimization-based and deep learning-based approaches, and on the CAVE dataset, it exhibits better or equivalent outcomes with the supervised deep learning approach. The suggested approach greatly outperforms all SoTA approaches for the Harvard dataset. Next, we further conducted experiments for much larger spatial upscale factor: 32. The compared results on two images from both CAVE and Harvard datasets are given in Table 4.3, where all compared methods are implemented for the well-registered HR-RGB and LR-HS pairs. It should be noted that the compared u^2-MDN method [110] was proposed for HSI SR for the unregistered HR-RGB and LR-HS observation. In order to give a fair comparison, the values of the u^2-MDN method in Table 4.3 are the results for the well-registered input pair.

4.3.3 Ablation Study

Different Settings in the Proposed ISGM (blind) Method

In our proposed DSSH method, we employ the fused context as the input of the generative network instead of noise as in DHP [101], which is expected to provide insight about spectral correlation and the high-resolution spatial structures existed in the observed LR-HS and HR-RGB images. In addidition, DHP simply adopts a

TABLE 4.3: Compared results of two representative images from both CAVE and Harvard datasets with the SoTA methods including the u^2-MDN [110] method for the up-scale factor 32.

	CAVE						Harvard					
	ballon			cloth			img1			imgb5		
	PSNR	SAM	ERGAS	PSNR	SAM	ERGAS	PSNR	SAM	ERGAS	PSNR	SAM	ERGAS
CNMF [93]	39.27	9.71	0.26	30.52	6.55	0.54	37.25	2.86	0.15	39.06	2.14	0.17
CSU [45]	41.52	4.68	0.19	33.47	5.52	0.40	39.12	2.30	0.12	39.01	2.37	0.18
NSSR [46]	43.20	3.35	0.16	33.30	4.58	0.31	39.91	2.24	0.14	39.12	2.17	0.17
uSDN [77]	41.54	4.56	0.20	33.48	4.16	0.35	39.30	2.27	0.12	39.72	2.10	0.16
u^2-MDN [110]	43.59	1.93	0.16	34.85	4.31	0.30	40.97	2.06	0.11	39.76	2.08	0.15
ISGM (blind)	46.12	2.44	0.11	36.80	3.62	0.26	48.38	1.24	0.07	49.98	1.36	0.13

TABLE 4.4: Compared results with the SoTA methods on the NUS dataset with the up-scale factor 16.

Dataset		NUS			
Method		RMSE↓	PSNR↑	SSIM↑	SAM↓
Mathematical optimization	CSU [45]	1.65	44.80	0.976	3.23
	NSSR [46]	1.21	47.56	0.972	2.78
Deep learning	DHSIS [57]	1.47	45.48	0.981	3.15
	ResNet [102]	1.19	43.06	0.975	2.83
	uSDN [77]	2.21	41.77	0.970	5.15
	ISGM (blind)	1.18	47.59	0.986	2.57

randomly generated noise as the input, which cannot leverage the spectral correlation in the observed LR-HS image and the HR spatial structure in the observed LR-HS image for constraining the network learning, while our method leverages the combined observations with a small perturbed noise as the input of the network, and is expected to reconstruct a more robust and stable HR-HS image. It is possible to replace the noise input as the fused context: $Z_{\{XY\}}$ in DHP to validate the effectiveness of different components in our proposed method. We provided the compared results using the DHP network with the perturbed combination of the observed images in Table 4.5. It can be seen that our method can achieve much better reconstruction performance. Also, we implement the spatial degradation model as a learnable module inside the DSSH framework. With the known kernel of the spatial degradation model, we simply set the trainable parameter as False. In contrast, we can simultaneously learn the kernel and the generative network with the unknown spatial kernel. Table 4.5 gives the compared results using different settings: the input of the generative network (noise or the fused context: $Z_{\{XY\}}$) and the spatial degradation kernels including the known kernel with Lanczos approximation, and the unknown kernel: the widely used Gaussian kernels with different standard deviations and automatically learned kernel. Regarding to the Gaussian kernel, the hyper-parameter: standard deviation is needed to be defined previously. Firstly, we conducted a pilot experiment with different parameters: $1/2$, $1/\sqrt{2}$ and

TABLE 4.5: Compared results using different settings: the input of the generative network (noise or the fused context: $Z_{\{XY\}}$) and the spatial degradation kernels including the known kernel (Lanczos), and the unknown kernels: the widely used Gaussian kernels with different standard deviations and automatically learned kernel, in our proposed DSSH method for both CAVE and Harvard datasets with the spatial expanding factors: 8 and 16.

Factor	Kernel	CAVE					Harvard				
		RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
8	DHP-Lanczos+$Z_{\{XY\}}$	3.33	38.36	0.961	4.73	1.51	3.81	37.26	0.942	2.14	1.41
	Lanczos+Noise	2.10	42.53	0.978	5.30	1.12	2.15	42.63	0.975	2.32	1.01
	Lanczos+$Z_{\{XY\}}$	1.44	45.60	0.992	3.27	0.80	1.17	48.27	0.993	1.78	0.77
	Gauss1+$Z_{\{XY\}}$	5.97	33.21	0.936	5.55	2.55	4.47	35.48	0.960	2.48	1.74
	Gauss2+$Z_{\{XY\}}$	6.17	33.05	0.932	6.05	2.60	6.56	32.51	0.943	5.06	2.36
	Gauss3+$Z_{\{XY\}}$	6.96	31.86	0.926	6.03	2.95	6.82	30.85	0.929	4.91	2.29
	Learned+$Z_{\{XY\}}$	1.70	44.05	0.988	3.70	0.75	1.33	46.28	0.992	1.95	0.93
16	DHP-Lanczos+$Z_{\{XY\}}$	4.20	36.26	0.948	5.53	0.95	4.71	35.32	0.992	2.52	0.90
	Lanczos+Noise	2.60	40.75	0.970	6.42	0.70	9.46	38.14	0.876	8.52	7.71
	Lanczos+$Z_{\{XY\}}$	1.77	43.85	0.989	3.76	0.50	1.32	47.16	0.992	1.99	0.47
	Gaussian+$Z_{\{XY\}}$	5.64	33.64	0.943	6.68	1.21	3.84	36.67	0.973	2.71	1.11
	Learned+$Z_{\{XY\}}$	2.17	41.95	0.983	4.56	0.47	1.51	45.21	0.990	2.24	0.52

$2 \times factor/6$ on the CAVE dataset with the expanding factor 8, and then utilize the parameter achieving the best performance on the CAVE dataset for other experiments. From Table 4.5, it is obvious that with the known spatial kernel, the best performance can be obtained for both CAVE and Harvard datasets with different spatial expanding factors. However, the performances with a wrong selected spatial kernel are decreased greatly while automatically learning the spatial kernel leads to a little performance decreasing compared with the known kernel results.

Next, we adjust the hyper-parameter α from 0 to 1.0 to verify the effectiveness of the loss terms in Eq. 4.3. The compared results on the CAVE dataset for the upscale factor 8 are shown in Table 4.6, which manifests the best performance can be achieved with $\alpha = 0.4$ and the second best one is obtained with $\alpha = 0.5$. When we set α as 0 or 1.0, which means only one loss term has been used while completely ignoring the other one, the reconstruction performance is significantly degraded especially with $\alpha = 0$. However, with α changing from 0.2 to 0.8, the reconstruction performance remains very stable, which means the plausible and robust HR-HS image can be achieved as long as the incorporated loss is adopted without large effect by the weight value.

Finally, we verified the effectiveness of our method on the NUS dataset. The compared results with different paradigms for HSI SR task are given in Table 4.4, which also manifests great improvement margin over the SoTA methods.

4.3.4 Perceptual Quality

Moreover, we show the visualization results (one band) of a representative image with several unsupervised model-based and deep learning-based methods in the

TABLE 4.6: Ablation study with different α in the loss function (Eq. 4.7) on the CAVE dataset for upscale factor: 8.

Spatial Expanding Factor = 8					
Dataset	CAVE				
alpha	RMSE↓	PSNR↑	SSIM ↑	SAM↓	ERGAS↓
0.0	25.98	19.97	0.631	40.02	12.50
0.2	1.52	44.99	0.990	3.24	0.67
0.4	**1.45**	**45.45**	**0.991**	3.16	**0.63**
0.5	1.46	45.35	0.991	**3.13**	0.64
0.6	1.49	42.26	0.991	3.15	0.66
0.8	1.47	45.20	0.991	3.13	0.66
1.0	3.33	38.36	0.961	4.73	1.51

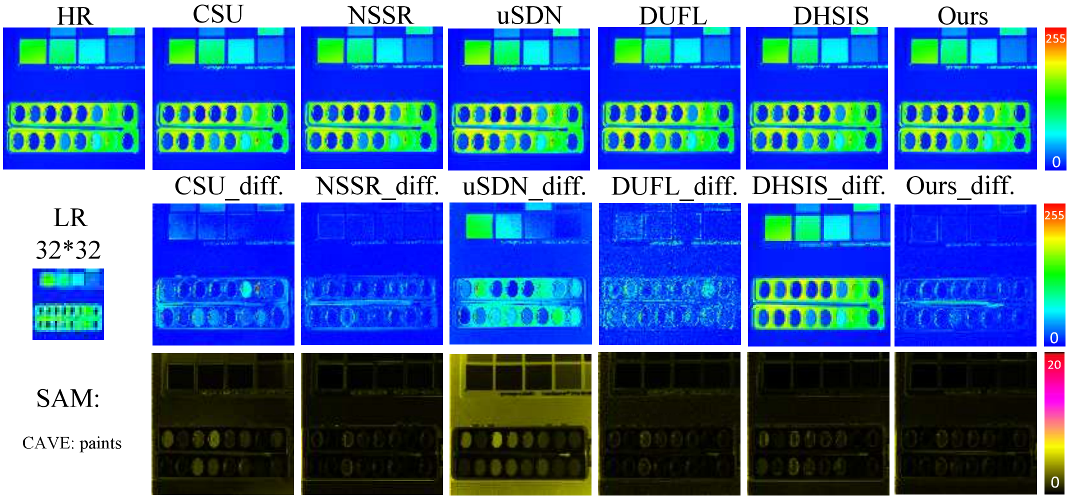


FIGURE 4.2: Visual difference results of the representative images: Paints in the CAVE dataset of mathematical optimization-based methods: CSU [45], NSSR [46] and deep learning-based methods: uSDN [77], DUFL [109], DHSIS [57] and our method on the CAVE dataset with the up-scale factor 16.

first two rows of Fig. 4.2 and Fig. 4.3 for both CAVE and Harvard dataset by the up-scale factor 16, which also demonstrate our proposed method achieves much smaller reconstruction errors on all spatial positions. To provide the compared spectral recovery fidelity of all wavelength bands instead of one band, we compute the SAM value (Angular degree: $0^o - 180^o$) of each pixel in the representative images, and visualize the SAM intensity as an image in the third row of Fig. 4.2 and Fig. 4.3 with the up-scale factor 16. Further, the spectral curves of one pixel in the representative images are also provided in Fig. 4.4 and Fig. 4.5, which manifests the high spectral fidelity of our proposed method.

