

PanNet: A deep network architecture for pan-sharpening

Junfeng Yang[†], Xueyang Fu[†], Yuwen Hu, Yue Huang, Xinghao Ding, John Paisley[†]
Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University, China
[†]Department of Electrical Engineering, Columbia University, USA

Abstract

We propose a deep network architecture for the pan-sharpening problem called PanNet. We incorporate domain-specific knowledge to design our PanNet architecture by focusing on the two aims of the pan-sharpening problem: spectral and spatial preservation. For spectral preservation, we add up-sampled multispectral images to the network output, which directly propagates the spectral information to the reconstructed image. To preserve spatial structure, we train our network parameters in the high-pass filtering domain rather than the image domain. We show that the trained network generalizes well to images from different satellites without needing retraining. Experiments show significant improvement over state-of-the-art methods visually and in terms of standard quality metrics.

1. Introduction

Multispectral images are widely used, for example in agriculture, mining and environmental monitoring applications. Due to physical constraints, satellites will often only measure a high resolution panchromatic (PAN) image (i.e., grayscale) and several low resolution multispectral (LRMS) images. The goal of pan-sharpening is to fuse this spectral and spatial information to produce a high resolution multispectral (HRMS) image of the same size as PAN.

With recent advances made by deep neural networks for image processing applications, researchers have begun exploring this avenue for pan-sharpening. For example, one deep pan-sharpening model assumes that the relationship between HR/LR multispectral image patches is the same

Figure 1. Examples from 425 satellite images used in experiments.

as that between the corresponding HR/LR panchromatic image patches, and uses this assumption to learn a mapping through a neural network [16]. The state-of-the-art pan-sharpening model, based on the convolutional neural network and called PNN [21], adopts an architecture previously proposed for image super-resolution [11].

These two methods regard the pan-sharpening problem as a simple image regression problem. That is, though they are able to obtain good results, they do not exploit the particular goals of the pan-sharpening problem—spectral and spatial preservation—but rather treat pan-sharpening as a black-box deep learning problem. However, for pan-sharpening it is clear that preserving spatial and spectral information are the primary goals of fusion, and so deep learning methods should explicitly focus on these aspects. This motivates our proposed deep network called “PanNet,” which has the following features:

1. We incorporate problem-specific knowledge about pan-sharpening into the deep learning framework. Specifically, we propagate spectral information through the network using up-sampled multispectral images, a procedure we’ll refer to as “spectra-mapping.” To focus on the spatial structure in the PAN image, we train the network in the high-pass domain rather than the image domain.

2. Our approach is an end-to-end system which automatically learns the mapping purely from the data. Convolutions allow us to capture intra-correlation across different bands of the MS images and the PAN image, unlike previous (non-deep) methods. Experiments show that PanNet achieves state-of-the-art performance compared with sev-

[†]co-first authors contributed equally, correspondence: dxh@xmu.edu.cn. This work was supported in part by the National Natural Science Foundation of China grants 61571382, 81671766, 61571005, 81671674, U1605252, 61671309 and 81301278, Guangdong Natural Science Foundation grant 2015A030313007, Fundamental Research Funds for the Central Universities grants 20720160075 and 20720150169, the CCF-Tencent research fund, and the Science and Technology funds from the Fujian Provincial Administration of Surveying, Mapping and Geoinformation. Xueyang Fu conducted portions of this work at Columbia University under China Scholarship Council grant No. [2016]3100.

Figure 2. The deep neural network structure of the proposed pan-sharpening framework called PanNet.

eral standard approaches, as well as other deep models.

3. Most conventional methods require parameter tuning for different satellites because the range of imaging values are inconsistent. However, training in the high-pass domain removes this factor, allowing for training on one satellite to generalize well to new satellites. This is not a feature of other deep approaches, which train on the image domain.

