ADASR:An Adversarial Auto-Augmentation Framework for Hyperspectral and Multispectral Data Fusion

Jinghui Qin†, Lihuang Fang†, Ruitao Lu, Liang Lin, and Yukai Shi*

Abstract—Deep learning-based hyperspectral image (HSI) super-resolution, which aims to generate high spatial resolution HSI (HR-HSI) by fusing hyperspectral image (HSI) and multispectral image (MSI) with deep neural networks (DNNs), has attracted lots of attention. However, neural networks require large amounts of training data, hindering their application in real-world scenarios. In this letter, we propose a novel adversarial automatic data augmentation framework ADASR that automatically optimizes and augments HSI-MSI sample pairs to enrich data diversity for HSI-MSI fusion. Our framework is sample-aware and optimizes an augmentor network and two downsampling networks jointly by adversarial learning so that we can learn more robust downsampling networks for training the upsampling network. Extensive experiments on two public classical hyperspectral datasets demonstrate the effectiveness of our ADASR compared to the state-of-the-art methods.

Index Terms—Adversarial training, data augmentation, hyperspectral, multispectral, deep learning

I. Introduction

RECENTLY, there has been a growing interest in developing deep neural networks [1], [2], [3], [4], [5], [6] for hyperspectral image (HSI) super-resolution which is a task of producing HSIs from contiguous spectral information in narrow spectral bands. The HSI can be expressed as 3D tensors with 2 spatial dimensions and 1 spectral dimension [7]. Training a neural network robustly often relies on massive and diverse data. However, unlike other image super-resolution tasks with much more synthetic or real training samples, hyperspectral image data is scarce, and the spectral dimension of hyperspectral image data is very high. Therefore, it is non-trivial to train a stable and effective deep neural network.

Nowadays, data augmentation (DA) is an efficient strategy to lift up the model generalization performance by artificially increasing the volume and diversity of the training data. Conventional DA strategies, such as image rotation [8], image flip [9], etc., often rotate the input image randomly in a pre-defined augmentation angle. Despite its effectiveness on image classification and image super-resolution tasks, this conventional DA approach may lead to insufficient training due to the following reasons: 1) the network training and DA are regarded as two independent phases without joint optimization; 2) the same fixed image rotation augmentation is applied to all input samples without considering the complexity of the samples. Different samples need different rotation angles. Hence, it is insufficient to apply conventional DA to augment the training samples [10].

To improve the effect of hyperspectral and multispectral image fusion by augmenting the input samples and training a more stable network, we propose a novel adversarial automatic augmentation framework that jointly optimizes an augmentor network and two downsampling networks, such that the augmentor can learn to produce augmented samples by rotating them at appropriate angles driven by their content to make the two downsampling networks more stable for training upsample network at the next stage. Specifically, in the first stage, the augmentor network learns data variations to enrich the input samples by using the loss from a spatial downsampling network and a spectral downsampling network as the feedback. Meanwhile, these downsampling networks take charge of learning a degradation procession on ensuring the generated augmented samples to be semantic-consistent with low-resolution multispectral images so that these downsampling networks can generate appropriate and valuable feedback to optimize the augmentor network. The augmentor network and the downsampling networks are trained in the adversarial learning setting. In the second stage, we train a spectral upsampling network by using the low-spatial-resolution multispectral images generated by the spatial downsampling network and the high spatial resolution multispectral images with reconstruction loss and consistency loss so that we can take full advantage of the priors learned by downsampling networks. The experimental results on two public classical hyperspectral benchamrks demonstrate the effectiveness of our method compared to the state-of-the-art HSI-MSI fusion methods.

II. Methodology

The main contribution of this work is the adversarial auto-augmentation framework that automatically optimizes the augmentation of the input samples for more effective

J. Qin, L. Fang, and Y. Shi are with the School of Information Engineering, Guangdong University of Technology, Guangzhou, 510006, China (email: qinjinghui@gdut.edu.cn; 3120002329@mail2.gdut.edu.cn; ykshi@gdut.edu.cn)

R. Lu is with the Department of Missile Engineering, Rocket Force University of Engineering, Xi'an 710025, China (email: lrt19880220@163.com)

L. Lin is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, 510006, China (email: linliang@ieee.org)

† The first two authors share equal contributions.