4.4 Conclusion

This chapter proposed a novel deep self-supervised learning framework for hyperspectral image reconstruction. The proposed framework is completely non-dependent

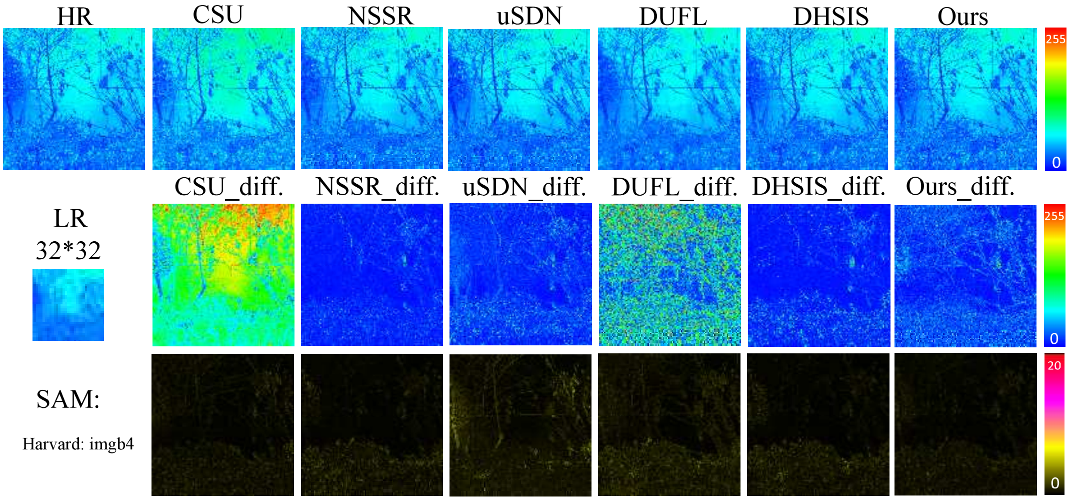


FIGURE 4.3: Visual difference results of the representative images: imgb4 in the Harvard dataset of mathematical optimization-based methods: CSU [45], NSSR [46] and deep learning-based methods: uSDN [77], DUFL [109], DHSIS [57] and our method on the CAVE dataset with the up-scale factor 16.

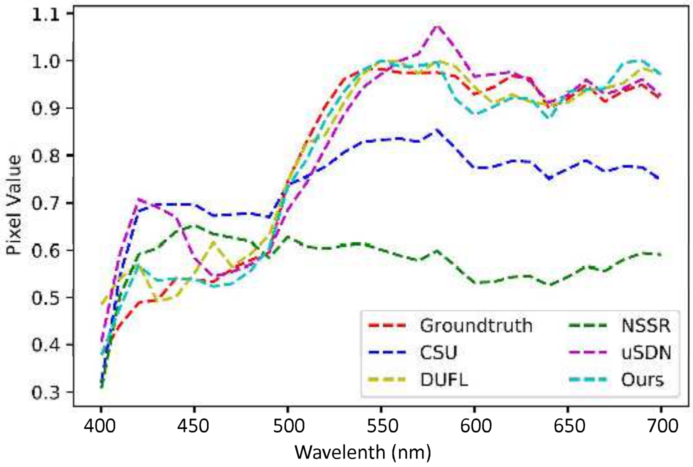


FIGURE 4.4: The recovered spectral curve of one pixel in the representative images: Paints in the CAVE dataset of mathematical optimization-based methods: CSU [45], NSSR [46] and deep learning-based methods: uSDN [77], DUFL [109], DHSIS [57] and our method on the CAVE dataset with the up-scale factor 16.

on any hand-crafted prior and previously collected training triplets. Via leveraging the designed architecture of generative network itself for capturing the prior of the underlying structure in the latent HR-HS image, we employed the observed LR-HS and HR-RGB images only for network parameter learning. Further, we implemented the degradation models in a learnable manner inside our proposed framework and is prospected to be used flexibly in different real scenarios. Experiments on two benchmark HS image datasets validated that the proposed DSSH method manifested very impressive reconstruction performance, and even better than most state-of-the-art supervised learning approaches.

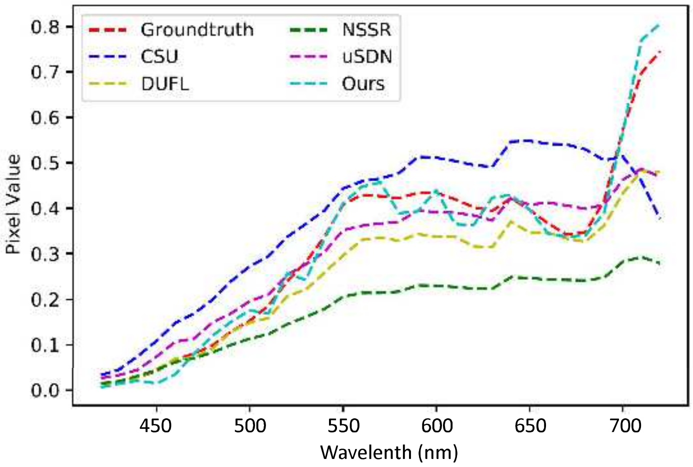


FIGURE 4.5: The recovered spectral curve of one pixel in the representative images: imgb4 in the Harvard dataset of mathematical optimization-based methods: CSU [45], NSSR [46] and deep learning-based methods: uSDN [77], DUFL [109], DHSIS [57] and our method on the CAVE dataset with the up-scale factor 16.

Chapter 5

Unsupervised Internal Learning of Hyperspectral Image Super-Resolution

Inspired by the fact that natural images have strong internal data repetition and the cross-scale internal recurrence, Shocher et al. [111] exploited a zero-shot super-resolution (ZSSR) for the RGB images, and aimed to learn an image-specific CNN model for each under-studying test data. Via extracting the LR-HR pairs from the down-sampled versions of the LR image and itself as training samples, ZSSR trained a CNN model to infer the complicated image-specific LR-HR relations, and then applied the learned relations (model) on the LR observation to provide the HR estimation. However, ZSSR has to down-sample the observation to lower-resolution data for extracting training pairs and would lead to a very limited amount of training samples, especially for large-upscale SR problems. As a result, ZSSR has usually applied for super-resolving the LR observation with small up-scale factors such as 2 or 4. It is well known that the up-scale spatial factor in the HSI SR scenario is required to be very large such as from 8 to 32, and then the naive adaptation of ZSSR to the HSI SR task would be impractical. Moreover, the essential attributes in HS images are the owed detail spectral distribution potentially for distinguishing the materials with a subtle difference, and thus lift the spectral fidelity in HSI SR task should be the concentrated aspect. With regard to naively adopting the internal spatial recurrence like in the ZSSR paradigm, the down-sampling operation on the observations usually causes severe spectral mixing of the surrounding pixels, and thus the deviation of the spectral mixing levels at the training phase and test phase would be great large. This domain shift in HSI SR possibly degrades the super-resolved performance in real experiments.

To overcome the above limitations, this study proposes a novel deep internal and self-supervised learning framework for HSI SR, which is image-specific generative model of generalized internal learning (ISGM (GIL)). On one hand, given the observations: the HR-HS and HR-RGB images for a specific scene, similar in ZSSR we down-sample them into LR-HS and HR-RGB sons and then extract the training triplets from the son images and the original LR-HS to conduct deep internal learning. On the other hand, we further extract the input pairs from the original observations: LR-HS and HR RGB images as training samples without the ground-truth to conduct deep self-supervised learning. To effectively leverage the unlabeled training samples, we specifically design the degradation blocks to transform the predicted HR-HS image to the LR-HS and HR-RGB images, which are used to formulate the loss function for network training. The proposed internal and self-supervised learnings are aggregated into a unified framework, where the deep network with an encoder-decoder architecture is shared for both branches of learning. By integrating the self-supervised learning without ground-truth label with the 'supervised ' internal learning, the domain shift caused by different spectral mixing levels can be expected to be significantly mitigated, and thus greatly lift the spectral recovering performance in the HSI SR task. We conduct extensive experiments on two benchmark HSI datasets, and demonstrate the significant superiority of our method over SoTA CNN-based HSI SR methods.

In summary, our main contributions are three-fold:

1) We present a novel deep internal learning method for unsupervised HSI SR, which extracts the training triplets from the down-sampled versions of the observations and the LR-HS image to train a specific CNN model for the under-studying scene.

2) We exploit deep self-supervised learning via leveraging the observed LR-HS and HR-RGB images without the corresponding ground-truth as the complementary training samples which can potentially mitigate the domain shift, especially for the great gap of the spectral mixing levels between the training samples in internal learning and the ongoing HR-HS prediction from the observations. Specifically, we design the special convolution blocks to implement the degradation operations inside the learning framework, and then produce the estimations of the observed LR-HS and HR-RGB images from the predicted HR-HS image to construct a loss function for network learning.

3) We combine the internal and self-supervised learning into a unified end-to-end framework for HSI SR. In detail, we adopt a weight-shared network with encoder-decoder architecture for both internal and self-supervised learning, where following the output of the self-supervised learning we append the convolution-based degradation blocks and investigate a joint optimization strategy for network training.

