IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 11, 2020, accepted October 30, 2020, date of publication November 4, 2020, date of current version November 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3035802

UPSNet: Unsupervised Pan-Sharpening Network With Registration Learning Between Panchromatic and Multi-Spectral Images

SOOMIN SEO[1], JAE-SEOK CHOI[2], JAEHYUP LEE[1], HYUN-HO KIM[3], DOOCHUN SEO[3], JAEHEON JEONG[3], AND MUNCHURL KIM[1], (Senior Member, IEEE)

[1] Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea
[2] Samsung Advanced Institute of Technology, Suwon 16678, South Korea
[3] School of Electrical Engineering, Korea Aerospace Research Institute, Daejeon 34133, South Korea

Corresponding author: Munchurl Kim (mkimee@kaist.ac.kr)

This work was supported by Korea Aerospace Research Institute (KARI) under the project of 'Super-Resolution and PAN Colorization for Satellite Imagery'. The work was done while the Jae-Seok Choi was at KAIST.

ABSTRACT Recent advances in deep learning have shown impressive performances for pan-sharpening. Pan-sharpening is the task of enhancing the spatial resolution of a multi-spectral (MS) image by exploiting the high-frequency information of its corresponding panchromatic (PAN) image. Many deep-learning-based pan-sharpening methods have been developed recently, surpassing the performances of traditional pan-sharpening approaches. However, most of them are trained in lower scales using misaligned PAN-MS training pairs, which has led to undesired artifacts and unsatisfying visual quality. In this paper, we propose an *unsupervised* learning framework with *registration learning* for pan-sharpening, called UPSNet. UPSNet can be effectively trained in the original scales, and *implicitly* learns the registration between PAN and MS images without any dedicatedly designed registration module involved. Additionally, we design two novel loss functions for training UPSNet: a guided-filter-based color loss between network outputs and aligned MS targets; and a dual-gradient detail loss between network outputs and PAN inputs. Extensive experimental results show that our UPSNet can generate pan-sharpened images with *remarkable* improvements in terms of visual quality and registration, compared to the state-of-the-art methods.

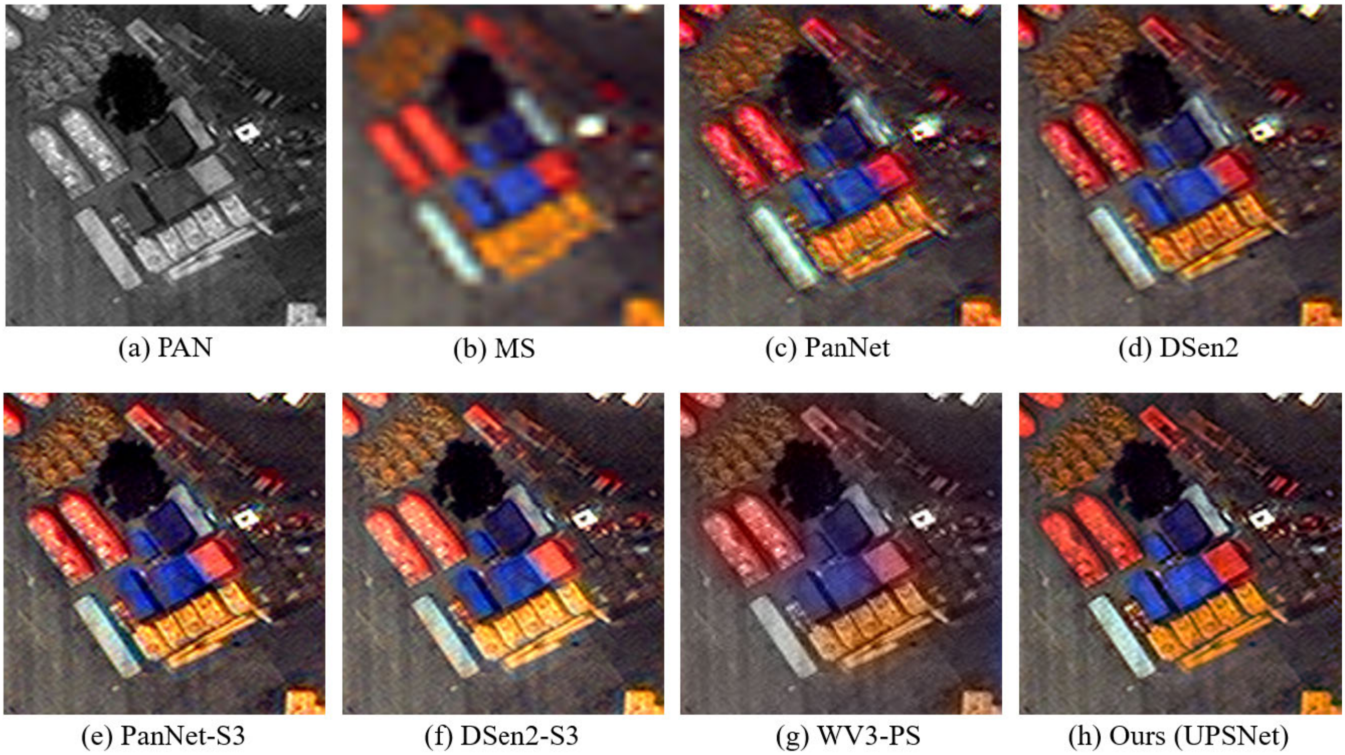
INDEX TERMS Pan-sharpening, pan-colorization, image restoration, deep-learning, convolutional neural networks (CNN), satellite imagery.