1.1. Related work

Various pan-sharpening methods have emerged in recent decades. Among these, the most popular are based on component substitution, including the intensity hue-saturation technique (IHS) [5], principal component analysis (PCA) [20] and the Brovey transform [14]. These methods are straightforward and fast, but they tend to succeed in approximating the spatial resolution of the HRMS image contained in PAN at the expense of introducing spectral distortions. To fix this problem, more complex techniques have been proposed, such as adaptive approaches (e.g., PRACS [8]) and band-dependent approaches (e.g., BDSF [13]). In multi-resolution approaches [19, 22], the PAN image and LRMS images are decomposed, e.g. using wavelets or Laplacian pyramids, and then fused. Other model-based methods encode beliefs about the relationships between PAN, HRMS and LRMS images in a regularized objective function, and then treat the fusion problem as an image restoration optimization problem [3, 4, 7, 9, 12, 18]. Many of these algorithms obtain excellent results. We choose the best among these methods for comparison in our experiments.

2. PanNet: A deep network for pan-sharpening

Figure 2 shows a high-level outline of our proposed deep learning approach to pan-sharpening called PanNet. We motivate this structure by first reviewing common approaches to the pan-sharpening problem, and then discuss our approach in the context of the two goals of pan-sharpening, which is to reconstruct high-resolution multi-spectral images that contain the spatial content of PAN and the spectral content of the low-resolution images.

2.1. Background and motivation

We denote the set of desired HRMS images as \mathbf{X} and let \mathbf{X}_b be the image of the b th band. For the observed data, \mathbf{P} denotes the PAN image and \mathbf{M} denotes the LRMS images, with \mathbf{M}_b the b th band. Most state-of-the-art methods treat fusion as minimizing an objective of the form

$$\mathcal{L} = \alpha_1 f_1(\mathbf{X}, \mathbf{P}) + \alpha_2 f_2(\mathbf{X}, \mathbf{M}) + f_3(\mathbf{X}), \quad (1)$$

where the term $f_1(\mathbf{X}, \mathbf{P})$ enforces structural consistency, $f_2(\mathbf{X}, \mathbf{M})$ enforces spectral consistency, and $f_3(\mathbf{X})$ imposes desired image constraints on \mathbf{X} . For example the first variational method P+XS [4] lets

$$f_1(\mathbf{X}, \mathbf{P}) = \sum_{b=1}^B \|\mathbf{X}_b - \mathbf{P}\|_2^2 \quad (2)$$

with α a B -dimensional probability weight vector. Other approaches use a spatial difference operator G to focus on high-frequency content, for example $\sum_{b=1}^B \|G(\mathbf{X}_b - \mathbf{P})\|_2^2$ or $\sum_{b=1}^B \|G(\mathbf{X}_b - \mathbf{P})\|_2^2$ [3, 9, 12]. A similar structural penalty uses a hyper-Laplacian $\sum_{b=1}^B \|\mathbf{X}_b - \mathbf{P}\|_{1/2}$ [18].

For spectral consistency, many methods define

$$f_2(\mathbf{X}, \mathbf{M}) = \sum_{b=1}^B \|\mathbf{K} \mathbf{X}_b - \mathbf{M}_b\|_2^2, \quad (3)$$

where \mathbf{K} indicates upsampling \mathbf{M}_b to be the same size as \mathbf{X}_b , which is smoothed by convolving with smoothing kernel \mathbf{k} [3, 4, 12, 18]. The term $f_3(\mathbf{X})$ is often total variation penalization.

A straightforward deep learning approach to the pan-sharpening problem can leverage a plain network architecture to learn a nonlinear mapping relationship between the inputs (\mathbf{P}, \mathbf{M}) and the outputs \mathbf{X} that minimizes

$$\mathcal{L} = \|\mathbf{f}_{\mathbf{W}}(\mathbf{P}, \mathbf{M}) - \mathbf{X}\|_F^2. \quad (4)$$

Here, $\mathbf{f}_{\mathbf{W}}$ represents a neural network and \mathbf{W} its parameters. This idea is used by PNN [21], which directly inputs (\mathbf{P}, \mathbf{M}) into a deep convolutional neural network to approximate \mathbf{X} . Although this direct architecture gives excellent results, it does not exploit known image characteristics to define the inputs or network structure.

Figure 3. Example of the three model structures we considered for pan-sharpening: (left-to-right) ResNet [15], ResNet+spectra-mapping, and the final proposed network, called PanNet. ResNet has been shown to improve CNN performance on image processing tasks, but has drawbacks in the pan-sharpening framework. The second network captures the goal of spectral preservation, while the final proposed network captures both spatial and spectral information. We experiment with all three, none of which have been applied to pan-sharpening.