5.1 Deep Supervised External Learning

In recent years, deep learning approaches have been widely studied for HSI SR tasks to automatically learn a common model for any observed LR-HS/HR-RGB pair without manual exploration of the image priors, and demonstrate remarkable performance gain over the traditional prior-based methods. Han et al. [71] firstly stacked the up-sampled LR-HS and the HR-RGB images together, and then adopted a simple 3-layer CNN to estimate the latent HR-HS image. Later, more complex CNN architectures with residual structure and dense connection [72] have been proposed for boosting SR performance. Palsson et al. [73] proposed performed a 3D CNN architecture to perform MS/HS fusion by first dimensionality reduction for decreasing computational time. Dian et al. [57] exploited a combined optimization and learning method for HSI SR, which firstly obtained an initial HR-HS estimation via solving a Sylvester equation, and then adopted a CNN network to refine the initial result. More recently, Wang et al. [74] proposed a coarse-to-fine HS image learning procedure by iteratively exploring the relationship between the target and the observations, and illustrated great performance improvement. Moreover, Han et al. [55] focused on handling the extreme difference issue of the spatial structure in two modalities of observations, and proposed a multi-scale and multi-level fusion learning framework. All of the mentioned networks are implemented in a fully-supervised way, and are needed to be trained using a large number of external samples, which consist of not only the easily captured LR-HS/HR-RGB images but also the high-cost HR-HS images. Furthermore, although the learned models potentially have the common modeling capability for dealing with any observation while cannot take account of the specific attributes for an under-study scene.

5.2 Zero-Shot Learning

Zero-shot learning is a type of machine learning technique where a model can recognize and classify objects or concepts it has never seen before by leveraging prior knowledge or information. In other words, the model can perform well on new tasks without requiring any training data or examples specific to that task. This is achieved by training the model to understand the relationships and similarities between different classes or concepts. For example, if a model has been trained on a dataset of animal images and their corresponding labels, it can still classify a new animal it has never seen before by using its understanding of the relationships between different animal categories (e.g. mammals, birds, reptiles) and the features they share (e.g. fur, wings, scales). Zero-shot learning has applications in various fields, such as natural language processing, computer vision, and robotics, where it can be used to improve the accuracy and efficiency of models by reducing the need for large amounts of training data. Recently, zero-shot learning is utilized in image processing, especially for natural images super-resolution. Zero-shot learning can be applied to natural image super-resolution, which is the process of generating high-resolution images from low-resolution inputs. In traditional approaches, super-resolution models are trained on pairs of high-resolution and low-resolution images to learn the mapping between them. However, zero-shot learning can be used to enhance super-resolution models by leveraging prior knowledge about the image content and structure. This is achieved by training the model on a set of high-resolution images and their corresponding attributes, such as edges, textures, and patterns, without explicitly providing low-resolution images. During inference, the model can then generate high-resolution images from low-resolution inputs by using its understanding of the relationships between different image features and their corresponding attributes. For example, if the model has learned that a certain edge pattern in a high-resolution image corresponds to a particular texture, it can use this knowledge to enhance the texture details in a low-resolution image with a similar edge pattern. Zero-shot learning in natural image super-resolution has the potential to improve the quality and accuracy of super-resolved images, particularly in scenarios where obtaining pairs of high- and low-resolution images for training is difficult or time-consuming.

5.3 Deep Internal Learning

Inspired by the strong internal data repetition and the cross-scale internal recurrence in a natural image, Shocher et al. [111] proposed a deep internal learning network for the RGB images, dubbed as zero-shot super-resolution (ZSSR). The ZSSR method aimed to learn an image-specific CNN model for each under-studying data via synthesizing the internal training samples, i.e. the internal LR-HR pairs from the LR image and its down-sampled version. This method was specifically proposed for natural RGB image super resolution, and demonstrated promising performance for super-resolving the LR observation with small up-scale factors such as 2 and 4. This study aims to integrate the internal learning for the HSI SR task. However, in the HSI SR scenario, the up-scale factor is usually large such as from 8 to 32, and then the synthesized internal training samples extracted from the observations and their down-sampled versions would be extremely small. The network training with the internal samples would not generate a model with sufficient modeling capability. Thus, the naive adaptation of the existing internal learning to the HSI SR task possible cannot work well. In addition, the beneficial attributes in HS images are the detail spectral distribution for effectively distinguishing the materials with a subtle difference, and thus enhancing the spectral reliability in HSI SR task should be more important. Whilst the down-sampling of the observation in the conventional internal learning usually causes severe spectral mixing in the synthesized samples, and then results in heavy domain shift between in training and test phases, which greatly degrades the super-resolved performance in real experiments. Therefore, this study proposes to leverage the observations without the ground-truth as the training samples for self-supervised learning, and guides the internal network learning to proper direction.

5.4 Proposed Image-Specific Generative Model of Generalized Internal Learning

Instead of utilizing hand-crafted image priors, the deep learning networks can automatically learn the intrinsic image priors hidden in the training data and have been successfully applied for HSI SR tasks to provide superior SR performance. The current dominated research paradigm mainly leverages the previously collected external dataset to off-line learn a common model while several works realize the deep learning framework in an unsupervised way with the help of the mathematical relation between the observations and the required HR-HS image. In this subsection, we survey both deep supervised external learning and unsupervised learning methods. In this section, we first introduce the formulation of the HSI SR problem to merge the observed LR-HS and HR-RGB images, and describe the motivation for our method. Then, we present our unsupervised CNN-based method for the HSI SR, which can build a specific model by conducting the internal learning and complementary self-supervised learning with the observed images only.

5.4.1 Motivation

In the conventional fully-supervised CNN based methods, it is necessary to previously build the external dataset including large number of training triplets $\{\mathbf{x}_n, \mathbf{y}_n, \mathbf{z}_n\}_{n=1}^{N}$, where $\mathbf{x}_n$ and $\mathbf{y}_n$ are $n-th$ LR-HS and HR-RGB images as the CNN network inputs while $\mathbf{z}_n$ is the corresponding label, and then learn an off-line HSI-SR model by minimizing the reconstruction errors of the external training HR-HS samples as the follows:

$$\theta^* = \arg\min_{\theta} \sum_{n=1}^{N} \|\mathbf{z}_n - f_{\theta}^{CNN}(\mathbf{x}_n, \mathbf{y}_n)\|_2^2, \tag{5.1}$$

where θ is the network parameters to be optimized. After finishing the network learning at the training phase, the observed LR-HS and HR-RGB images: $\mathbf{x}_t$, $\mathbf{y}_t$ of an arbitrary test scene are inputted to the CNN model with the fixed parameters θ^* to produce the corresponding HR-HS estimation: $\mathbf{z}_t = f_{\theta^*}^{CNN}(\mathbf{x}_t, \mathbf{y}_t)$.

Unlike the supervised paradigm above, this study leverages the observed images only to train a specific CNN for the under-studying data via synthesizing the training triplets from the observation and their down-sampled versions for 'supervised' learning and producing un-labeled training samples from the observations only for unsupervised learning, which can also be called as internal and self-supervised learning, respectively. Fundamental of the internal learning is the fact that there is strong internal data repetition in natural images, such as a large number of repeated small patches inside a single image within the same scale as well as across various scales. Thus, on the one hand, we take the observed LR-HS image as the HR-HS training label and down-sample the observed HR-RGB and LR-HS images to synthesize its corresponding training inputs, which have been verified to possess similar super-resolving relations to some extent in spite of different resolution scales. On the other hand, in contrast with the required high-fidelity on the spatial structures in the natural image SR task, the HSI SR expects more to recover the high-reliable spectral characteristics. However, the spectral mixing levels may be significantly made heavier by down-sampling the observed LR-HS image, which itself has a very low spatial resolution, and thus results in a great gap of the spectral mixing between the internal samples and the true un-available samples in the real HSI SR scenario. Then, this study appeals to the un-labeled observations, and uses them as the complementary training samples without ground-truth to conduct self-supervised learning, which is expected to bridge the great gap of the spectral mixing level between the internal samples and the true-scale samples. The conceptual architecture of our deep internal and self-supervised learning framework is shown in Fig. 5.1, which mainly consists of two branches of training flows with the shared encode-decode network. In detail, we adopt a simple encoder-decoder network architecture for learning multi-level contexts and conducting fusion between encode-decoder paths. Concretely, the encoder and decoder paths, contain the same number of blocks, and a point-wise convolution-based bridge (PWCB) between the corresponding blocks of the two paths is adopted to transfer the learned detail features of the encoder to the decoder path. Each block in the encoder path is composed of 2 convolution layers with kernel size $3 * 3$, following the batch-normalization and LeakyRELU activation layers, where the first convolution has the stride parameter 2 to decrease the feature map size of the previous one to half. In addition, the block in the decoder path firstly concatenates the transfered feature of the PWCB and the up-sampled feature of the previous decoder's block, which doubly recovers the feature map size between the adjacent decoder blocks, and then serially employ two convolution layers with kernel sizes $3 * 3$ and $1 * 1$, respectively, for feature learning. Finally, a convolution reconstruction layer is used for predicting the latent target as the output of the specific CNN model. According to the above description, we aim to learn a specific CNN model for each under-studying scene via leveraging the observed images and their transformations. Generally, following the similar expression of Eq. 5.1, we formulate the loss function of our internal and self-supervised learning as:

$$\begin{aligned}\theta^* = &\arg\min_{\theta} \|\mathbf{x}_t - T_2(f_{\theta}^{SCNN}(T_1(\mathbf{x}_t, \mathbf{y}_t)))\|_2^2 \\ &+ \|\mathbf{y}_t - T_3(f_{\theta}^{SCNN}(T_1(\mathbf{x}_t, \mathbf{y}_t)))\|_2^2,\end{aligned} \tag{5.2}$$

where $T_1(\cdot)$, $T_2(\cdot)$, and $T_3(\cdot)$ represent the optional spatial or spectral transformations according to the internal and self-supervised learning while f^{SCNN} denotes the transformation of the specific CNN model. In the following subsection, we substantiate the implementation of the transformation operations and the concrete architecture of the specific CNN model.