I. INTRODUCTION

With the advent of deep-learning, many deep-learning-based methods have been proposed to solve various image restoration problems, i.e., super-resolution [7], [18], [20], [22], [35], showing state-of-the-art performances in terms of reconstruction quality. Likewise, the growing usage of deep-learning for satellite imagery research can be observed recently. Satellite imageries contain various scenes around the world. The research areas for satellite imagery include prediction of forest growth, classification of crops, buildings and roads, environmental monitoring, and many other applications. To achieve high performance for solving such problems, it is essential to obtain high-quality, high-resolution satellite image datasets. However, due to the constraints of intrinsic satellite sensor resolutions and transmission bandwidths, most satellites acquire multi-spectral images with varying resolutions for the same geographical regions. In general, satellite images are comprised of pairs of low-resolution (LR) multi-spectral (MS) images of a larger ground sample distance (GSD) and high-resolution (HR) panchromatic (PAN) images of a smaller GSD. Pan-sharpening or pan-colorization is the task of generating pan-sharpened (PS) multi-spectral images which have the same spatial resolutions as the PAN images, by fusing the high-frequency details from the PAN images and the color information from the MS images. Fig. 1 shows an example pair of PAN, MS and PS results from various pan-sharpening approaches, including the proposed method.

Recently, several works on pan-sharpening have been proposed that incorporate learning models with convolutional neural networks (CNN) [4], [6], [12], [15], [19], [21], [28],

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


(a) PAN (b) MS (c) PanNet (d) DSen2

(e) PanNet-S3 (f) DSen2-S3 (g) WV3-PS (h) Ours (UPSNet)

FIGURE 1. Pan-sharpening results using various methods and the proposed method.

[33], [38], [43], [44]. These methods are based on supervised learning (Fig. 2-a) that often requires a degradation model to prepare a training dataset of PAN-MS pairs. For this, the original PAN-MS pairs are degraded (down-scaled) to LR PAN-MS pairs which are then used as inputs to the networks, and the original MS images are used as pseudo ground truth for training. In doing so, the networks are trained to output down-scaled PS images of input MS scales in such a lower scale scenario. Therefore, when these networks are tested under the original scale scenario, they perform poorly where the networks yield the PS images of input PAN scales. To overcome the scale (resolution) mismatch between training and testing, we propose an effective *unsupervised* learning framework for pan-sharpening, where a ground truth is not required for training. This enables the network to be trained and tested on the same scales, resulting in better visual quality.

Since the ground truth data are not available in pan-sharping, conventional supervised PS methods could not help but utilize the lower scale scenario. These methods optimize their PS outputs with mean absolute error (MAE) or mean squared error (MSE) loss using pseudo ground truth MS image. In our unsupervised PS (Fig. 2-b), where no ground truth image is required, we design two novel loss functions so that our UPSNet can effectively learn the high-frequency details from PAN inputs and color information from MS inputs in the original scale scenario without any pseudo ground truth: one is a dual-gradient detail loss between network outputs and PAN inputs; and the other is a guided-filter-based color loss between network outputs and our aligned MS targets.