2.2. PanNet architecture

We are motivated to build on the success of PNN in defining PanNet. As with PNN, we also use a convolution neural network (CNN), but our specific structure differs from PNN in using the recently proposed ResNet structure [15] as our neural network. Convolutional filters are particularly useful for this problem, since they can exploit the high correlation across different bands of the multispectral images, something shown to be useful by the SIRF algorithm [7]. As with other pan-sharpening approaches, our deep network aims to preserve both spectral and spatial information. We discuss these separately below.

The high-level idea is represented in the sequence of potential network structures shown in Figure 3. The first is vanilla ResNet, while the second network only focuses on spectral information. We propose the third network called PanNet, which performs the best. We experiment with all three, none of which have been applied to the pan-sharpening problem.

2.2.1 Spectral preservation

To fuse spectral information, we up-sample \mathbf{M} and add a skip connection to the deep network of the form

$$\mathcal{L} = f_W(\mathbf{P}, \mathbf{M}) + \|\mathbf{M} - \mathbf{X}\|_F^2. \quad (5)$$

\mathbf{M} represents the up-sampled LRMS image and f_W represents ResNet, discussed later. This term is motivated by the same goal as represented in Equation (3). As we will see, it enforces that \mathbf{X} shares the spectral content of \mathbf{M} . Unlike variational methods, we do not convolve \mathbf{X} with a smoothing kernel, instead allowing the deep network to correct for the high-resolution differences. In our experiments, we refer to this model as “spectra-mapping,” and use the ResNet model for f_W ; it corresponds to the middle network in Figure 3. For PanNet, we include this spectra-mapping procedure, as well as the following modification.

2.2.2 Structural preservation

As discussed in Section 2.1, many variational methods use the high-pass information contained in the PAN image to enforce structural consistency. These methods generate much clearer details than the original P+XS approach which directly uses images as in Equation (2). Based on this motivation, we input to the deep network f_W the high-pass content of the PAN image and of the up-sampled LRMS images. The modified model is

$$\mathcal{L} = f_W(G(\mathbf{P}), G(\mathbf{M})) + \|\mathbf{M} - \mathbf{X}\|_F^2. \quad (6)$$

To obtain this high pass information, represented by the function G , we subtract from the original images the low-pass content found using an averaging filter. For LRMS images, we up-sample to the size of PAN *after* obtaining the high-pass content. We observe that, since \mathbf{M} is low resolution, it can be viewed as containing the low-pass *spectral* content of \mathbf{X} , which the term $\|\mathbf{M} - \mathbf{X}\|_F^2$ models. This frees the network f_W to learn a mapping that fuses the high-pass *spatial* information contained in PAN into \mathbf{X} . We input $G(\mathbf{M})$ to the network in order to learn how the spatial information in PAN maps to the different spectral bands in \mathbf{X} . This objective corresponds to PanNet in Figure 3.

In Figure 4 we show an initial comparisons of the left and right networks of Figure 3. The HRMS and LRMS images have 8 spectral bands, which we represent as an average shown in grayscale. Figure 4(c) shows the mean absolute error (MAE) image of the ResNet reconstruction of (a), while (d) shows the MAE image for the proposed PanNet. As is evident, spectra-mapping can model spectral content better (evident in darker smooth regions), while training a network on the high-pass domain can preserve edges and details. These conclusions are supported by our extensive quantitative experiments. As mentioned in the introduction, another advantage of training a deep network on the high-pass domain is to remove inconsistencies between PAN and

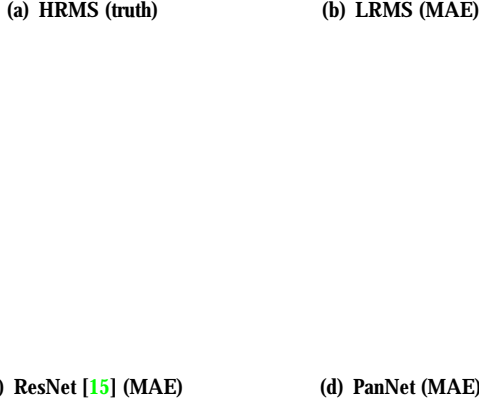


Figure 4. Example result on an 8-channel satellite image. (HRMS shown as average across channels.) For LRMS, ResNet and PanNet we show the MAE across these channels. We note the spectral distortion of ResNet. (Best viewed zoomed in on computer.)