* Corresponding author: Yukai Shi

Fig. 1. Overview. (a) The adversarial auto-augmentation framework: We jointly optimize a data augmentor G and two downsampling networks, a speDnet D_y and a spaDnet D_z at the first stage. The augmentor G learns to generate a sample-specific augmentation function that takes into account the transformation of the image rotation angle, increasing the data diversity for training a better downsampling network. In the second stage, we optimize the spectral upsampling network (SpeUnet) U with the help of the low-resolution multispectral images generated by the fixed downsampling network D_z. (b) The design of our data augmentor G. The data augmentor G learns to predict the angle to augment the input sample for the degradation model learning.

training of the spatial downsampling network. As shown in Fig. 1, our adversarial auto-augmentation framework consists of four modules, including an adversarial data augmentor G, a spatial downsampling network (SpaDnet) D_y, a spectral downsampling network (SpeDnet) D_z, and a spectral upsampling network (SpeUnet) U. These modules will be optimized in two training stages. At the first stage, G, D_y, and D_z are optimized jointly in the adversarial training setting. The augmentor G takes charge of learning a sample-specific rotation angle transformation to increase the data diversity for optimizing downsampling networks D_y and D_z so that we can generate high-quality low-resolution multispectral images for upsampling. Meanwhile, the losses generated by the two downsampling networks will be as feedback to optimize the augmentor G. In the second stage, we optimize the speUnet U with the help of the low-resolution multispectral images generated by the fixed spatial downsampling network D_z and the high spatial resolution multispectral images by reconstruction loss and consistency loss.

A. Downsampling Networks

Let the low-spatial-resolution HSI $\mathbf{Y} \in \mathbb{R}^{wh \times C}$ and the low-spectral-resolution MSI $\mathbf{Z} \in \mathbb{R}^{WH \times C_m}$ be the spatially degraded version and spectrally degraded version of ground-truth HR-HSI $\mathbf{X} \in \mathbb{R}^{WH \times C}$. Here, W, H, and C are the width, height and spectral bands of $\mathbf{X}$ while w and h are the width and height of $\mathbf{Y}$, and C_m is the spectral band of $\mathbf{Z}$ ($w \ll W, h \ll H$, $C_m \ll C$), respectively. Then, the degradation models HSI and MSI can be modeled as follows:

$$\mathbf{Y} = \mathbf{PX}, \qquad (1)$$
$$\mathbf{Z} = \mathbf{XS}, \qquad (2)$$

where $\mathbf{P} \in \mathbb{R}^{wh \times WH}$ denote the spatial degradation pipeline, which consists of a convolution operation using Point Spread Function (PSF) and a spatial downsample operation, and $\mathbf{S} \in \mathbb{R}^{C \times C_m}$ is a band-level spectral response in MSI $\mathbf{Z}$. With the

$\mathbf{Y}$ and $\mathbf{Z}$, the HSI-MSI fusion task targets at reconstructing the latent $\mathbf{X}$. In addition, the HSI $\mathbf{Y}$'s spectrally degraded result $\mathbf{M_Y}$ should be equal to the MSI $\mathbf{Z}$'s spatially degraded result $\mathbf{M_Z}$:

$$\mathbf{M_Y} = \mathbf{YS} = \mathbf{PZ} = \mathbf{M_Z}, \qquad (3)$$

where $\mathbf{M_Y} \in \mathbb{R}^{wh \times Cm}$ and $\mathbf{M_Z} \in \mathbb{R}^{wh \times Cm}$.

To model the degradation process, we follow prior work [3] to design the spectral downsampling network (SpeDnet) D_y with one convolutional layer with the kernels shape $N_k \times C_{i,in} \times C_{i,o} \times 1 \times 1$ and stride size 1 to model the integral procedure with Spectral Response Function (SRF), where N_k denotes both the number of convolutional kernels in SpeDnet and the band number of MSI $\mathbf{Z}$. i is the index of the kernels. $C_{i,in}$ is determined by the number of hyperspectral bands that are covered by each band's spectral response in MSI $\mathbf{Z}$. $C_{i,o}$ is constrained to 1. That is, each kernel generates only one feature map. 1×1 is each kernel's spatial size. Therefore, the SpeDnet can be modeled as follows:

$$\mathbf{M}_{\mathbf{Y}(i,j)} = \mathrm{SpeDnet}(\mathbf{Y}, \theta) = \frac{\sum_{t \in \Theta_j} \mathbf{Y}_{i,t}\omega_j}{\sum \omega_j}, \qquad (4)$$

where θ represents the weights of the SpeDnet, i and j represent the index of row and column, respectively. Θ_j represents the j-th support set that the band of $\mathbf{Y}$ appertains, and ω_j represents the weights of the j-th $C_{in} \times 1 \times 1$ convolution kernel.