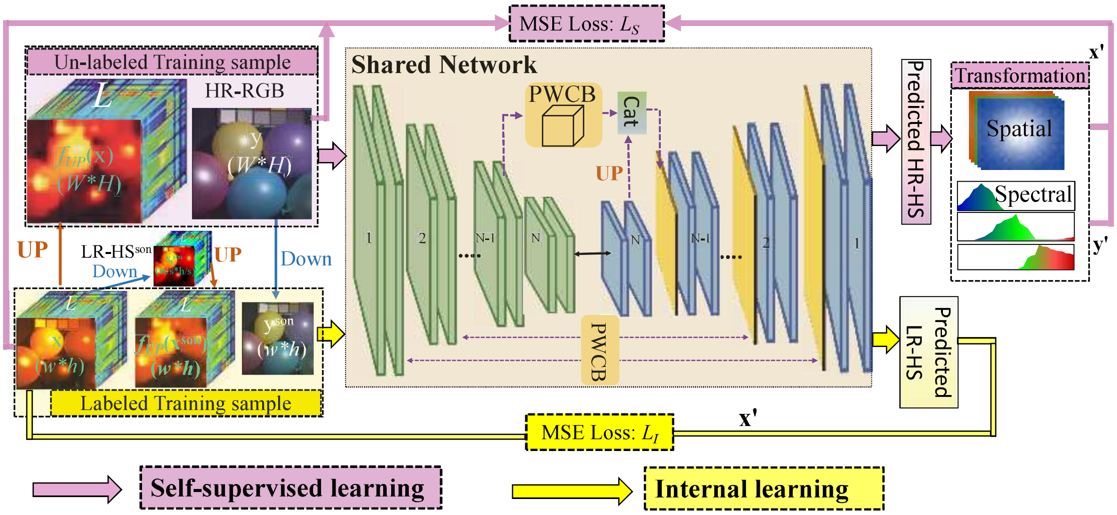


FIGURE 5.1: Conceptual diagram of the proposed ISGM (GIL) method.

5.4.2 Unsupervised Internal Learning: UIL

Our unsupervised internal learning (UIL) aims to combine the powerful modeling capability of CNN and the internal recurrence characteristics inside a single image, and construct an image-specific CNN model for super-resolving this under-studying image with no external training samples. Specifically, we train the CNN on examples extracted from the test image itself. Such examples are produced by down-sampling the observed LR-HS and HR-RGB images: $\mathbf{x}_t$, $\mathbf{y}_t$, to synthesize the lower-resolution versions of themselves, $\mathbf{x}_t^{son}$, $\mathbf{y}_t^{son}$, which are expressed as the follows:

$$(\mathbf{x}_t^{son}, \mathbf{y}_t^{son}) = (\mathbf{x}_t \downarrow^s, \mathbf{y}_t \downarrow^s) \tag{5.3}$$

where $\downarrow^s$ denotes the spatial down-sampling operation of the s scale factor as the desired one in the HSI SR task. Then, we can obtain the pseudo-supervised sample triplets $(\mathbf{x}_t^{son}, \mathbf{y}_t^{son}, \mathbf{x}_t)$ to train our specific CNN model just like in the fully-supervised learning. From Eq. 5.2, we set $T_1(\mathbf{x}_t, \mathbf{y}_t)$ as $(\mathbf{x}_t^{son}, \mathbf{y}_t^{son})$ while deleting $T_2(\cdot)$ since the corresponding training label $\mathbf{x}_t$ is available, and then the loss function for our UIL can be formulated as:

$$\mathcal{L}_I = \|\mathbf{x}_t - f_\theta^{SCNN}(\mathbf{x}_t^{son}, \mathbf{y}_t^{son})\|_2^2 \tag{5.4}$$

It is possible to build the specific CNN model with the UIL only, and then deploy the resulting learned CNN to the observed images: $\mathbf{x}_t$ and $\mathbf{y}_t$ as the LR input to the network for predicting the desired HR-HS output. It should be noted that the learned CNN model is fully convolutional, and hence can be directly applied to the observed images of different sizes with the training samples. However, in the HSI SR scenario, the upscale factor s is usually very large while the observed LR-HS image itself is of small size with low spatial resolution and then results in an extremely small size of the down-sampled version from the LR-HS image to extract enough training samples for the internal learning. Moreover, the spectral mixing gap between the cross-scale images would significantly degrade the spectral recovery fidelity with the internal samples. Therefore, we next integrate the self-supervised learning by using the observed images without the corresponding ground-truth as training samples.

5.4.3 Pseudo-Supervised Internal Learning: SIL

We use the observed LR-HS and HR-RGB images: $(\mathbf{x}_t, \mathbf{y}_t)$ as the un-labeled complimentary samples to guide the network learning process in the proper direction for reliable spectral recovery. In spite of the in-availability of the corresponding groundtruth of $(\mathbf{x}_t, \mathbf{y}_t)$, the degradation models from the underlying HR-HS image to the observations are usually mathematically formulated as in Eq. 3.4. Thus, this study leverages the mathematical relation of the degradation model between the underlying HR-HS image and the observed LR-HS and HR-RGB images, and implements them using specially designed convolution blocks to transform the network output into the approximated LR-HS and HR-RGB observations, which then can be applied for evaluating the reconstruction errors of the un-labeled network inputs. In the aspect of network evaluation with the inputs only, we dub this un-supervised method as deep self-supervised learning. Specifically, we design a special depthwise convolution layer with the same kernel for all spectral bands to approximate the spatial degradation operation while adopting a point-wise convolution layer to approximate the spectral transformation operation inside our network. By instituting the spatial and spectral degradation operations for $T_2(\cdot)$ and $T_3(\cdot)$ in Eq. 5.2 and leaving out $T_1(\cdot)$ with directly using $(\mathbf{x}_t, \mathbf{y}_t)$ as the inputs, the loss function for our self-supervised learning is expressed as:

$$\begin{aligned} \mathcal{L}_S = & \|\mathbf{x}_t - D_{Spat}(f_\theta^{SCNN}(\mathbf{x}_t, \mathbf{y}_t))\|_2^2 + \\ & \|\mathbf{y}_t - D_{Spec}(f_\theta^{SCNN}(\mathbf{x}_t, \mathbf{y}_t))\|_2^2, \end{aligned} \tag{5.5}$$

where D_{Spat} and D_{Spec} represent the transformations of the spatial and spectral degradation blocks in our SIL, respectively. In general, with the known imaging conditions for the observed LR-HS and HR-RGB images, we can feasibly set the kernel weights of the depth-wise convolution layer as the point spread function of the HS sensor and the kernel weights of the point-wise convolution layer as the camera spectral function of the color sensor whilst we impose the bias parameters for both layers as False. With the simple implementation using the special convolution layers for both spatial and spectral degradations, we can conduct the self-supervised learning in an end-to-end manner, and produce the specific CNN model with the un-labeled samples.

5.4.4 Image-Specific Generative Network of Generalized Internal Learning: ISGM (GIL)

Via combining unsupervised and pseudo-supervised internal learning and leveraging two kinds of data: $(\mathbf{x}_t^{son}, \mathbf{y}_t^{son}, \mathbf{x}_t)$ and $(\mathbf{x}_t, \mathbf{y}_t)$ as training samples, we propose a unified framework for simultaneously conducting UIL and SIL. The combined framework uses a shared network to carry out both internal and pseudo-supervised learning. The loss $\mathcal{L}_I$ of UIL is formulated using the training sample: $(\mathbf{x}_t^{son}, \mathbf{y}_t^{son}, \mathbf{x}_t)$ while the loss $\mathcal{L}_S$ of SIL is obtained using $(\mathbf{x}_t, \mathbf{y}_t)$. Compared with the input sample $(\mathbf{x}_t, \mathbf{y}_t)$ in the SIL, the input of $(\mathbf{x}_t^{son}, \mathbf{y}_t^{son})$ to the UIL branch has much less pixel number, and is one fraction of s^2 for s-upscale HSI SR problem. To this end, the naive integration of the losses $\mathcal{L}_I$ and $\mathcal{L}_S$ would significantly decrease the impact of the UIL on the learned model compared with the SIL. To mitigate this issue, we firstly augment the pseudo supervised samples $(\mathbf{x}_t^{son}, \mathbf{y}_t^{son}, \mathbf{x}_t)$ using flipping and rotation operations, and incorporate all augmented samples for formulating the UIL loss. The total loss of our overall learning method is formulated as:

$$
\begin{aligned}
\mathcal{L}_{total} =& \mathcal{L}_I + \mathcal{L}_S \\
=& \|\mathbf{x}_t - f_\theta^{SCNN}(\mathbf{x}_t^{son}, \mathbf{y}_t^{son})\|_2^2 \\
&+ \|\mathbf{x}_t - D_{Spat}(f_\theta^{SCNN}(\mathbf{x}_t, \mathbf{y}_t))\|_2^2 \\
&+ \|\mathbf{y}_t - D_{Spec}(f_\theta^{SCNN}(\mathbf{x}_t, \mathbf{y}_t))\|_2^2
\end{aligned}
\tag{5.6}
$$

As shown in Eq. 5.6, our deep ISGM of generalized internal learning (GIL) adopts a shared network to carry out both UIL and SIL. Note that the conventional supervised CNNs for HSI SR, which is trained on a large-scale external dataset of LR-HS, HR-RGB and HR-HS triplets, have to capture the rich diversity of all potential relations among the observations and the target, and thus these supervised methods prefer much deeper and more complex network architectures. In contrast, the estimation relations from the observed LR-HS and HR-RGB images to its corresponding HR-HS image for a specific scene is significantly simpler, and hence could be well modeled by a much shallower and simpler network structure. In our experiments, we adopt a simple encoder-decoder network architecture for the specific CNN model f_θ^{SCNN}. In detail, the encoder and decoder paths, respectively, contain 5 blocks, and the skip connections are used to bridge between the corresponding blocks of the two paths for reusing the learned detail features of the encoder. Each block in both paths is composed of 3 convolution layers, following the RELU activation function. A max-pooling layer with a 2×2 kernel is adopted to decrease the feature map size to half between adjacent encoder blocks whilst an up-sampling layer is used to doubly recover the feature map size between the adjacent decoder blocks. Finally, we reconstruct the latent target using a convolution output layer. Moreover, in both unsupervised and pseudo-supervised internal learning, there are two available modalities of data: $(\mathbf{x}_t^{son}, \mathbf{y}_t^{son})$ or $(\mathbf{x}_t, \mathbf{y}_t)$ with very large differences in the spatial resolution, and cannot employ equal operations on them as the network input. Therefore, we first conduct a simple up-sampling on the HS image with lower spatial resolution $\mathbf{x}_t^{son}$ $(\mathbf{x}_t)$ to the same spatial size with the RGB image $\mathbf{y}_t^{son}$ $(\mathbf{y}_t)$, and then concatenate them together as the input to our network:

$$
\begin{aligned}
\mathbf{xy}_{son} &= f_{concat}(f_{UP}(\mathbf{x}_{son}), \mathbf{y}_{son}) \\
\mathbf{xy} &= f_{concat}(f_{UP}(\mathbf{x}), \mathbf{y}),
\end{aligned}
\tag{5.7}
$$

where f_{concat} and f_{UP} represent the concatenating and up-sampling transformation, respectively. After a predefined iteration of network training for our specific CNN model, the concatenated data $\mathbf{xy}$ is inputted to the network to predict the latent HR-HS image.

