- Supplemental Material -
SIPSA-Net: Shift-Invariant Pan Sharpening with Moving Object Alignment for Satellite Imagery

Jaehyup Lee Soomin Seo Munchurl Kim*

Korea Advanced Institue of Science and Technology (KAIST)
{woguq365, ssm9462, mkimee} @kaist.ac.kr

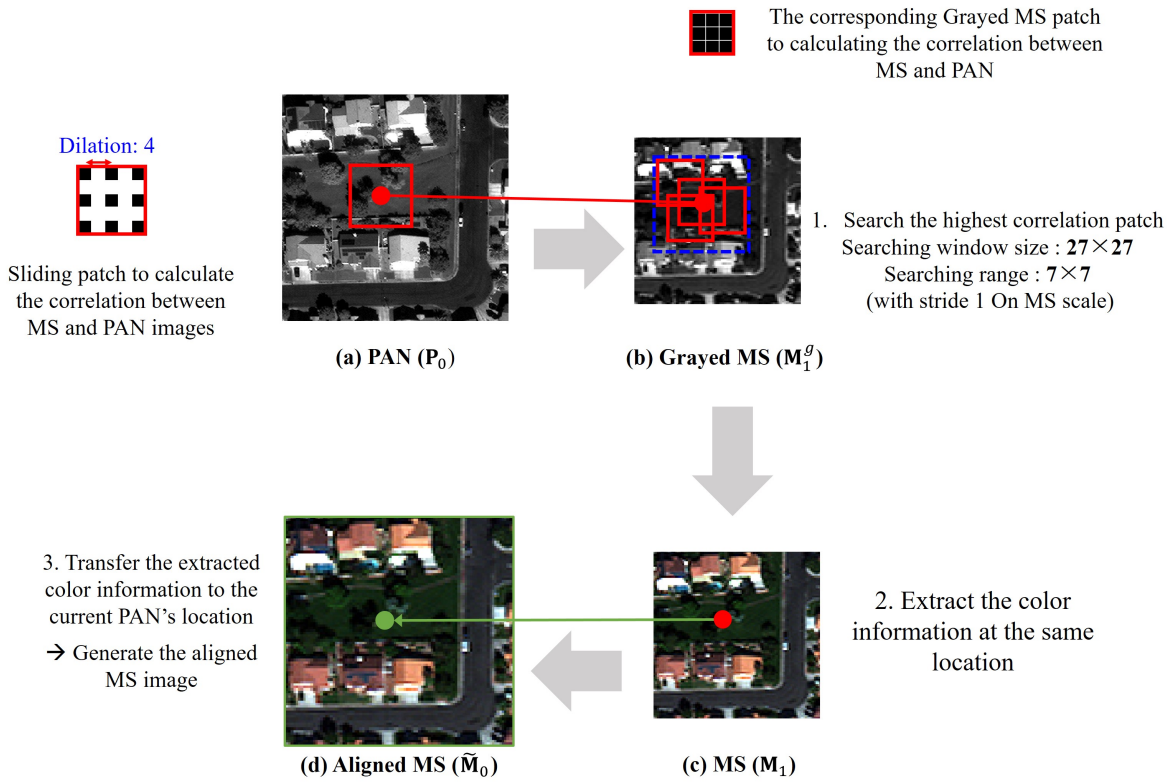
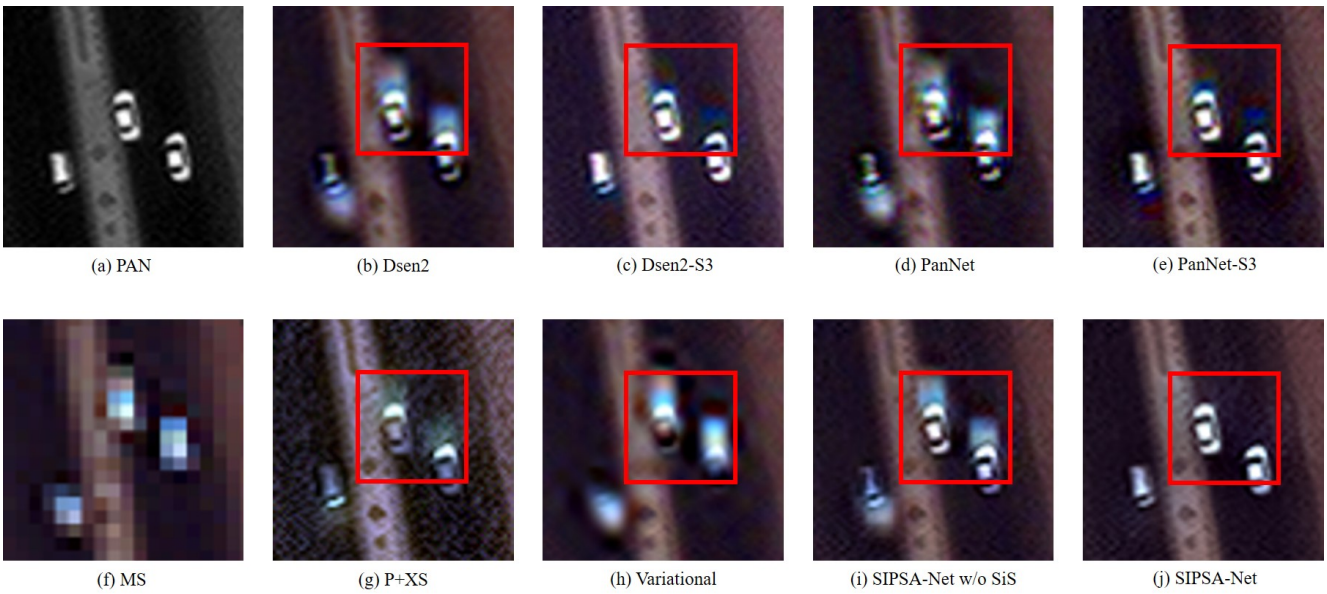