One of the main difficulties of the pan-sharpening task is a misalignment between PAN and MS image pairs. PAN and MS images often have the misalignment of some pixel distances due to inherent limitations in satellite sensor arrays and acquisition time difference. A misaligned dataset used for training often entails undesired artifacts in pan-sharpened results such as double edge and color spread artifacts. To remedy this problem, we incorporate a preprocessing step only during the training where each MS image is registered to its corresponding PAN image in the sense of correlation maximization. The aligned MS images are not used as inputs to the network but are used as targets for the color loss. By doing so, our UPSNet can learn to implicitly match the high-frequency information from PAN inputs and color information from misaligned MS inputs during training, without any dedicatedly designed registration module. The trained UPSNet can then properly handle misaligned PAN-MS input pairs during testing. As shown in Fig. 1, the output image from UPSNet shows that structures and colors of the objects are better well-aligned compared to the other five methods. We can also observe that the produced pan-sharpened image from UPSNet has the most similar color compared with that of the input MS image while preserving the strong edges from the corresponding PAN image.

Furthermore, we found that a patch-based normalization can effectively deal with non-stationary PAN and MS input images of various pixel intensity distributions depending on geographical features, which often leads to color distortion in the pan-sharpened results. Similar to a batch normalization [13], this reduces the internal covariate shift and enables faster and more stable training of the network, which could possibly result in higher performance. Besides, applying local normalization helps maintain the color information of the MS input. This allows the network trained on the images acquired by a specific satellite to be well generalized for unseen images of other satellites. Our contributions can be summarized as follows:

- We propose a novel *unsupervised* learning framework for pan-sharpening where our proposed UPSNet can achieve state-of-the-art performance for most metrics and shows significantly better visual quality when tested on the original scale.
- Two novel loss functions for pan-sharpening are proposed, which effectively fuse the high-frequency details from PAN images and color information from MS images: a dual gradient detail loss and a guided-filter-based color loss. The dual gradient detail loss can appropriately handle different characteristics of PAN and MS image signals, so that UPSNet can effectively learn the details of PAN images. The guided-filter-based color loss allows UPSNet to effectively learn the color information from aligned and upscaled target MS images.
- With a preprocessing step of correlation-based alignment between PAN and MS images only for training, UPSNet can be trained to implicitly handle the inherent misalignment between PAN and MS input images without the preprocessing step in testing.
- We propose a simple yet very effective patch-based normalization technique that boosts up the generalization capability of our UPSNet for PAN-MS images of various satellites.

II. RELATED WORKS

A. TRADITIONAL PAN-SHARPENING METHODS

Before the advent of deep-learning, pan-sharpening algorithms were based on component substitution, multiresolution analysis, and model learning. Component substitution methods [5], [9], [17], [34], [42] apply spectral transformations on an interpolated MS input, and its spatial channel is replaced with a modified PAN. Multiresolution analysis based methods [27], [36] fuse the high-frequency details of PAN images into up-sampled MS input images. To decompose such high-frequency components, wavelet or undecimated decomposition techniques are utilized. Then these decomposed components are incorporated into interpolated MS input images to form pan-sharpened images. These methods have relatively low computational complexity but tend to produce the resulting images with mismatched spectral information and artifacts because they do not consider local properties of MS and PAN images. Model learning-based methods [11], [29], [31] learn pan-sharpening models by using regularization terms. In these methods, pan-sharpening is defined as an ill-posed problem, where a certain model is optimized to generate an output image so that a similarity metric between the output and target pan-sharpened image is maximized. These methods tend to produce pan-sharpened images with better quality having well-preserved spectral information, but require high computational complexity compared to the previously mentioned methods.