Similarly, the spatial downsampling network (SpaDnet) D_z is designed to act as PSF. Each band in the spatial dimension is convolved with the same convolutional kernel of size $1 \times r \times r$ and stride r, where r ($r = W/w = H/h$) is the spatially dimensional scale factor and the size of convolutional kernels. The SpaDnet can be modeled as follows:

$$\mathbf{M_Z} = \mathrm{SpaDnet}(\mathbf{Z}, \beta), \qquad (5)$$

where β is the weight of SpaDnet.

B. Sample Augmentor

To train the downsampling networks more effectively, the augmentor G learns to generate a sample-specific image rotation angle function for augmenting each input sample. Our augmentor G takes HSI $\mathbf{Y}$ and MSI $\mathbf{Z}$ as input and output the augmented samples $\mathbf{Y}_G$ and $\mathbf{Z}_G$ respectively. The overall architecture of our augmentation procedure is shown in Figure 1 (b). First, a pixel-wise feature extraction unit is deployed to extract features $F \in \mathbb{R}^{\hat{w}\hat{h} \times \hat{C}}$, where $\hat{h}$, $\hat{w}$, and $\hat{C}$ denote the height of the feature map, the width of the feature map, and the number of feature channels, respectively. Then, the adaptive average pooling operation is applied to build a one-pixel and multi-channels feature map $F^{'}$. Furthermore, a multilayer perceptron takes the $F^{'}$ as input to generate a suitable rotated angle. Finally, the augmentor G generates augmented samples with the affine transform and the generated rotation angle. Meanwhile, we also use the same rotation angle and affine transform on the ground-truth spatially degraded version $\mathbf{M}$, which is given in the training set, so that we can train the downsampling networks and augmentor with adversarial learning strategy. We label the augmented $\mathbf{M}$ as $\mathbf{M}_G$. The augmented samples $\mathbf{Y}_G$ and $\mathbf{Z}_G$ generated by our augmentor can satisfy two following requirements to maximize the network learning: (i) $\mathbf{Y}_G$ and $\mathbf{Z}_G$ can be more challenging than $\mathbf{Y}$ and $\mathbf{Z}$ for downsampling networks since they are rotated and deformed; (ii) $\mathbf{Y}_G$ and $\mathbf{Z}_G$ do not lose any semantic information in the original $\mathbf{Y}$ and $\mathbf{Z}$.

C. Adversarial Learning in First Stage

To maximize the network learning and generate more challenging samples, we train the augmentor G and the two downsampling networks D_z and D_y with adversarial learning strategy [11]. The augmentor G and the two downsampling networks D_z and D_y are trained alternately and iteratively.

To train the augmentor G, D_z and D_y are fixed. The hyperspectral HSI $\mathbf{Y}$, high-spatial-resolution MSI $\mathbf{Z}$ are fed into data augmentor G and generate new augmented images $\mathbf{Y}_G$ and $\mathbf{Z}_G$. Then the $\mathbf{Y}_G$ is fed into the SpeDnet D_y to generate low-spectral version $\mathbf{M}_{\mathbf{Y}_G}$ while the $\mathbf{Z}_G$ is fed into the SpaDnet D_z to generate low-spatial version $\mathbf{M}_{\mathbf{Z}_G}$. Finally, we constrain $\mathbf{M}_{\mathbf{Y}_G}$, $\mathbf{M}_{\mathbf{Z}_G}$, and $\mathbf{M}_G$ to be consistent by minimizing the following loss:

$$\mathcal{L}_G = \rho \mathcal{L}\left(\mathbf{M}_{\mathbf{Y}_G}\right) + \rho \mathcal{L}\left(\mathbf{M}_{\mathbf{Z}_G}\right), \quad (6)$$

where

$$\mathcal{L}\left(\mathbf{M}_{\mathbf{Y}_G}\right) = log\left(\frac{1}{whC_m} \left\|\mathbf{M}_{\mathbf{Y}_G} - \mathbf{M}_G\right\|_1\right), \quad (7)$$

$$\mathcal{L}\left(\mathbf{M}_{\mathbf{Z}_G}\right) = log\left(\frac{1}{whC_m} \left\|\mathbf{M}_{\mathbf{Z}_G} - \mathbf{M}_G\right\|_1\right), \quad (8)$$