HRMS images that arise in different satellites. This is not demonstrated here, but shown in Section 3.4.

2.2.3 Network architecture

The goal of recovering spatial information while preserving spectral information has motivated the objective proposed in Equation (6). In addition, previous variational methods try to improve performance by using prior image assumptions [3, 7], corresponding to f_3 in Equation (1). Here, we instead take advantage of deep learning to directly learn a function that captures the relationship between the PAN and LRMS inputs, and the HRMS outputs.

In [15], the authors propose a Residual Network (ResNet) for the image recognition task. While deep neural networks can learn complex functions of data, there are still some issues with their design. The ResNet framework was designed to ensure that the input information can be sufficiently propagated through all parameter layers when training a very deep network. We adopt the ResNet structure with convolutional neural networks as our network model f_W in Equation (6). The convolutional operation can help to model coupling between different bands of the multispectral images. Accordingly, our network structure is expressed

through the following operations:

$$\begin{aligned} Y^1 &= \max(W^1 \text{ stack}(G(P), G(M)) + b^1, 0), \\ Y^{2l} &= \max(W^{2l} Y^{2l-1} + b^{2l}, 0), \\ Y^{2l+1} &= \max(W^{2l+1} Y^{2l} + b^{2l+1}, 0) + Y^{2l-1}, \\ X &= W^L Y^{L-1} + b^L + M. \end{aligned} \quad (7)$$

Here, W denotes the weights and b denotes biases of the network, $l = 1, \dots, \frac{L-2}{2}$, and Y^l represents the output of the l th layer. For the first layer, we compute a_1 feature maps using a $S_1 \times S_1$ receptive field and a rectified linear unit, $\max(0, x)$. Filters are of size $c \times S_1 \times S_1 \times a_1$. $c = B + 1$, representing the fact that we stack the B LRMS images with the PAN image as shown in Figure 2. In layers 2 to $L - 1$ we compute a_2 feature maps using a $S_2 \times S_2$ receptive field and a rectified linear unit. The filters are size $a_1 \times S_2 \times S_2 \times a_2$. Finally, the last layer uses a $S_3 \times S_3$ receptive field and contains mostly spectral information. This is seen from the fact that we add the upsampled LRMS images M (spectra-mapping) to obtain an approximation to the ground truth HRMS image X .² Therefore, the network is modeling high-frequency edge information not contained in M . The penalty of the approximation is the Frobenius norm shown in Equation (6).

Although the parameter layers of our architecture follow ResNet, the two are different in the spectra-mapping procedure (bottom equation) and the high-pass inputs to the network (top equation). We compare this PanNet framework with ResNet directly applied to the image domain in our experiments to show the clear advantage of incorporating this additional domain knowledge. (We again recall that neither have been applied to the pan-sharpening problem.) We also compare with the state-of-the-art PNN [21], which uses a different deep CNN learning approach from ResNet.

3. Experiments

We conduct several experiments using data from the Worldview3 satellite. The resolution of PAN in this satellite ranges from 0.41m to 1.5m. We use stochastic gradient descent (SGD) to minimize the objective function in Equation (6). For our experiment we extract 18,000 PAN/LRMS/HRMS patch pairs of size 64×64 . We split this into 90/10% for training/validation. We compare with six widely used pan-sharpening methods: PRACS [8], Indusion [19], PHLP [18], BDSD [13], SIF [6, 7] and PNN [21]. Several parameter settings were used for each and the best performance selected.

We also experiment with the three networks of Figure 3: ResNet [15], ResNet+spectra-mapping and our final PanNet model. We use the Caffe package [17] to train these models. For SGD we set the weight decay to 10^{-7} , momentum to

²Alternatively, one could learn the upsampling from a CNN [10].

Table 1. Quality metrics at the lower scale of different methods on 225 satellite images from WorldView3.