5.5 Experiment Results

In this section, we will conduct extensive experiments to demonstrate the effectiveness of our proposed deep internal and self-supervised learning method. We first introduce the experimental setting, including the same used datasets and evaluation metrics, and then provide the comparisons with the state-of-the-art (SoTA) methods and the ablation study.

5.5.1 Comparisons with the State-of-the-art Methods

To verify the effectiveness of our ISGM (GIL), we compare the HSI SR performance with different SoTA paradigms including the unsupervised prior-based methods: Generalization of Simultaneous Orthogonal Matching Pursuit (GOMP) [96], Sparse Non-negative Matrix Factorization (SNNMF) [97], Bayesian sparse representation (BSR) [96], Non-Negative Structured Sparse Representation (NSSR) [46] and couple spectral unmixing (CSU) [93], supervised deep learning methods: SSFNet [56], ResNet [102], DHSIS [57], and unsupervised deep learning methods: uSDN [77], DHP [59], DUFL [109], GDD [81]. The compared results with the scale factors $s = 8$ and $s = 16$ for both CAVE and Harvard datasets are shown in TABLE 5.1. The up arrow in TABLE 5.1 indicates that the larger the value, the better the HSI SR performance; the down arrow is the opposite. From TABLE 5.1, we can observe that our proposed method achieves the best performance than all state-of-the-art (SoTA) methods of different paradigms on the upscale 8 and 16 of both CAVE and Harvard datasets. Compared with the best unsupervised deep learning paradigm: GDD [81], our method can lift the PSNR 1.9dB/1.65dB and 3.06dB/3.8dB, respectively for the upscale factors 8/16 of both CAVE and Harvard datasets.

5.5.2 Ablation Study

As mentioned in Section 5.4, the network learning can be implemented with the synthesized internal training triplets (unsupervised internal learning: UIL), the observations without ground-truth data (pseudo-supervised internal learning: SIL) and our ISGM (GIL). We conducted experiments with different learning conditions, and provided the ablation studies for both CAVE and Harvard datasets. We extensively carried out verification using three upscale factors $s = 4, 8, 16$. Moreover, we also validated the performance w/o data augmentation on the UIL, and provided the compared results on the upscale factors: 8 and 16. All the results of the ablation studies are shown in TABLE 5.2. TABLE 5.2 (a) obviously manifests that the UIL results are in very limited performance due to the small number of synthesized training triplets and the domain shift between training and testing phases. Although the data augmentation on UIL does lift the HSI SR performance at some extent, it is far from enough compared with the SoTA methods as shown in TABLE 5.1. The SIL by taking the relation between the latent HR-HS and the observations into account can greatly improve the SR performance in spite of the unsupervised learning without any label data. The incorporation of the UIL and SIL can further boost the HS image resolved results for all upscale factors: 4, 8 and 16 in both CAVE and Harvard datasets.

Next, we validate the performance effect by varying the network architectures. As we mentioned above, we employed an encoder-decoder architecture to serve as our specific CNN model, where both encoder and decoder paths consist of multiple blocks for extracting multi-scale contexts in different receptive fields. Thus, we change the block numbers from 3 to 5, and manifest the possible performance variation. Furthermore, the learned features in the encoder path have been transferred to the decoder path using a point-wise convolution-based bridge (PWCB), where the channel number can be adjusted to turn the balance between the transferred encoder feature and the learned feature of the decoder’s previous block. We carried out the experiments by setting the channel number as 4 and 64 in the PWCB, and demonstrated the compared performance for the upscale factor 8 of the CAVE dataset as

TABLE 5.1: Compared evaluation results between unsupervised prior-based methods: BSR [96], CSU [93], GOMP [96], SNNMF [97], NSSR [46], supervised deep learning methods: SSFNet [56], ResNet [102], DHSIS [57], unsupervised deep learning-based methods: uSDN [77], DHP [59], DUFL [109], GDD [81] and our proposed method with the scale factors: 8 and 16 in CAVE and Harvard datasets.

Spatial Expanding Factor = 8											
Dataset		CAVE					Harvard				
Method		RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
Unsupervised Prior-based	GOMP	5.69	33.64	-	11.86	2.99	3.79	38.89	-	4.00	1.65
	SNNMF	1.89	43.53	-	3.42	1.03	1.79	43.86	-	2.63	0.85
	BSR	1.75	44.15	-	3.31	0.97	1.71	44.51	-	2.51	0.84
	CSU	2.56	40.74	0.985	5.44	1.45	1.40	46.86	0.993	1.77	0.77
	NSSR	1.45	45.72	0.992	2.98	0.80	1.56	45.03	0.993	2.48	0.84
Supervised Learning-based	SSFnet	1.89	44.41	0.991	3.31	0.89	2.18	41.93	0.991	4.38	0.98
	ResNet	1.47	45.90	0.993	2.82	0.79	1.65	44.71	0.984	2.21	1.09
	DHSIS	1.46	45.59	0.990	3.91	0.73	1.37	46.02	0.981	3.54	1.17
Unsupervised Learning-based	uSDN	4.37	35.99	0.914	5.39	0.66	2.42	42.11	0.987	3.88	1.08
	DHP	7.60	31.40	0.871	8.25	4.20	7.94	30.86	0.803	3.53	3.15
	DUFL	2.08	42.50	0.975	5.36	1.16	2.38	42.16	0.965	2.35	1.09
	GDD	1.68	44.22	0.987	3.81	0.96	1.30	47.02	0.990	1.94	0.90
	ISGM (GIL)	1.39	46.10	0.993	3.12	0.77	1.00	50.08	0.995	1.47	0.56
Spatial Expanding Factor = 16											
Unsupervised Prior-based	GOMP	6.08	32.96	-	12.60	1.43	3.85	38.56	-	4.16	0.77
	SNNMF	2.45	42.21	-	4.61	0.66	1.93	43.31	-	2.85	0.45
	BSR	2.36	41.57	-	4.57	0.58	1.93	43.56	-	2.74	0.42
	CSU	2.87	39.83	0.983	5.65	0.79	1.60	45.50	0.992	1.95	0.44
	NSSR	1.78	44.01	0.990	3.59	0.49	1.65	44.51	0.993	2.48	0.41
Supervised Learning-based	SSFnet	2.18	41.93	0.991	4.38	0.98	1.94	43.56	0.980	3.14	0.98
	ResNet	1.93	43.57	0.991	3.58	0.51	1.83	44.05	0.984	2.37	0.59
	DHSIS	2.36	41.63	0.987	4.30	0.49	1.87	43.49	0.983	2.88	0.54
Unsupervised Learning-based	uSDN	3.60	37.08	0.969	6.19	0.41	9.31	39.39	0.931	4.65	1.72
	DHP	11.31	27.76	0.805	10.66	3.09	10.38	38.44	0.754	4.57	2.08
	DUFL	2.61	40.71	0.967	6.62	0.70	2.81	40.77	0.953	3.01	0.75
	GDD	2.12	42.24	0.983	4.41	0.61	1.66	44.64	0.986	2.50	0.64
	ISGM (GIL)	1.73	44.17	0.990	3.73	0.48	1.14	48.84	0.994	1.68	0.32

TABLE 5.2: Ablation studies with the three learning strategies: UIL, SIL and ISGM (GIL) and w/o data augmentation of the UIL for different upscale factors: 4, 8, 16 in both CAVE and Harvard datasets.

(a) Comparisons with the three learning strategies: UIL, SIL and ISGM (GIL) and w/o data augmentation for the upscale factors: 8 and 16.

Spatial Expanding Factor = 8										
Dataset	CAVE					Harvard				
Method	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
UIL	8.78	29.73	0.868	17.58	4.71	8.93	30.80	0.926	8.31	3.54
SIL	1.46	45.47	0.992	3.27	0.81	1.16	48.37	0.993	1.74	0.79
GIL	1.40	45.92	0.993	3.14	0.79	1.01	49.95	0.995	1.48	0.56
UIL + Aug	7.75	31.16	0.870	17.49	3.81	7.84	31.36	0.934	8.42	3.57
ISGM (GIL)+Aug	1.39	46.10	0.993	3.12	0.77	1.00	50.08	0.995	1.47	0.56
Spatial Expanding Factor = 16										
UIL	16.65	24.64	0.785	21.72	4.69	12.25	28.04	0.899	9.85	2.50
SIL	1.83	43.50	0.989	3.92	0.51	1.31	47.19	0.992	1.98	0.46
GIL	1.77	43.89	0.990	3.79	0.49	1.19	48.44	0.994	1.77	0.33
UIL + Aug	11.07	27.88	0.832	20.45	2.97	10.60	28.44	0.908	9.13	2.38
ISGM (GIL)+Aug	1.73	44.17	0.990	3.73	0.48	1.14	48.84	0.994	1.68	0.32

(b) Comparisons with the three learning strategies: UIL, SIL and ISGM (GIL) for the upscale factor 4.

Spatial Expanding Factor = 4										
Dataset	CAVE					Harvard				
Method	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
UIL	9.51	31.70	0.884	15.73	9.84	4.24	36.69	0.968	5.57	4.18
SIL	1.19	47.32	0.994	2.91	1.32	0.99	49.71	0.994	1.52	1.17
ISGM (GIL)	1.10	47.91	0.995	2.80	1.26	0.87	51.79	0.996	1.32	0.97

TABLE 5.3: Ablation studies for network architectures.