Figure 1: Aligned MS generation by correlation maximization

1. Aligned MS generation

To evaluate the degree of distortion of PS output images, the aligned MS images should be generated from the misaligned MS images. This is because the proposed SIPSA-Net is designed to correct the location of misplaced colors in MS images. Therefore, it is more proper to evaluate the performance using the aligned MS images rather than the original skewed MS images. The aligned MS images are created from the perspective of correlation maximization between the PAN and aligned MS image pair.

*Corresponding author.



(a) PAN (b) Dsen2 (c) Dsen2-S3 (d) PanNet (e) PanNet-S3

(f) MS (g) P+XS (h) Variational (i) SIPSA-Net w/o SiS (j) SIPSA-Net

Figure 2: Pan-sharpened images of pan-sharpening methods including SIPSA-Net.

The procedure of generating an aligned MS image is depicted in Fig. 1. First, the MS image to be aligned is converted to grayscale to match the channel with the PAN image. Then for each pixel location of the PAN image, we replace the current pixel with the optimal MS pixel that is aligned to the shape of the PAN through the correlation score maximization process. For a single-pixel location on the PAN image, let us assume that there is a patch with size 27×27 having dilation of 4 which has the center pixel located at the current pixel location. Then on the corresponding position at the grayed MS image, consider a patch with size 27×27 which has the center pixel located at the same pixel location. Then, we have two patches of the same size, one from the PAN image and the other from the grayed MS image. The correlation value is calculated between the two patches from the PAN image and grayed MS image.

The sliding window in the MS image is moved within the searching range of size 7×7 around the center pixel. Overall, for a single patch (single-pixel) on the PAN image, 49 patches within the search range (7×7) are used for calculating the correlation value with the patch from the PAN image. Then, the MS patch with the highest correlation value is selected, and the center pixel of the MS patch is selected as the optimal MS pixel. The searching process is repeated for all pixel positions of the PAN images. The aligned MS images of the PAN scales will then be used to evaluate the degree of spectral distortion of pan-sharpened images.