Algorithm	Q8	QAVE	SAM	ERGAS	SCC	Time
BDSB [13]	0.871 ± 0.010	0.867 ± 0.013	7.158 ± 1.909	3.631 ± 0.621	0.856 ± 0.032	0.2s (CPU)
PRACS [8]	0.836 ± 0.023	0.822 ± 0.025	6.675 ± 1.628	3.834 ± 0.718	0.835 ± 0.040	0.9s (CPU)
Indusion [19]	0.799 ± 0.017	0.799 ± 0.015	6.385 ± 1.544	4.340 ± 0.699	0.825 ± 0.026	0.4s (CPU)
PHLP [18]	0.859 ± 0.013	0.835 ± 0.011	5.748 ± 0.926	3.747 ± 0.590	0.845 ± 0.024	18s (CPU)
SIRF [6, 7]	0.863 ± 0.013	0.859 ± 0.002	6.140 ± 1.416	3.564 ± 0.553	0.866 ± 0.019	27s (CPU)
PNN [21]	0.882 ± 0.005	0.891 ± 0.003	4.752 ± 0.870	3.277 ± 0.473	0.915 ± 0.009	0.2s (GPU)
ResNet	0.847 ± 0.005	0.886 ± 0.009	4.940 ± 0.941	3.838 ± 0.355	0.917 ± 0.012	0.2s (GPU)
ResNet+spectra-map	0.905 ± 0.010	0.905 ± 0.014	4.730 ± 0.959	2.933 ± 0.406	0.918 ± 0.019	0.2s (GPU)
PanNet (proposed)	0.925 ± 0.005	0.928 ± 0.010	4.128 ± 0.787	2.469 ± 0.347	0.943 ± 0.018	0.2s (GPU)
ideal value	1	1	0	0	1	

(a) LRMS

(b) PAN

(c) BDSB

(d) PRACS

(e) Indusion

(f) PHLP

(g) SIRF

(h) PNN

(i) PanNet (proposed)

(j) GroundTruth

Figure 5. Comparison of fusion results (source: WorldView3). Size of the PAN image is 400×400 .

0.9 and use a mini-batch size of 16. We start with a learning rate of 0.001, dividing it by 10 at 10^5 and 2×10^5 iterations, and terminate training at 2.5×10^5 iterations. We set the network depth $L = 10$, the filter sizes $s_1 = s_2 = s_3 = 3$ and filter numbers $a_1 = a_2 = 32$. The radius of the low-pass averaging filter used to calculate $G(\mathbf{P})$ and $G(\mathbf{M})$ is 5. It took about 8 hours to train each network.

3.1. Evaluation at lower scale

Since generally speaking HRMS images are not available in pan-sharpening datasets, we follow Wald's protocol [24] for our simulated experiments. In this case, all original images are filtered by a 7×7 Gaussian smoothing kernel and downsampled by a factor of 4. Experiments are conducted on these downsampled images. This way, we can treat the original LRMS images as though they were

the ground truth HRMS images. We simulate experiments on 225 test images from Worldview3 satellite using this experimental framework.

Each image contains eight spectral bands. For visualization we only show the three color bands, but our quantitative evaluation considers all spectral bands. We use five widely used quantitative performance measures to evaluate performance: relative dimensionless global error in synthesis (ERGAS) [23], spectral angle mapper (SAM) [26], universal image quality index [25] averaged over the bands (QAVE) and x-band extension of Q8 (for 8 bands) [2], and the spatial correlation coefficient (SCC) [27].

Table 1 shows the average performance and standard deviation of each method we compare with across the 225 satellite images. We observe that, not considering the networks of Figure 3, the deep PNN performs the best. It can

Table 2. Quality metrics tested at the original scale of different methods on 200 satellite images from WorldView3.