ρ is an adjustable hyperparameter.

Similarly, to train the D_z and D_y, we fix the augmentor G. Then, the original sample $\mathbf{Y}$ and its augmented sample $\mathbf{Y}_G$ are fed into SpeDnet D_y to obtain $\mathbf{M}_{\mathbf{Y}}$ and $\mathbf{M}_{\mathbf{Y}_G}$ while the original sample $\mathbf{Z}$ and its augmented sample $\mathbf{Z}_G$ are fed into

SpaDnet D_z to obtain $\mathbf{M}_{\mathbf{Z}}$ and $\mathbf{M}_{\mathbf{Z}_G}$. Finally, we adopt L1 loss to optimize these two downsampling networks as follows:

$$\mathcal{L}_D = \left\|\mathbf{M}_{\mathbf{Y}} - \mathbf{M}\right\|_1 + \left\|\mathbf{M}_{\mathbf{Y}_G} - \mathbf{M}_G\right\|_1 \\ + \left\|\mathbf{M}_{\mathbf{Z}} - \mathbf{M}\right\|_1 + \left\|\mathbf{M}_{\mathbf{Z}_G} - \mathbf{M}_G\right\|_1, \quad (9)$$

D. Spectral Upsample Network in Second Stage

The low-spatial-resolution version $\mathbf{M}_{\mathbf{Z}}$ can be produced by applying the spatial degradation pipeline $\mathbf{P}$ in Eq (3) to the MSI $\mathbf{Z}$, where $\mathbf{M}_{\mathbf{Z}} \in \mathbb{R}^{wh \times C_m}$. From Eq (2) and Eq (3), $\mathbf{M}_{\mathbf{Z}}$ and $\mathbf{Z}$ are generated by applying the same spectral degradation operation $\mathbf{S}$ to $\mathbf{Y}$ and $\mathbf{X}$, respectively. The latent HR-HSI $\mathbf{X}$ can be reconstructed if we apply the spectral inverse mapping from ow-spatial-resolution version $\mathbf{M}_{\mathbf{Z}}$ to hyperspectral HSI $\mathbf{Y}$, which is learned in the low resolution, to high-spatial-resolution MSI $\mathbf{Z}$. Therefore, we use the $\mathbf{M}_{\mathbf{Z}}$ to learn the inverse mapping of the spectrum from $\mathbf{M}_{\mathbf{Z}}$ to $\mathbf{Y}$ by a SpeUnet. It can be modeled as follows:

$$\hat{\mathbf{Y}}_{\mathbf{M}_{\mathbf{Z}}} = SpeUnet\left(\mathbf{M}_{\mathbf{Z}}\right), \quad (10)$$

where the SpeUnet contains 1×1 convolution kernels. To optimize the $SpeUnet$, we apply L1 loss to learn spectral inverse mapping as follows:

$$\mathcal{L}_{U_1} = \frac{1}{whC} \left\|\mathbf{Y} - \hat{\mathbf{Y}}_{\mathbf{M}_{\mathbf{Z}}}\right\|_1, \quad (11)$$

With an appropriate optimization, HR-HSI $\hat{\mathbf{X}}$ can be reconstructed coarsely by inputting $\mathbf{Z}$ into $SpeUnet$. However, there exist limitations to the performance of the learned low-resolution spectral inverse mapping. For reconstructing fine-grain HR-HSI $\hat{\mathbf{X}}$, we also use the paired $\mathbf{Z}$ and $\mathbf{X}$ as training data to train the $SpeUnet$ while introducing a consistency loss to optimize $SpeUnet$ for improving its reconstruction performance on high resolution. The consistency loss can be modeled as follows:

$$\hat{\mathbf{Z}}_{\mathbf{X}_{\mathbf{Z}}} = SpeDnet\left(SpeUnet\left(\mathbf{Z}\right)\right), \quad (12)$$

$$\mathcal{L}_{U_2} = \frac{1}{WHC_m} \left\|\mathbf{Z} - \hat{\mathbf{Z}}_{\mathbf{X}_{\mathbf{Z}}}\right\|_1, \quad (13)$$