Table 1: QNR [1] and JQM [7] comparison (measured with original misaligned MS and aligned MS images).

	with aligned MS		with misaligned MS	
	$QNR\uparrow$	$JQM\uparrow$	$QNR\uparrow$	$JQM\uparrow$
P+XS [2]	0.860	0.921	0.869	0.919
Variational [4]	0.894	0.913	0.902	0.926
PanNet [9]	0.836	0.883	0.844	0.891
PanNet-S3 [3]	<u>0.947</u>	0.947	<u>0.939</u>	0.941
DSen2 [6]	0.848	0.894	0.856	0.900
DSen2-S3 [3]	0.898	<u>0.956</u>	0.889	0.947
SIPSA-Net (full)	0.899	**0.962**	0.890	**0.955**
SIPSA-Net (w/o SiS)	**0.954**	0.951	**0.947**	**0.955**

2. Analysis and Comparison of QNR and JQM

In this paper, we measure two no-reference metrics, joint quality measure(JQM) [7] and quality with no-reference (QNR) [1]. Several previous works have pointed out its drawbacks and unexpected properties [5, 7, 8], especially when perfect alignment between the MS and PAN images is not assured. As known, PAN and MS images in the WorldView-3 dataset are not well-aligned, so the values of the QNR metric does not correlate well with the observed visual quality. We also have intensively investigated this discrepancy between QNR metric and subjective quality for

PS output. As shown in Fig. 2, SIPSA-Net has better perceived visual quality than the SIPSA-Net trained without the SiS loss. SIPSA-Net (w/o SiS) suffers from the artifact that comes from the misaligned MS colors, as shown in the red boxes in Fig. 2. However, in Table 1, SIPSA-Net (w/o SiS) has higher QNR values when measured with both aligned and misaligned MS images. This is not coincident with the human perceived visual quality.

The inconsistency between QNR metric and perceptual visual quality comes from that QNR does not directly take account the spectral and spatial distortions when calculating the metric [1]. In order to remedy this problem, we additionally adopted another metric (JQM) which is known to be better agreed with the perceived visual quality on PS images [7]. As shown in Fig. 2, it can be easily noticed that the values of the JQM metric are very well agreed with the perceived visual qualities of the PS output. The PS results from DSen2-S3, PanNet-S3, and SIPSA-Net (full) have better visual quality compared to the others. The JQM metric shows higher values for the results from those methods, showing a similar tendency with perceived visual quality. Also, the PS output by our SIPSA-Net yields the highest JQM score, which is coincided with the perceived visual quality.

We analyze QNR and JQM for comparison, which can help understand why JQM is better correlated with perceptual visual quality than QNR.

2.1. Basis Functions of QNR and JQM

QNR utilizes the Q index as a basis function when measuring the difference between two images to be compared. The quality index Q between a reference original image x and a distorted image y is defined as:

$$Q(x,y) = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \times \frac{2\mu_x\mu_y}{\mu_x^2+\mu_y^2} \times \frac{2\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} \tag{1}$$

where μ_x, μ_y are means, σ_x, σ_y are standard deviations, and σ_{xy} is covariance for two image patches x, y. The first term ($\sigma_{xy}/\sigma_x\sigma_y$) is the correlation coefficient between x and y. The second term is always less than 1, from Cauchy-Schwartz inequality, and is sensitive to the bias in the mean of y with respect to x. The third term is also less than 1 and accounts for relative changes in the contrast between x and y. Q index has a dynamic range of [-1, 1], and the best value of $Q = 1$ is achieved *iff* $x = y$ for all pixels.

JQM utilizes composite measure based on means, standard deviations and correlation coefficient (CMSC), which is translation invariant with respect to means and standard deviations, thus enhanced over Q index. CMSC for a reference original image x and a distorted image to be tested y is defined as:

$$CMSC(x,y) = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}\left(1 - \frac{(\mu_x-\mu_y)^2}{R^2}\right)\left(1 - \frac{(\sigma_x-\sigma_y)^2}{(R/2)^2}\right) \tag{2}$$

where $R = 2^8 - 1 = 255$ for 8-bit data.

It should be noted that both QNR and JQM share an idea of measuring two different types of distortion (spectral distortion and spatial distortion) and merging these two distortion index values to obtain a final quality score for a PS image.

2.2. Spectral Distortion Index

First, we now compare the two spectral distortion terms of QNR and JQM. Spectral distortion of a PS image is measured with respect to an original input MS image. However, there exists a scale difference between the PS and MS images. Therefore, one should match the scale between them by upscaling the MS image or downscaling the PS image to calculate some metric.