Algorithm	QNR	Q8	QAVE	SAM	ERGAS	SCC
BSDS [13]	0.803±0.048	0.832±0.054	0.795±0.061	5.295±1.595	5.039±1.130	0.802±0.049
PRACS [8]	0.871±0.021	0.962±0.012	0.954±0.013	2.540±0.641	2.340±0.580	0.980±0.004
Indusion [19]	0.876±0.034	0.958±0.017	0.952±0.017	2.330±0.620	2.382±0.594	0.970±0.008
PHLP [18]	0.896±0.035	0.962±0.015	0.954±0.017	2.430±0.567	2.194±0.525	0.980±0.003
SIRF [6, 7]	0.849±0.047	0.974±0.010	0.969±0.012	2.127±0.557	2.102±0.455	0.982±0.004
PNN [21]	0.880±0.022	0.946±0.028	0.938±0.026	3.284±0.596	2.758±0.659	0.979±0.004
ResNet	0.866±0.022	0.913±0.050	0.931±0.029	3.108±0.533	3.329±0.973	0.976±0.006
ResNet+spectra-map	0.891±0.017	0.974±0.010	0.969±0.012	2.234±0.526	1.895±0.517	0.989±0.003
PanNet (proposed)	0.908±0.015	0.976±0.011	0.972±0.013	1.991±0.511	1.780±0.494	0.991±0.002
ideal value	1	1	1	0	0	1

(a) LRMS

(b) PAN

(c) PHLP

(d) SIRF

(e) PNN

(f) PanNet (proposed)

Figure 6. Comparison of fusion results on a subset of best performing algorithms (source: WorldView3). The size of PAN is 400×400 .

be seen that PanNet significantly improves the results of PNN, which we believe is because of the additional features of spectra-mapping and high resolution inputs discussed in Sections 2.2.1 and 2.2.2. We also show the feed-forward computation time (the size of PAN is 400×400). On a GPU, the deep methods are very fast for practical use. We show a specific example at the reduced scale in Figure 5. It can be seen that Indusion, PRACS and PHLP introduce different levels of blurring artifacts. BSDS, SIRF and PNN show greater spectral distortion, especially in the left smooth area, but have good visual performance in terms of spatial resolution. To highlight the differences, we show the residuals of these images below in Figure 7. It is clear that our network architecture PanNet has the least spatial and spectral distortion when compared to the reference image, while PNN has significant spectral distortion.

3.2. Evaluation at the original scale

Since all models require an input and output and HRMS images are not available, we use Wald’s training strategy outlined in Section 3.1. To assess how these above models translate to the original resolution for *testing*, we use the resulting models to pan-sharpen 200 *different* WorldView3 satellite images at the original scale. That is, while we trained the model in the subsampled scale to have input/output pairs, we now show how this translates to testing new images at the higher scale. Since we only perform

testing in this experiment, we directly input the PAN and LRMS images into the learned PanNet model to obtain the predicted pan-sharpened HRMS images, and similarly, for PNN as well as the other models.

We show a typical example of the results in Figure 6. It can be seen that Indusion, PHLP still exhibit different degrees of blurring, while BSDS introduces block artifacts. Since we do not have the ground truth HRMS images, we show the residuals to the upsampled LRMS images in Figure 8. This residual analysis is different from our earlier experiments; since the up-sampled multispectral images lack high-resolution spatial information but contain precise spectral information, ideal reconstructions will have smooth regions close to zero while the edges of structures should be apparent.

To quantitatively evaluate, we follow [6] and downsample the output HRMS images and compare with the LRMS as ground truth. We also use a reference-free measure QNR [1], which doesn’t require ground truth. These results are shown in Table 2, where again we see the good performance of PanNet. We also observe that PNN does not translate as well from good training performance (Section 3.1) to testing performance on new images at a higher resolution.

3.3. Analysis of different network structures

We assess the training and testing convergence of different network structures, including ResNet, ResNet+spectra

(a) BDSD (b) PRACS (c) Indusion (d) PHLP (e) SIRF (f) PNN (g) PanNet

Figure 7. The residual images for Figure 5. In general, PanNet has better spatial resolution with less spectral distortion.

(a) BDSD (b) PRACS (c) Indusion (d) PHLP (e) SIRF (f) PNN (g) PanNet

Figure 8. The residual images corresponding to Figure 6, including for some reconstructions not shown above.

mapping and PanNet, using pixel-wise MSE on the 225 training (i.e., pixel average of Equation (6)), as well as on the 200 testing satellite images. These are shown in Figure 9 as a function of iteration. As is evident, PanNet exhibits considerably lower training and testing error (as expected given the previous quantitative evaluation). All algorithms converge at roughly the same speed.