Block Number	Channel Number in PWCB: 4					Channel Number in PWCB: 64				
	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
3	1.38	46.07	0.993	3.10	0.78	1.37	46.19	0.993	3.06	0.77
4	1.35	46.23	0.993	3.08	0.75	1.35	46.31	0.993	3.00	0.77
5	1.39	46.10	0.993	3.12	0.77	1.36	46.23	0.993	3.01	0.77

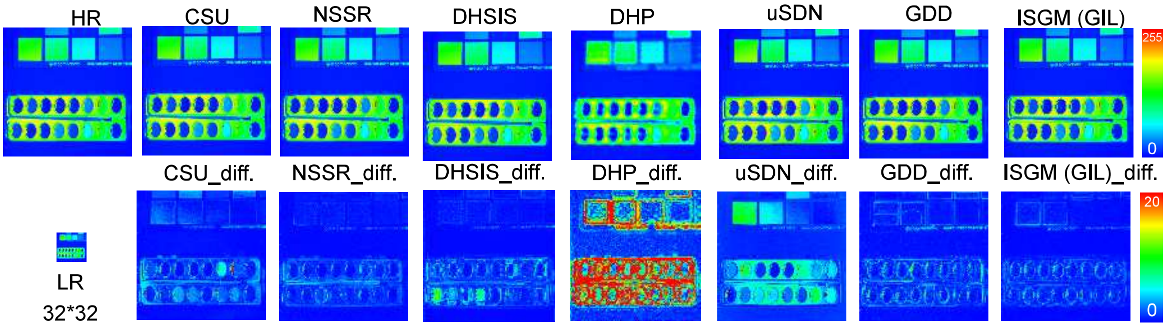


FIGURE 5.2: Visualization error map results of an example image: paints from CAVE dataset compared with unsupervised prior-based methods: CSU [93], NSSR [46], supervised deep learning methods: DHSIS [57], unsupervised deep learning-based methods: DHP [59], uSDN [77], GDD [81] and the proposed method.

shown in Table 5.3. From Table 5.3, it can be seen that the best reconstruction performance is achieved when the block number of the network is set as 4. And the channel number increasing in the PWCB from 4 to 64 has only a slight positive effect on the reconstruction performance, and some results are almost same with different channel numbers. However, a larger channel number would increase the number of parameters, which results in a heavier computational cost. Therefore, from Table 5.3, it can be concluded that the block in the encoder/decoder paths and channel in the PWCB has the suitable number 4 for providing acceptable reconstruction performance.

5.5.3 Perceptual Quality

We also provide some compared visualization results with the unsupervised prior-based CSU [93] and NSSR [46], the supervised deep learning-based DHSIS [57] and the unsupervised deep learning-based methods: uSDN [77], DHP [59], GDD [81] in Fig. 5.2 and 5.3, which further verify that our proposed ISGM (GIL) method has achieved great performance improvement compared with state-of-the-art methods.

5.6 Conclusion

This chapter investigated a novel unsupervised learning network for multispectral and hyperspectral image fusion by conducting both internal and self-supervised learnings. Specifically, we first down-sampled the observations and produced the

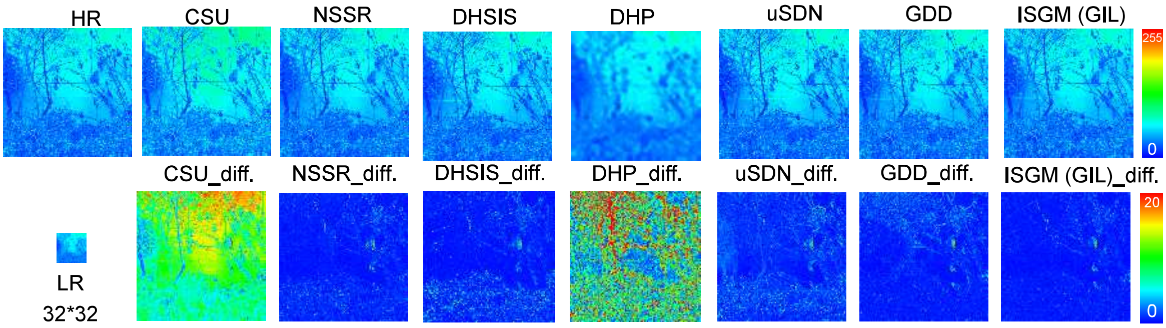


FIGURE 5.3: Visualization error map results of an example image: imgb4 from Harvard dataset compared with unsupervised prior-based methods: CSU [93], NSSR [46], supervised deep learning methods: DHSIS [57], unsupervised deep learning-based methods: DHP [59], uSDN [77], GDD [81] and the proposed method.

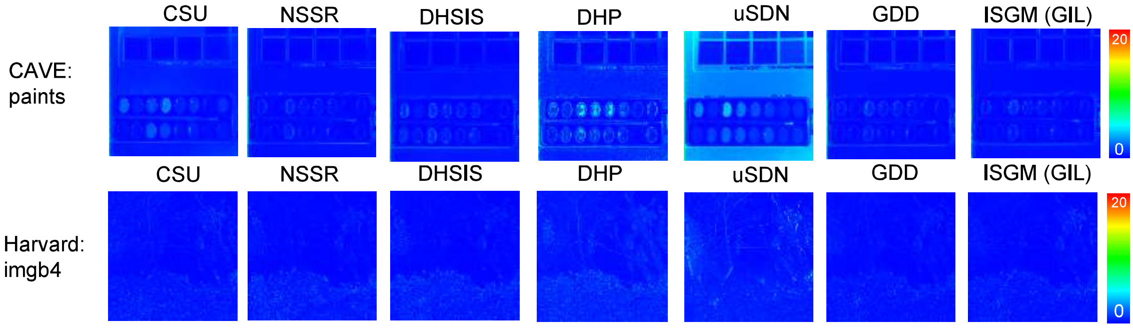


FIGURE 5.4: Visualization SAM results of example images from CAVE and Harvard datasets compared with unsupervised prior-based methods: CSU [93], NSSR [46], supervised deep learning methods: DHSIS [57], unsupervised deep learning-based methods: DHP [59], uSDN [77], GDD [81] and the proposed method.

LR-HS son, HR-RGB son images to synthesize the training triplets, which are intuitively adopted to train a specific CNN model for the under-studying image. To increase the robustness of the specific CNN, we further leveraged the observed data without the ground-truth image to carry out unsupervised learning, which can tune the rough internal CNN revolving to the optimal parameter space. Extensive experiments on two benchmark datasets demonstrated that our proposed method achieved a significant improvement compared with the SoTA HSI SR methods.

Chapter 6

Conclusion

Hyperspectral (HS) imaging can capture the detailed distribution in the spectral direction, and obtain an abundant spectral signature with dozens or even hundreds of bands at each spatial position of a scene, which greatly benefits performance improvement in various HS processing systems. However, existing HS imaging sensors usually get the HS data in a low spatial resolution and greatly restrict the wide applicability in the reality. Thus, generating a high-resolution hyperspectral (HR-HS) by merging the degraded observations: a low-resolution hyperspectral (LR-HS) image and a high-resolution Multispectral/RGB (HR-MS/RGB) image, called as HS image super resolution (HSI SR). Depending on the reconstruction principle, HSI SR is divided into two main categories: traditional mathematical model-based methods and deep supervised learning-based methods. For the mathematical model-based methods, most HSI SR methods aim to explore various hand-crafted priors for regularizing the mathematical model, and employ optimization procedures to solve this problem. Specifically, such methods mainly focus on constructing a mathematical formula to model the degradation procedure of HR-HS images into LR-HS images and HR-RGB images. Since the known variables in the observed LR-HS/HR-RGB images is much less than the underestimation in the latent HR-HS image, this task is a severely ill-posed problem, and direct optimization of the formulated mathematical model would lead to a very unstable solution. Therefore, existing methods often exploit various priors to regularize the mathematical model, i.e. imposing constrain on the solution space. Although the improvements to some extent have been achieved by elaborating the hand-crafted priors, the super-resolving performances are usually unstable according to the content of the under-studying images, and heavy spectral distortion may be caused due to the insufficient representative capability of the empirically designed priors. For the deep supervised learning-based methods, motivated by the tremendous success of the DCNN on different vision tasks, DCNN-based methods have been proposed for the HSI SR task to automatically learn the inherent priors in the latent HR-HS image. Although the reconstruction performance has remarkably progressed, all the above DCNN-based methods are required to be trained with large-scale external datasets including the degraded LR-HS/HR-RGB images and their corresponding HR-HS images, which are difficult to be collected especially for the HSI SR scenario. To solve these problems, we proposed three frameworks to achieve the goal of unsupervised hyperspectral image super-resolution. Overall, the main contributions of this dissertation are three-fold and summarized as follows.

In Chapter 3, we proposed a deep unsupervised fusion learning framework for HSI SR. This chapter suggests an unsupervised framework to automatically generate HS target images using only LR-HS and HR-RGB observations without using an external training database. The framework is motivated by the fact that convolutional neural networks have a significant amount of underlying image statistics (a prior) and are more likely to generate images with regular spatial structure and spectral patterns than noisy data. We specifically look into two HS image generation paradigms: To train HR-HS targets from a data generation perspective, 1) randomly chosen noise is used as input to the generative network, and 2) background fusion of LR-HS and HR-RGB observations is used as input to the generative network to reconstruct targets from a self-supervised learning viewpoint. By focusing on maximizing the generative network's parameters rather than the initial HR-HS aim, both approaches can produce a priori models that are automatically tailored to the object scene under investigation. To construct our generative network and create target HR-HS images from noisy or merged backdrop backgrounds, specifically, we employ an encoder-decoder architecture. Assuming the methods for the underestimated LR-HS and HR-RGB observations' spatial and spectral degradation are known, we can then produce approximations of the observations from the degraded generated HR-HS images, which can be intuitively used to derive the observation reconstruction error as a network training loss function. Our unsupervised learning framework not only enables us to model the prior information of the specific scene under study to reconstruct a trustworthy HR-HS estimate without the need for external datasets, but it also readily adapts to observations made under various imaging conditions, which can be accomplished naively by altering the degradation operations in our framework.