The spectral distortion measure of QNR, D_λ, avoids downscaling of the PS image by comparing inter-band Q values, separately for both the MS and PS images which have different resolutions. D_λ is defined as follows:

$$D_\lambda = \frac{1}{N(N-1)} \sum_{l,k=1}^{N} |Q(MS_l, MS_k) - Q(PS_l, PS_k)| \tag{3}$$

where N is the number of bands. It should be noted for Eq. 3 that D_λ does not directly measure the spectral difference between the MS and PS images, but indirectly calculates it in terms of the difference between the inter-band Q values of the MS and PS images. Unfortunately, the evidence or proof that such inter-band relations hold between resolution scales is not studied intensively.

On the other hand, the JQM's spectral distortion index, called quality measure at low resolution (QLR), is defined in a low-resolution scale that makes a direct comparison between MS and PS images. QLR is defined as follows:

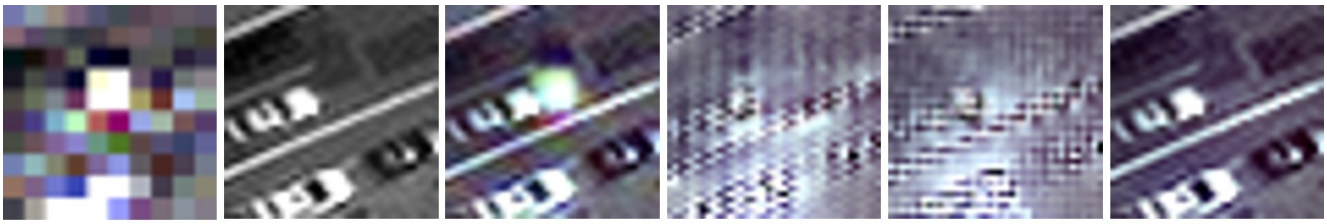
$$QLR = \frac{1}{N} \sum_{k=1}^{N} CMSC(MS_k, PS_{k,lpf\downarrow}) \tag{4}$$

$$PS_{k,lpf\downarrow} = (PS_k * lpf_k)\downarrow \tag{5}$$

where lpf_k is a Gaussian low pass filter, $*$ is a convolution operator, and $\downarrow$ indicates a down-scaling operation. As indicated in Eq. 4, the QLR of JQM directly reflects the spectral difference between the MS and PS images, as opposed to D_λ of QNR with the difference of the inter-band spectral differences.

(a) MS (b) PAN (c) SIPSA-Net w/o SiS (d) SIPSA-Net w/o FAM (e) SIPSA-Net w PAM (f) SIPSA-Net (full)

Figure 3: PS result images from the ablation study on SiS loss and FAM



(a) MS (b) PAN (c) SIPSA-Net w/o SiS (d) SIPSA-Net w/o FAM (e) SIPSA-Net w PAM (f) SIPSA-Net (full)

Figure 4: PS result images from the ablation study on SiS loss and FAM

2.3. Spectral Distortion Index

Next, we now compare the two spatial distortion terms of QNR and JQM. The spatial distortion of a PS image is to measure how much a pan-sharpening method can maintain the sharpness of the input PAN image in the PS output image. Therefore, the spatial distortion is measured between the PS and PAN images. However, the number of bands is different between the PS image (3 channels) and the PAN input image (1 channel). Therefore, one should fuse all three bands of the PS image to a single band and then measure the spatial distortion between them, or may measure the spatial distortion for each band of the PS image with the PAN image and then take a weighted sum to obtain a single value.

The spectral distortion measure of QNR, D_S, compares inter-band Q index values in a pairwise manner: between the PS and PAN images, and the MS image and a low-pass-filtered and downscaled PAN image. D_S is defined as follows:

$$D_S = \frac{1}{N}\sum_{k=1}^{N} |Q(MS_k, PAN_{lpf\downarrow}) - Q(PS_k, PAN)| \tag{6}$$

where $PAN_{lpf\downarrow} = (PAN * lpf) \downarrow$. In Eq. 6, it is worthwhile to note that the first spatial relation $Q(MS_k, PAN_{lpf\downarrow})$ is measured in terms of the difference between the two relations: the first relation is the Q index between each channel of the MS image and the low-pass-filtered and downscaled PAN image; and the second relation is the Q index between each channel of the PS image and the PAN image. Likewise for D_λ, D_S does not directly reflect the spatial distortion between the PS and PAN images at the full (PAN) resolution. Unfortunately, the evidence or proof that such a comparison of different spectral bands (narrow multispectral band and broad panchromatic band) is legitimate is not well studied.