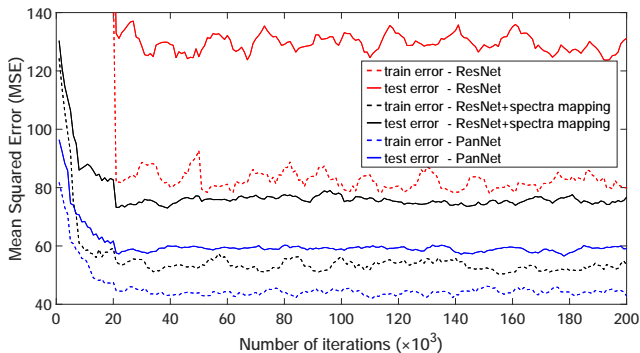


Figure 9. Convergence of different network structures.

3.4. Generalization to new satellites

We’ve motivated PanNet as being more robust to differences across satellites because it focuses on high-frequency content, and so networks trained on one satellite can generalize better to new satellites. To empirically show this we compare PanNet with PNN using data from the WorldView2 and WorldView3 satellites. For our comparison, we use two PNN-trained models: one we call PNN-WV2 which trains the PNN model using WorldView2 data; the other called PNN-WV3 is the model trained above on the WorldView3 dataset. We use the PanNet model trained on the

WorldView3 dataset in both cases. This section also shows some more qualitative evaluation of PanNet’s performance. To verify that the following results weren’t due to constant shift in spectrum between satellites that could be easily addressed, we tested PNN using various normalizations and obtained similar outcomes.

We show two typical visual results in Figures 10 and 11. The first contains a WorldView2 image, and the second a WorldView3 image. In both figures, we see that PNN does well on the image from the same satellite used during training, but that network does not translate to good performance on another satellite. In both figures, the cross-satellite PNN result suffers from obvious spectral distortions. PanNet, which is trained on WorldView3, translates much better to a new WorldView2 satellite image. This supports our motivation that performing the proposed “spectra-mapping” procedure allows the network to bypass having to model spectral information and focus on the structural information. This is not possible with the PNN model, which inputs the raw image into the deep CNN and models all information.

We also considered how PanNet and PNN generalize to a third satellite IKONOS. Since IKONOS data contains only 4 bands, we select R,G,B and infrared bands from WorldView3 and retrain the models. PNN-IK is trained on IKONOS, while PNN-WV3 is trained on the WorldView3 data using the same four bands as PanNet. As shown in Figure 12, although PNN-IK trains directly on IKONOS, our method still has clearer results. In the corresponding LRMS residual images in Figure 13, our method preserves spectral information (less color difference in smooth regions) and achieves better spatial resolution (clearer structures around edge regions). Again this supports that using a high-pass component to train PanNet can remove inconsistencies across different satellites.

(a) PNN-WV2 [21] (b) PNN-WV3 (c) PanNet (proposed) (d) Ground Truth

Figure 10. Example of network generalization (test on **WorldView2**). PNN-WV3 does not transfer well to WorldView2 image.

(a) PNN-WV2 [21] (b) PNN-WV3 (c) PanNet (proposed) (d) Ground Truth

Figure 11. Example of network generalization (test on **WorldView3**). PNN-WV2 does not transfer well to WorldView3 image.

(a) PNN-IK [21] (b) PNN-WV3 (c) PanNet (proposed) (d) LRMS

Figure 12. Example of network generalization (test on **IKONOS**).

(a) PNN-IK (b) PNN-WV3 (c) PanNet

Figure 13. The residual to the LRMS images. Our method preserves spectral information (less color difference in smooth regions) and achieves better spatial resolution (clearer structures around edge regions). This, and Figures 10–12, show that PanNet is fairly immune to inconsistencies across different satellites.

4. Conclusion

We have proposed PanNet, a deep model motivated by the two goals of pan-sharpening: spectral and spatial preservation. For spectral preservation, we introduce a technique called “spectra-mapping” that adds upsampled LRMS images to the objective function, allowing the network to focus only on details in the image. For spatial preservation, we train network parameters on the high-pass components of the PAN and upsampled LRMS images. We use ResNet as a deep model well-suited to this task. Compared with state-of-the-art methods, including PNN and vanilla ResNet, PanNet achieves significantly better image reconstruction and generalizes better to new satellites.