In Chapter 4, we proposed a novel blind learning method for unsupervised HSI SR. Understanding the spatial and spectral degradation mechanisms is necessary for deeply unattended HSI SR. The spatial blur kernel in LR-HS imaging and the camera spectral response function (CSF) in RGB sensors are two degradation processes that are difficult for regular users to understand in depth because they result from the various optical designs of HS imaging devices and RGB cameras. Furthermore, particular estimations of degradation processes under various imaging circumstances may further skew the outcomes. This makes it challenging to learn something about the deterioration of each scene under investigation in practical applications. In this work, a unique unsupervised blind technique is employed to automatically learn degradation parameters simultaneously and construct a grid in order to address the aforementioned issues. We specifically suggest three strategies to address various issues, based on the unknown components: 1) a spatially blind technique that, when the LR-HS observation is finished, automatically learns the spatial blur kernel since the RGB sensor's CSF is known; 2) a spectrally blind technique that, after the HR-RGB observation is over but the burr kernel of the HS image is known, automatically learns the CSF transformation matrix; 3) A totally blind technique that learns both the CSF matrix and the spatial blur kernel. In order to execute the spatial and spectral decomposition processes, we have developed specific convolutional layers based on our previously presented unsupervised system. The parameters of the layers are automatically processed as the weights of the learnt sum kernels and CSF matrices. The spectral decomposition procedure has been implemented using a pointwised convolutional layer in output channel 3 to obtain approximations of HR-RGB images, while the spatial decomposition procedure has been implemented using a deep convolutional layer in which the kernels of the various spectral channels are identical and the range parameters are defined as extended scale factors. In order to simultaneously learn the unique pre- and degradation knowledge of the HR-HS images and to construct a highly generic HSI SR system, an integrated framework has been constructed using the learning implementation of the degradation mechanism. Furthermore, the suggested framework can be used uniformly to various types of blind HSI SR and is extremely scalable to arbitrary HSI SR observations by employing the applied convolution parameters as known blur kernels or CSFs.

In Chapter 5, we proposed a generalized internal learning method for unsupervised HSI SR. We synthesize a labeled training triplet using only LR-HS and HR-RGB observations and use them as training data for supervised and unsupervised learning along with the unlabeled observations to build a more potent image-based CNN model for the under-utilized HR-HS data since natural images have strong intrinsic and internal recurrence at various scales. We constructed training triples of LR-HS/HR-RGB subversions and LR-HS observations that have the same correlation with LR-HS/HR-RGB observations and HR-HS objects, despite their varied resolutions, by first reducing the observed LR-HS and HR-RGB pictures to their subversions. It is feasible to train an image-specific CNN model for the HR-HS object using artificial training examples. This process is known as internal learning. However, there are rarely many synthetically labeled training samples, particularly for large spatial expansion factors, and further reducing LR-HS observations results in significant spectral blurring of neighboring pixels, leading to biased spectral blurring levels in the training and testing phases. As a result, these restrictions could make naïve internal learning super-resolution less effective. We present a generalized internal learning technique for more trustworthy HR-HS image reconstruction in order to overcome these constraints. This method combines naïve internal learning with our self-supervised learning method for unsupervised HSI SR.

Bibliography

[1] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[2] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[3] C. A. Bishop, J. G. Liu, and P. J. Mason, "Hyperspectral remote sensing for mineral exploration in pulang, yunnan province, china," *International Journal of Remote Sensing*, vol. 32, no. 9, pp. 2409–2426, 2011.

[4] J.-L. Xu, C. Riccioli, and D.-W. Sun, "Comparison of hyperspectral imaging and computer vision for automatic differentiation of organically and conventionally farmed salmon," *Journal of Food Engineering*, vol. 196, pp. 170–182, 2017.

[5] G. Lu and B. Fei, "Medical hyperspectral imaging: a review," *Journal of biomedical optics*, vol. 19, no. 1, p. 010901, 2014.

[6] K. Jensen and D. Anastassiou, "Spatial resolution enhancement of images using nonlinear interpolation," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1990, pp. 2045–2048.

[7] G. Licciardi, G. Vivone, M. D. Mura, R. Restaino, and J. Chanussot, "Multiresolution analysis techniques and nonlinear pca for hybrid pansharpening applications," *Multidimensional Systems and Signal Processing*, vol. 27, pp. 807–830, 2016.

[8] V. Patel and K. Mistree, "A review on different image interpolation techniques for image enhancement," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 12, pp. 129–133, 2013.

[9] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based superresolution," *IEEE Computer graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[10] J. Vrabel, "Multispectral imagery band sharpening study," *Photogrammetric engineering and remote sensing*, vol. 62, no. 9, pp. 1075–1084, 1996.

[11] M. A. Bendoumi, M. He, and S. Mei, "Hyperspectral image resolution enhancement using high-resolution multispectral image based on spectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 10, pp. 6574–6583, 2014.

[12] G. Vivone, A. Garzelli, Y. Xu, W. Liao, and J. Chanussot, "Panchromatic and hyperspectral image fusion: Outcome of the 2022 whispers hyperspectral pansharpening challenge," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 166–179, 2022.

[13] M. D. Sacchi, T. J. Ulrych, and C. J. Walker, "Interpolation and extrapolation using a high-resolution discrete fourier transform," *IEEE Transactions on Signal Processing*, vol. 46, no. 1, pp. 31–38, 1998.

[14] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6421–6433, 2019.

[15] M. Moeller, T. Wittman, and A. L. Bertozzi, "A variational approach to hyperspectral image fusion," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, vol. 7334. SPIE, 2009, pp. 502–511.

[16] Y.-H. Wang, J. Qiao, J.-B. Li, P. Fu, S.-C. Chu, and J. F. Roddick, "Sparse representation-based mri super-resolution reconstruction," *Measurement*, vol. 47, pp. 946–953, 2014.

[17] B. Tu, X. Zhang, J. Wang, G. Zhang, and X. Ou, "Spectral–spatial hyperspectral image classification via non-local means filtering feature extraction," *Sensing and Imaging*, vol. 19, pp. 1–25, 2018.

[18] W. Huang, H. H. Wang, Z. Liu, and L. Wang, "Image de-noising and enhancement based on rough set and principal component analysis," in *2017 International Conference on Inventive Computing and Informatics (ICICI)*. IEEE, 2017, pp. 536–539.

[19] Y. Choi, E. Sharifahmadian, and S. Latifi, "Performance analysis of contourlet-based hyperspectral image fusion methods," *International Journal on Information Theory*, vol. 2, no. 1, pp. 1–14, 2013.

[20] H. Gao, J.-F. Cai, Z. Shen, and H. Zhao, "Robust principal component analysis-based four-dimensional computed tomography," *Physics in Medicine & Biology*, vol. 56, no. 11, p. 3181, 2011.

[21] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[22] Z. An and Z. Shi, "Hyperspectral image fusion by multiplication of spectral constraint and nmf," *Optik*, vol. 125, no. 13, pp. 3150–3158, 2014.

[23] Z. Cai, Z. Huang, M. He, C. Li, H. Qi, J. Peng, F. Zhou, and C. Zhang, "Identification of geographical origins of radix paeoniae alba using hyperspectral imaging with deep learning-based fusion approaches," *Food Chemistry*, p. 136169, 2023.

[24] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and L. Wang, "Superpca: A superpixelwise pca approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4581–4593, 2018.

[25] E. Fakiris, G. Papatheodorou, M. Geraga, and G. Ferentinos, "An automatic target detection algorithm for swath sonar backscatter imagery, using image texture and independent component analysis," *Remote Sensing*, vol. 8, no. 5, p. 373, 2016.

[26] D. Li, F. Kong, and Q. Wang, "Hyperspectral image classification via nonlocal joint kernel sparse representation based on local covariance," *Signal Processing*, vol. 180, p. 107865, 2021.

[27] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3373–3388, 2014.

[28] C. Urbina Ortega, E. Quevedo Gutiérrez, L. Quintana, S. Ortega, H. Fabelo, L. Santos Falcón, and G. Marrero Callico, "Towards real-time hyperspectral multi-image super-resolution reconstruction applied to histological samples," *Sensors*, vol. 23, no. 4, p. 1863, 2023.

[29] J. Qu, J. Lei, Y. Li, W. Dong, Z. Zeng, and D. Chen, "Structure tensor-based algorithm for hyperspectral and panchromatic images fusion," *Remote Sensing*, vol. 10, no. 3, p. 373, 2018.

[30] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Information Fusion*, vol. 69, pp. 40–51, 2021.

[31] Y.-Z. Feng and D.-W. Sun, "Application of hyperspectral imaging in food safety inspection and control: a review," *Critical reviews in food science and nutrition*, vol. 52, no. 11, pp. 1039–1058, 2012.

[32] J. Qin, M. S. Kim, K. Chao, D. E. Chan, S. R. Delwiche, and B.-K. Cho, "Line-scan hyperspectral imaging techniques for food safety and quality applications," *Applied Sciences*, vol. 7, no. 2, p. 125, 2017.

[33] G. ElMasry, P. Gou, and S. Al-Rejaie, "Effectiveness of specularity removal from hyperspectral images on the quality of spectral signatures of food products," *Journal of Food Engineering*, vol. 289, p. 110148, 2021.

[34] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, 2005.

[35] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3639–3655, 2017.

[36] M. P. Uddin, M. A. Mamun, M. I. Afjal, and M. A. Hossain, "Information-theoretic feature selection with segmentation-based folded principal component analysis (pca) for hyperspectral image classification," *International Journal of Remote Sensing*, vol. 42, no. 1, pp. 286–321, 2021.

[37] J. Liang, J. Zhou, X. Bai, and Y. Qian, "Salient object detection in hyperspectral imagery," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 2393–2397.

[38] X. Kang, P. Duan, X. Xiang, S. Li, and J. A. Benediktsson, "Detection and correction of mislabeled training samples for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5673–5686, 2018.

[39] Y. Tian, Z. Li, Y. Lin, L. Xiang, X. Li, Y. Shao, and J. Tian, "Metal object detection for electric vehicle inductive power transfer systems based on hyperspectral imaging," *Measurement*, vol. 168, p. 108493, 2021.

[40] M. A. Calin, S. V. Parasca, D. Savastru, and D. Manea, "Hyperspectral imaging in the medical field: present and future," *Applied Spectroscopy Reviews*, vol. 49, no. 6, pp. 435–447, 2014.