On the other hand, the JQM's spatial distortion index, called quality measure at high resolution (QHR), is defined at a high (PAN) resolution scale that makes a direct comparison between the PAN image and an intensity image calculated as a weighted sum of the PS image channels (as a simulated PAN image). QHR is defined as follows:

$$QHR = CMSC(PAN, \sum_{k=1}^{N} w_k \cdot PS_k) \tag{7}$$

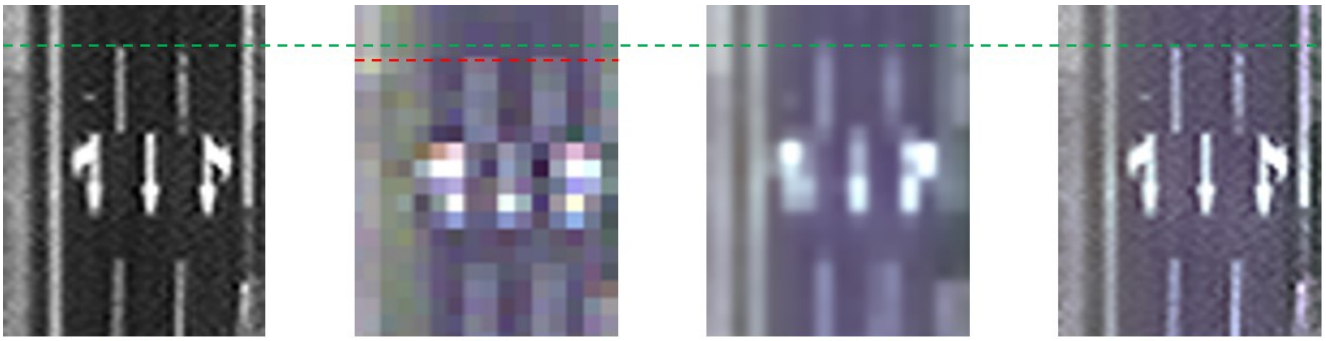
where w_k is a spectral response weight for the band k, which is calculated from spectral response functions of a data provider. As indicated in Eq. 7, the QHR of JQM directly reflects the spatial difference between the PAN and PS images, as opposed to D_S of QNR with the difference between the two relations: MS-PAN$_{lpf\downarrow}$ and PS-PAN.

2.4. Joint Quality Measures Based on Spectral and Spatial Index

Finally, QNR is defined as a product of two separate measure presented in 3 and 6 as

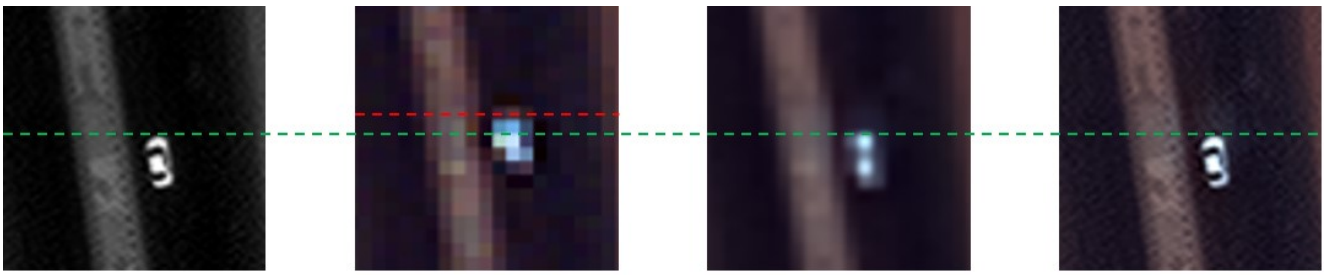
$$QNR = (1 - D_\lambda) \cdot (1 - D_s) \tag{8}$$

whereas JQM is defined as a weighted sum of separate measures presented in Eqs. 4 and 7 as



(a) PAN (b) MS (c) Our aligned MS (d) Our PS

Figure 5: Result images for alignment and pan-sharpening using our SIPSA-Net with global misaligned MS and PAN image pair.