Finally, $SpeUnet$ is optimized as follows:

$$\mathcal{L}_U = \mathcal{L}_{U_1} + \alpha \mathcal{L}_{U_2}, \quad (14)$$

where α is an adjustable hyperparameter.

To obtain the final HR-HSI $\hat{\mathbf{X}}$, we applied the learned SpeUnet to the original MSI $\mathbf{Z}$ as follows:

$$\hat{\mathbf{X}} = SpeUnet\left(\mathbf{Z}\right). \quad (15)$$

III. EXPERIMENTS

A. Datasets, Baselines, and Metrics

Datasets. To evaluate the efficiency of our design strategy, we adopted two widely used hyperspectral datasets. The first one called Houston18 for short has 48 bands with wavelengths ranging from 380 to 1050 nm. This dataset contains 1202×4172 pixels with a spatial resolution of 1 m. The second one, named Chikusei for short, has 128 bands with wavelengths ranging from 363 to 1018 nm and contains

TABLE I
QUANTITATIVE PERFORMANCE OF VARIOUS METHODS ON CHIKUSEI AND HOUSTON18 DATASETS

| Methods | × 5 | | | | | | | | | | × 8 | | | | | | | | | |
| | Chikusei | | | | | houston18 | | | | | Chikusei | | | | | houston18 | | | | |
	SAM↓	ERGAS↓	PSNR↑	RMSE↓	CC↑	SAM↓	ERGAS↓	PSNR↑	RMSE↓	CC↑	SAM↓	ERGAS↓	PSNR↑	RMSE↓	CC↑	SAM↓	ERGAS↓	PSNR↑	RMSE↓	CC↑
HySure	1.1953	1.0697	42.4036	0.0081	0.9974	1.5485	0.6756	43.0993	0.0052	0.9989	1.5365	0.8894	40.0119	0.0107	0.9968	2.1120	0.5719	40.2963	0.0070	0.9984
FUSE	1.3413	1.2055	41.1566	0.0094	0.9967	1.6768	0.9117	40.5935	0.0072	0.9979	1.4428	0.8323	40.5532	0.0098	0.9972	1.8498	0.5673	40.7906	0.0066	0.9980
G-SOMP+	1.2874	1.3306	41.2452	0.0091	0.9945	1.4909	0.6880	42.8371	0.0054	0.9989	1.5247	1.0934	38.8789	0.0117	0.9914	1.8978	0.5429	40.5534	0.0068	0.9984
CSU	1.4397	1.7031	39.8464	0.0097	0.9898	1.4980	0.6912	42.5571	0.0054	0.9987	1.6817	1.1877	38.3814	0.0117	0.9879	1.9006	0.5476	40.3440	0.0068	0.9981
CNMF	1.0918	1.0752	42.6555	0.0079	0.9964	1.2197	0.6054	43.9170	0.0047	9.9991	1.2458	0.9126	39.8480	0.0105	0.9951	1.5856	0.4972	41.3301	0.0063	0.9986
STEREO	0.8801	0.8282	49.7968	0.0043	0.9968	1.0094	0.3691	51.0273	0.0028	0.9994	1.0282	0.5957	48.8520	0.0050	0.9958	1.1090	0.2525	50.4403	0.0030	0.9992
CSTF	1.2306	1.3534	45.1193	0.0059	0.9932	1.4285	0.5509	46.9136	0.0037	0.9988	1.2458	0.8431	45.1089	0.0060	0.9930	1.4539	0.3513	46.8232	0.0038	0.9988
DHIF-Net	1.4113	1.6151	39.6355	0.0096	0.9904	1.4108	0.6578	42.8076	0.0052	0.9987	1.7132	1.3255	37.7378	0.0118	0.9841	1.7075	0.5127	40.8392	0.0064	0.9979
CUCaNet	1.0353	0.8356	48.4793	0.0054	0.9973	1.7031	0.7984	42.2861	0.0066	0.9986	0.8561	0.4843	49.6982	0.0044	0.9974	1.7450	0.4818	43.2826	0.0064	0.9987
UDALN	0.7127	0.6851	52.3858	0.0037	0.9980	0.8770	0.6698	42.6841	0.0053	0.9994	0.7504	0.4574	49.2979	0.0045	0.9976	0.8896	0.5452	40.2321	0.0070	0.9994
ADASR(OUR)	0.7032	0.5742	53.6463	0.0035	0.9983	0.8234	0.3132	53.5342	0.0024	0.9995	0.7395	0.4347	52.0179	0.0038	0.9977	0.8438	0.2007	53.1627	0.0025	0.9995