References

- [1] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva. Multispectral and panchromatic data fusion assessment without reference. *Photogrammetric Engineering & Remote Sensing*, 74(2):193–200, 2008. **6**
- [2] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini. A global quality measurement of pan-sharpened multispectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 1(4):313–317, 2004. **5**
- [3] H. A. Aly and G. Sharma. A regularized model-based optimization framework for pan-sharpening. *IEEE Transactions on Image Process.*, 23(6):2596–2608, 2014. **2, 4**
- [4] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé. A variational model for p+ xs image fusion. *International Journal of Computer Vision*, 69(1):43–58, 2006. **2**
- [5] W. J. Carper. The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data. *Photogrammetric Engineering and Remote Sensing*, 56(4):457–467, 1990. **2**
- [6] C. Chen, Y. Li, W. Liu, and J. Huang. Image fusion with local spectral consistency and dynamic gradient sparsity. In *CVPR*, pages 2760–2765, 2014. **4, 5, 6**
- [7] C. Chen, Y. Li, W. Liu, and J. Huang. Sirf: simultaneous satellite image registration and fusion in a unified framework. *IEEE Transactions on Image Process.*, 24(11):4213–4224, 2015. **2, 3, 4, 5, 6**
- [8] J. Choi, K. Yu, and Y. Kim. A new adaptive component-substitution-based satellite image fusion by using partial replacement. *IEEE Transactions on Geoscience and Remote Sensing*, 49(1):295–309, 2011. **2, 4, 5, 6**
- [9] X. Ding, Y. Jiang, Y. Huang, and J. Paisley. Pan-sharpening with a bayesian nonparametric dictionary learning model. In *AISTATS*, pages 176–184, 2014. **2**
- [10] C. Dong, C. Loy, K. He, and X. Tang. Learning a deep convolutional neural network for image super-resolution. In *ECCV*, 2014. **4**
- [11] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. **1**
- [12] F. Fang, F. Li, C. Shen, and G. Zhang. A variational approach for pan-sharpening. *IEEE Transactions on Image Process.*, 22(7):2822–2834, 2013. **2**
- [13] A. Garzelli, F. Nencini, and L. Capobianco. Optimal mmse pan sharpening of very high resolution multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(1):228–236, 2008. **2, 4, 5, 6**
- [14] A. R. Gillespie, A. B. Kahle, and R. E. Walker. Color enhancement of highly correlated images. ii. channel ratio and chromaticity transformation techniques. *Remote Sensing of Environment*, 22(3):343–365, 1987. **2**
- [15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. **3, 4**
- [16] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang. A new pan-sharpening method with deep neural networks. *IEEE Geoscience and Remote Sensing Letters*, 12(5):1037–1041, 2015. **1**
- [17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, 2014. **4**
- [18] Y. Jiang, X. Ding, D. Zeng, Y. Huang, and J. Paisley. Pan-sharpening with a hyper-laplacian penalty. In *ICCV*, pages 540–548, 2015. **2, 4, 5, 6**
- [19] M. M. Khan, J. Chanussot, L. Condat, and A. Montanvert. Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique. *IEEE Geoscience and Remote Sensing Letters*, 5(1):98–102, 2008. **2, 4, 5, 6**
- [20] P. Kwarteng and A. Chavez. Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. *Photogrammetric Engineering and Remote Sensing*, 55:339–348, 1989. **2**
- [21] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa. Pan-sharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016. **1, 2, 4, 5, 6, 8**
- [22] X. Otazu, M. González-Audiciana, O. Fors, and J. Núñez. Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods. *IEEE Transactions on Geoscience and Remote Sensing*, 43(10):2376–2385, 2005. **2**
- [23] L. Wald. *Data fusion: definitions and architectures: Fusion of images of different spatial resolutions*. Presses des MINES, 2002. **5**
- [24] L. Wald, T. Ranchin, and M. Mangolini. Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, 63(6):691–699, 1997. **5**
- [25] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002. **5**
- [26] R. H. Yuhas, A. F. Goetz, and J. W. Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL Airborne Geoscience Workshop: AVIRIS Workshop: Pasadena, CA, USA*, pages 147–149, 1992. **5**
- [27] J. Zhou, D. Civco, and J. Silander. A wavelet transform method to merge landsat tm and spot panchromatic data. *International Journal of Remote Sensing*, 19(4):743–757, 1998. **5**