(a) PAN (b) MS (c) Our aligned MS (d) Our PS

Figure 6: Result images for alignment and pan-sharpening using our SIPSA-Net with locally misaligned MS and PAN image pair.

$$JQM = v_1 \cdot QLR + v_2 \cdot QHR \tag{9}$$

where $v_1 + v_2 = 1$, and the weights are set to $v_1 = v_2 = 0.5$ as a default.

As described so far, the spectral distortion index (D_λ) of QNR is indirectly obtained by taking the difference between inter-band similarity measures of the MS and pan-sharpened images [1]. Similarly, the spatial distortion index (D_S) is measured in an indirect manner by taking the difference between the two relations: (i) each channel of an MS image and its corresponding low-pass-filtered and downscaled PAN image; (ii) each channel of a PS output image and a PAN image. These indirect measurement of spectral and spatial distortions leads to discrepancy between the QNR metric and human-perceived visual quality.

On the other hand, JQM [7] directly measures both the spectral distortion between MS and downscaled PS images, and the spatial distortion between PAN and fused PS images. These direct measurement of spectral and spatial distortions leads to better agreement with perceived visual quality than QNR. Throughout our intensive experiments, we also have found that the JQM is better correlated with perceived visual quality for PS output images from various methods, as shown in Fig. 2 and Table. 1.

3. Ablation study

To show the effectiveness of the key points of our SIPSA-Net, we provide a quantitative comparison among different versions of proposed network with and without feature alignment module (FAM) and SiS loss function. As shown in Fig. 3 and Fig. 4, only the full version of SIPSA-Net can effectively generate the pan-sharpened images of high quality. The SIPSA-Net trained without the SiS loss suffers from the misalignment between the MS and PAN image pair, therefore ghosting artifact appear. The SIPSA-Net trained without FAM and the version that was trained with the pixel alignment module (PAM) as described in the main paper could not be trained properly as can be seen in both Fig. 3 and Fig. 4. The result of ablation study shows the effectiveness of the proposed SiS loss and FAM in SIPSA-Net.

4. Effectiveness of Feature Alignment Module

Our SIPSA-Net generates two kinds of output images: an aligned MS image and a pan-sharpened image. As shown

in Fig. 5 and Fig. 6, the feature alignment module (FAM) can effectively generate the aligned MS image from the misaligned MS and PAN images. In Fig. 5 and Fig. 6, the red lines indicate the starting point of the misaligned MS's shape, and the green lines indicate the starting point of the aligned MS's shape (aligned to PAN). The visualization results show that the MS image alignment by FAM works well as we originally intended. Now the shapes in aligned MS images are well aligned to the details of PAN images.

References

[1] Luciano Alparone, Bruno Aiazzi, Stefano Baronti, Andrea Garzelli, Filippo Nencini, and Massimo Selva. Multispectral and panchromatic data fusion assessment without reference. *Photogrammetric Engineering & Remote Sensing*, 74(2):193–200, 2008.

[2] Coloma Ballester, Vicent Caselles, Laura Igual, Joan Verdera, and Bernard Rougé. A variational model for p+ xs image fusion. *International Journal of Computer Vision*, 69(1):43–58, 2006.

[3] Jae-Seok Choi, Yongwoo Kim, and Munchurl Kim. S3: A spectral-spatial structure loss for pan-sharpening networks. *IEEE Geoscience and Remote Sensing Letters*, 2019.

[4] Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. A variational pan-sharpening with local gradient constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10265–10274, 2019.

[5] Manjunath V Joshi and Kishor P Upla. *Multi-resolution Image Fusion in Remote Sensing*. Cambridge University Press, 2019.

[6] Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319, 2018.

[7] Gintautas Palubinskas. Joint quality measure for evaluation of pansharpening accuracy. *Remote Sensing*, 7(7):9292–9310, 2015.

[8] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2014.

[9] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5449–5457, 2017.