B. DEEP-LEARNING-BASED PAN-SHARPENING METHODS

Recent pan-sharpening methods incorporate various types of CNN structures. Pan-sharpening CNN (PNN) [28] is known to be the first CNN-based pan-sharpening method, showing competitive performance compared to conventional methods. The PNN adopted a shallow 3-layered network structure from SRCNN [7], which is the first super-resolution method to use CNN. Inspired by the success of ResNet [10] in classification, Yang *et al.* [43] proposed PanNet that has adopted the ResNet structure as their backbone network, where residual connection enables the network to focus on preserving the high-frequency details. PanNet applies high-pass filtering to MS and PAN inputs, and their edge components are used as network inputs. This enables better network generalization, being robust for unseen satellite datasets.

By adopting the network architecture of the state-of-the-art SR network, EDSR [22], Lanaras *et al.* [19] proposed a deep network (DSen2) and a deeper network (VDSen2) for super-resolution of the Sentinel-2 satellite images. DSen2 and VDSen2 are not exactly pan-sharpening methods since they super-resolve the images in 9 lower-resolution bands using the images in 4 higher-resolution bands as guidance. PAN images are not included in the Sentinel-2 dataset. PanNet and DSen2 show top performance in various quantitative metrics, producing PS images with high visual quality. Zhang *et al.* proposed a bidirectional pyramid network [45] that processes the MS and PAN images in two separate branches, which allows the spatial detail features from the PAN branch to be fused into the spectral information features of the MS branch, finally generating the output pan-sharpened images. This type of feature fusion has improved the preservation of high-frequency spatial information from PAN images.

Recently, Choi *et al.* proposed an S3 [6] loss, which considers the correlation between PAN and MS images. The S3 loss is devised to be applied adaptively for the areas according to the correlation values between MS and PAN images, thus reducing the ghosting artifacts around moving objects such as cars on the roads. Although the aforementioned deep-learning-based methods have greatly enhanced the performances and visual qualities over the traditional methods, they still have some limitations that those methods were trained in lower scales in a supervised manner, resulting in suboptimal PS outputs.

Recently, a few attempts have been made to tackle the drawbacks that come from supervised learning with pseudo ground truth. Ma *et al.* [26] proposed an unsupervised scheme based on a generative adversarial network with spatial and spectral discriminators. PercepPan [46] adopted an auto-encoder architecture into their unsupervised PS network