2517×2335 pixels with a spatial resolution of 2.5 m. After removing some noisy and water absorption bands, the sub-images of $240 \times 240 \times 46$ on Houston18 and $240 \times 240 \times 110$ on Chikusei are chosen for the test. The two sub-images are acted as reference images for comparison. In synthesizing the RGB images from the Houston18 and Chikusei datasets, we used 46th and 110th spectral bands as red-band, 30th and 75th spectral bands as green-band, and 14th and 30th spectral bands as blue-band.

Baselines. To demonstrate the effectiveness of our ADASR, we compare our ADASR with 10 SOTA HSI-MSI fusion methods, including 1 Bayesian representation based method (FUSE [12]), 4 matrix factorization based methods (HySure [13], CNMF [14], CSU [15] and G-SOMP+ [16]), 2 tensor factorization based methods (CSTF [17] and STEREO [1]), 1 supervised deep learning method (DHIF-Net [4]), 2 unsupervised deep learning methods (CUCaNet [2] and UDALN [3]). We abbreviate our method as ADASR in the following description.

Metrics. We deploy 5 metrics, including spectral angle mapper (SAM), error relative to the global adimensional synthesis (ERGAS), peak signal-to-noise ratio (PSNR), root mean square error (RMSE), and correlation coefficients (CC) to evaluate the performance.

B. Implementation Details

The ADASR is implemented by using PyTorch [18], and trained on the Linux server with an NVIDIA Titan XP GPU. We set the training step as 40,000 and deployed the Adam [19] optimizer with a learning rate of 0.0001. In addition, in Equation 14, we set the parameter α to 0.3. To train the downsampling networks, we take the HSI and MSI pairs as input and use trainable PSF and SRF for downsampling networks. To reduce model oscillations [11], we follow the prior work [20] and train the downsampling networks D_z and D_y by using mixed original and augmented training samples rather than using original training samples only.

C. Main Results

The quantitative results for the Housoton18 and Chikusei datasets with scale factors of 5 and 8 are shown in Table I.

Fig. 2. Visual results generated by different methods on the Housoton18 and Chikusei datasets in the scale factor 8. The 1st row shows the reconstruction results for the whole image, the 2nd row shows the MAE heatmap, and the 3rd row shows the SAM error heatmap.

From the results, we can observe that our method outperforms all baselines on both two datasets. The improvements (%) on SAM/ERGAS/PSNR/RMSE/CC metrics for Houston18 and Chikusei datasets in the scale 8 are 5.1/20.5/5.4/16.7/0.01 and 1.5/5.0/5.5/15.6/0.01, respectively. Meanwhile, the improvements (%) in the scale 5 are 6.1/15.1/4.9/14.3/0.01 and 1.3/16.2/2.4/5.4/0.03. These quantitative improvements demonstrate the superiority of our method. Besides, we also conduct the qualitative analysis by showing the synthetic RGB images of HR-HSI and the mean absolute error (MAE) heatmap and SAM heatmap between the reconstructed HR-HSI and the reference HR-HSI as in the Fig. 2. We can see that our method can achieve better SAM and MAE. The qualitative results also show the reconstruction superiority of our method to reconstruct details more efficiently.

D. Ablation Study

1) The effects of the data augmentation and the consistency loss: In our framework, the data augmentor G and the consistency loss $\mathcal{L}_{U_2}$ are two key components, so we conduct ablation studies on these two components. The experimental results are shown in Table II. We can observe that either the auto data augmentation or the consistency loss can improve the performance, while the best performance can be achieved by applying both auto data augmentation and the consistency loss.