[41] D. Zhang, J. Zhang, Z. Wang, and M. Sun, "Tongue colour and coating prediction in traditional chinese medicine based on visible hyperspectral imaging," *IET Image Processing*, vol. 13, no. 12, pp. 2265–2270, 2019.

[42] V. Dremin, Z. Marcinkevics, E. Zherebtsov, A. Popov, A. Grabovskis, H. Kronberga, K. Geldnere, A. Doronin, I. Meglinski, and A. Bykov, "Skin complications of diabetes mellitus revealed by polarized hyperspectral imaging and machine learning," *IEEE Transactions on Medical Imaging*, 2021.

[43] H. Zhang, L. Zhang, and H. Shen, "A super-resolution reconstruction algorithm for hyperspectral images," *Signal Processing*, vol. 92, no. 9, pp. 2082–2096, 2012.

[44] N. Akhtar, F. Shafait, and A. Mian, "Hierarchical beta process with gaussian process prior for hyperspectral image super resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 103–120.

[45] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3586–3594.

[46] 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*, vol. 25, no. 5, pp. 2337–2352, 2016.

[47] W. He, H. Zhang, L. Zhang, and H. Shen, "Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration," *IEEE transactions on geoscience and remote sensing*, vol. 54, no. 1, pp. 178–188, 2015.

[48] X.-H. Han, B. Shi, and Y. Zheng, "Self-similarity constrained sparse representation for hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5625–5637, 2018.

[49] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 528–537, 2011.

[50] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y.-W. Tai, and K. Ikeuchi, "High-resolution hyperspectral imaging via matrix factorization," in *CVPR 2011*. IEEE, 2011, pp. 2329–2336.

[51] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE transactions on geoscience and remote sensing*, vol. 49, no. 10, pp. 3973–3985, 2011.

[52] Y. Zhang, B. Du, L. Zhang, and S. Wang, "A low-rank and sparse matrix decomposition-based mahalanobis distance method for hyperspectral anomaly detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1376–1389, 2015.

[53] W. Wei, L. Zhang, Y. Jiao, C. Tian, C. Wang, and Y. Zhang, "Intracluster structured low-rank matrix analysis method for hyperspectral denoising," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 866–880, 2018.

[54] S. Mei, J. Hou, J. Chen, L.-P. Chau, and Q. Du, "Simultaneous spatial and spectral low-rank representation of hyperspectral images for classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2872–2886, 2018.

[55] X.-H. Han, Y. Zheng, and Y.-W. Chen, "Multi-level and multi-scale spatial and spectral fusion cnn for hyperspectral image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[56] X.-H. Han, B. Shi, and Y. Zheng, "Ssf-cnn: Spatial and spectral fusion with cnn for hyperspectral image super-resolution," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2506–2510.

[57] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5345–5355, 2018.

[58] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[59] O. Sidorov and J. Yngve Hardeberg, "Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[60] R. Imamura, T. Itasaka, and M. Okuda, "Self-supervised hyperspectral image restoration using separable image prior," *arXiv preprint arXiv:1907.00651*, 2019.

[61] Z. Chen, H. Pu, B. Wang, and G.-M. Jiang, "Fusion of hyperspectral and multispectral images: A novel framework based on generalization of pansharpening methods," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 8, pp. 1418–1422, 2014.

[62] G. A. Licciardi, M. M. Khan, J. Chanussot, A. Montanvert, L. Condat, and C. Jutten, "Fusion of hyperspectral and panchromatic images using multiresolution analysis and nonlinear pca band reduction," *EURASIP Journal on Advances in Signal processing*, vol. 2012, no. 1, pp. 1–17, 2012.

[63] G. Vivone, R. Restaino, G. Licciardi, M. Dalla Mura, and J. Chanussot, "Multiresolution analysis and component substitution techniques for hyperspectral pansharpening," in *2014 IEEE Geoscience and Remote Sensing Symposium*. IEEE, 2014, pp. 2649–2652.

[64] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," Jan. 4 2000, uS Patent 6,011,875.

[65] X.-H. Han, J. Wang, B. Shi, Y. Zheng, and Y.-W. Chen, "Hyper-spectral image super-resolution using non-negative spectral representation with data-guided sparsity," in *2017 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2017, pp. 500–506.

[66] E. Wycoff, T.-H. Chan, K. Jia, W.-K. Ma, and Y. Ma, "A non-negative sparse promoting algorithm for high resolution hyperspectral imaging," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 1409–1413.

[67] Y. Chen, W. He, N. Yokoya, and T.-Z. Huang, "Hyperspectral image restoration using weighted group sparsity-regularized low-rank tensor decomposition," *IEEE transactions on cybernetics*, 2019.

[68] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 63–78.

[69] X. X. Zhu, C. Grohnfeldt, and R. Bamler, "Exploiting joint sparsity for pan-sharpening: The j-sparsefi algorithm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 2664–2681, 2015.

[70] K. Wei and Y. Fu, "Low-rank bayesian tensor factorization for hyperspectral image denoising," *Neurocomputing*, vol. 331, pp. 412–423, 2019.

[71] X.-H. Han, B. Shi, and Y. Zheng, "Residual hsrcnn: Residual hyper-spectral reconstruction cnn from an rgb image," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2664–2669.

[72] X.-H. Han and Y.-W. Chen, "Deep residual network of spectral and spatial fusion for hyperspectral image super-resolution," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2019, pp. 266–270.

[73] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Model-based fusion of multi-and hyperspectral images using pca and wavelets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2652–2663, 2014.

[74] Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, and P. K. Varshney, "Fusionnet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 7565–7577, 2020.

[75] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by ms/hs fusion net," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1585–1594.

[76] Z. Zhu, J. Hou, J. Chen, H. Zeng, and J. Zhou, "Residual component estimating cnn for image super-resolution," vol. 30, 2020, pp. 1423–1428.

[77] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse dirichlet-net for hyperspectral image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2511–2520.

[78] L. Zhang, J. Nie, W. Wei, Y. Zhng, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," *CVPR*, 2020.

[79] J. Nie, L. Zhang, W. Wei, Z. Lang, and Y. Zhang, "Unsupervised alternating optimization for blind hyperspectral imagery super-resolution," *arXiv preprint arXiv:2012.01745*, 2020.

[80] Z. Liu, Y. Zheng, and X.-H. Han, "Unsupervised multispectral and hyperspectral image fusion with deep spatial and spectral priors," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[81] T. Uezato, D. Hong, N. Yokoya, and W. He, "Guided deep decoder: Unsupervised image pair fusion," in *European Conference on Computer Vision*. Springer, 2020, pp. 87–102.

[82] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3073–3082.

[83] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Hyperspectral image super-resolution with optimized rgb guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 661–11 670.

[84] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. R. Bradski, "Self-supervised monocular road detection in desert terrain." in *Robotics: science and systems*, vol. 38. Philadelphia, 2006.

[85] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[86] K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M. A. van Gerven, "Generative adversarial networks for reconstructing natural images from brain activity," *NeuroImage*, vol. 181, pp. 775–785, 2018.

[87] S. H. Bach, B. He, A. Ratner, and C. Ré, "Learning the structure of generative models without labeled data," in *International Conference on Machine Learning*. PMLR, 2017, pp. 273–282.

[88] Q. Huang, W. Li, T. Hu, and R. Tao, "Hyperspectral image super-resolution using generative adversarial network and residual learning," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3012–3016.

[89] Z. He, H. Liu, Y. Wang, and J. Hu, "Generative adversarial networks-based semi-supervised learning for hyperspectral image classification," *Remote Sensing*, vol. 9, no. 10, p. 1042, 2017.

[90] C. Zou and X. Huang, "Hyperspectral image super-resolution combining with deep learning and spectral unmixing," *Signal Processing: Image Communication*, p. 115833, 2020.

[91] 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*, vol. 19, no. 9, pp. 2241–2253, 2010.

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

[93] N. Yokoya, X. X. Zhu, and A. Plaza, "Multisensor coupled spectral unmixing for time-series analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2842–2857, 2017.

[94] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[95] J. Wang, S. Kwon, and B. Shim, "Generalized orthogonal matching pursuit," *IEEE Transactions on signal processing*, vol. 60, no. 12, pp. 6202–6216, 2012.

[96] M. Şımşek and E. Polat, "The effect of dictionary learning algorithms on super-resolution hyperspectral reconstruction," in *2015 XXV International Conference on Information, Communication and Automation Technologies (ICAT)*. IEEE, 2015, pp. 1–5.

[97] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[98] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3631–3640.

[99] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive cnn-based pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, pp. 5443–5457, 2018.

[100] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 639–643, 2017.

[101] O. Sidorov and J. Y. Hardeberg, "Deep hyperspectral prior: Denoising, inpainting, super-resolution," *CoRR*, vol. abs/1902.00301, 2019. [Online]. Available: http://arxiv.org/abs/1902.00301

[102] X.-H. Han, Y. Sun, and Y.-W. Chen, "Residual component estimating cnn for image super-resolution," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2019, pp. 443–447.

[103] B. Liu, X. Yu, A. Yu, P. Zhang, G. Wan, and R. Wang, "Deep few-shot learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 2290–2304, 2018.

[104] R. Imamura, T. Itasaka, and M. Okuda, "Zero-shot hyperspectral image denoising with separable image prior," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[105] Y. Li, J. Hu, X. Zhao, W. Xie, and J. Li, "Hyperspectral image super-resolution using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 29–41, 2017.

[106] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by cnn denoiser," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 3, pp. 1124–1135, 2020.

[107] R. Dian, S. Li, and L. Fang, “Learning a low tensor-train rank representation for hyperspectral image super-resolution,” *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2672–2683, 2019.

[108] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, “Model-guided deep hyperspectral image super-resolution,” *IEEE Transactions on Image Processing*, vol. 30, pp. 5754–5768, 2021.

[109] Z. Liu, Y. Zheng, and X.-H. Han, “Deep unsupervised fusion learning for hyperspectral image super resolution,” *Sensors*, vol. 21, no. 7, p. 2348, 2021.

[110] Y. Qu, H. Qi, C. Kwan, N. Yokoya, and J. Chanussot, “Unsupervised and unregistered hyperspectral image super-resolution with mutual dirichlet-net,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[111] A. Shocher, N. Cohen, and M. Irani, ““zero-shot” super-resolution using deep internal learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3118–3126.