TABLE II
ABLATION STUDY OF DIFFERENT COMPONENTS. '-' MEANS THE COMPONENT IS REMOVED

Model	Metric				
	SAM↓	ERGAS↓	PSNR↑	RMSE↓	CC↑
$-G$-$\mathcal{L}_{U_2}$	0.8655	0.5891	43.9732	0.0046	0.9995
$-G$	0.8591	0.3602	52.3446	0.0028	0.9995
$-\mathcal{L}_{U_2}$	0.8255	0.3160	53.5126	0.0024	0.9995
ADASR	0.8234	0.3132	53.5342	0.0024	0.9995

2) Can the learned augmentor G work better?: We also explore whether our learned augmentor G can work better than the conventional image augmentation method - random rotation augmentation or without any image augmentation. The results on the Houston18 dataset are shown in Table III. We can observe that the conventional image augmentation method can not improve the performance, while our method can improve the performance on all metrics. These results show the effectiveness of our adversarial auto-augmentation framework.

TABLE III
STUDY OF THE EFFECTIVENESS OF DATA AUGMENTOR

Method	Metric				
	SAM↓	ERGAS↓	PSNR↑	RMSE↓	CC↑
No augmentation	0.8591	0.3602	52.3446	0.0028	0.9995
Random rotation	0.8452	0.3351	53.0553	0.0025	0.9994
ADASR	0.8234	0.3132	53.5342	0.0024	0.9995

IV. CONCLUSION

In this letter, to improve the effect of hyperspectral and multispectral image fusion by augmenting the input samples and training a more stable network, we propose a novel adversarial automatic augmentation framework ADASR that jointly optimizes an augmentor network and two downsampling networks so that the augmentor network can augmented samples automatically by rotating them at appropriate angles driven by their content to make the two downsampling networks more stable for training upsample network at the next stage. Specifically, the augmentor network and the downsampling networks are trained by reconstructing low-spatial resolution multispectral images in the adversarial learning setting. Then, we train a spectral upsampling network by both high spatial resolution multispectral images and their generated low spatial resolution multispectral images with reconstruction loss and consistency loss, so that we can take full advantage of the priors learned by downsampling networks. The experimental results on two public classical hyperspectral datasets demonstrate the effectiveness of our ADASR compared to the state-of-the-art HSI-MSI fusion methods.

REFERENCES

[1] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Hyperspectral super-resolution: A coupled tensor factorization approach," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6503–6517, 2018.

[2] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer, 2020, pp. 208–224.

[3] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[4] T. Huang, W. Dong, J. Wu, L. Li, X. Li, and G. Shi, "Deep hyperspectral image fusion network with iterative spatio-spectral regularization," *IEEE Transactions on Computational Imaging*, vol. 8, pp. 201–214, 2022.

[5] X. Yang, M. Zhu, B. Sun, Z. Wang, and F. Nie, "Fuzzy c-multiple-means clustering for hyperspectral image," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[6] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, 07 2022.

[7] D. Hong, W. He, N. Yokoya, J. Yao, L. Gao, L. Zhang, J. Chanussot, and X. Zhu, "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 2, pp. 52–87, 2021.

[8] S. Bucci, M. R. Loghmani, and T. Tommasi, "On the effectiveness of image rotation for open set domain adaptation," in *Computer Vision– ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020, pp. 422–438.

[9] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[10] B. Graham, "Fractional max-pooling," *arXiv preprint arXiv:1412.6071*, 2014.

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2672–2680.

[12] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Fast fusion of multi-band images based on solving a sylvester equation," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4109–4121, 2015.

[13] M. Simões, 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, 2015.

[14] 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, 2012.

[15] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3586–3594.

[16] N. Akhtar, F. Shafait, and A. S. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *ECCV*, 2014.

[17] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.

[18] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[19] C. Chen, D. Carlson, Z. Gan, C. Li, and L. Carin, "Bridging the gap between stochastic gradient mcmc and stochastic optimization," in *Artificial Intelligence and Statistics*. PMLR, 2016, pp. 1051–1060.

[20] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2242–2251.