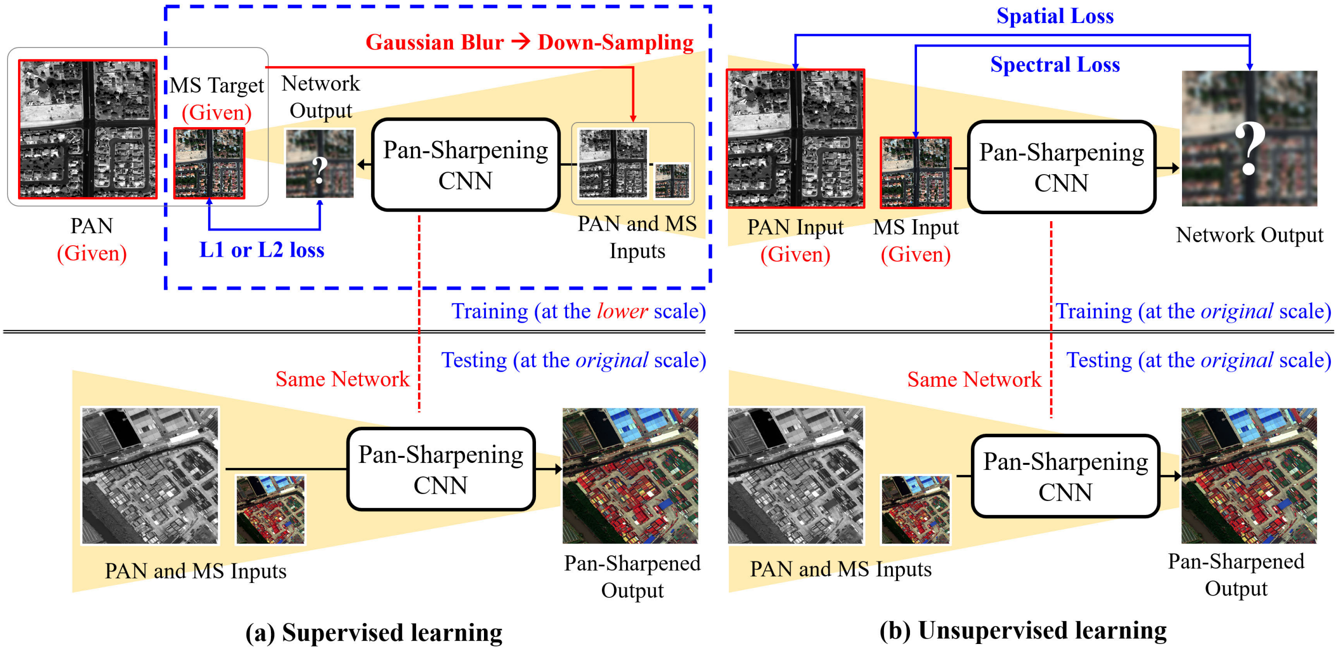


FIGURE 2. **Comparison between two different learning frameworks (a) Conventional supervised learning framework for pan-sharpening (b) Unsupervised learning framework.**

design, and utilized a perceptual loss to improve visual quality. Qu *et al.* incorporated a self-attention mechanism [32] that estimates spatially varying detail extraction and injection functions. Luo *et al.* also proposed an unsupervised pan-sharpening method [25] with an iterative fusion network. Although these unsupervised PS methods resolved the drawbacks of training in lower scales, none of them considered the inherent misalignment between MS and PAN inputs.

III. PROPOSED METHOD

As aforementioned, the pan-sharpening (PS) is defined as a task to obtain high-quality PS images using high-resolution (HR) PAN images and their corresponding low-resolution (LR) MS images. The resulting PS images should have the high-frequency detail information of the PAN images and the color information of the MS images as similar as possible. To avoid the drawbacks that come from training PS networks using pseudo ground truth images, our UPSNet learns the pan-sharpening in the original scale scenario, as shown in Fig. 2(b). Another root cause of inferior visual quality of previous pan-sharpening methods is a misaligned PAN-MS input pair. To allow UPSNet to implicitly handle the misalignment between PAN and MS images, which we call ‘‘registration learning’’, a data preparation step is introduced with a correlation-based alignment between PAN and MS images, which is only used during the training. To effectively train our UPSNet, we present two different types of loss functions, which allow the network to learn spatial information from PAN inputs and spectral information form MS inputs to produce high-quality PS images. Note that the training of UPSNet is done in the original scales of PAN and MS images, where the testing is also taken place. In order to handle diverse characteristics of PAN and MS images taken from different satellites with UPSNet, we propose a simple but very effective patch-based normalization technique to have a generalization capability for PAN-MS images from various satellites. More details for loss functions, registration method, and normalization will be thoroughly explained in the following subsections.

A. FORMULATIONS

In general, satellite imagery datasets include PAN images of higher resolution (smaller GSD), denoted as $\mathbf{P}_0$, and the corresponding MS images of lower resolution (larger GSD), denoted as $\mathbf{M}_1$. The subscript number denotes a level of resolution, where a smaller number indicates a higher resolution. Our final goal in pan-sharpening is to utilize $\mathbf{P}_0$ and $\mathbf{M}_1$ to generate a high-quality pan-sharpened image $\mathbf{S}_0$ which has the same resolution as $\mathbf{P}_0$ and similar spectral information of $\mathbf{M}_1$. This requires a pan-sharpening model $\mathbf{g}$ which takes $\mathbf{P}_0$ and $\mathbf{M}_1$ as inputs and yields a pan-sharpened image $\mathbf{S}_0$ as an output. In the conventional CNN-based pan-sharpening based on supervised learning, their models are trained using $\mathbf{P}_0$ and $\mathbf{M}_1$ as targets and their down-scaled version $\mathbf{P}_1$ and $\mathbf{M}_2$ as inputs, where their training is done in a lower scale scenario.

B. UNSUPERVISED LEARNING FRAMEWORK FOR PAN-SHARPENING

One of the main limitations of the previous CNN-based pan-sharpening methods is that the PAN-MS pairs are down-scaled to enable supervised learning. These networks are only trained in the lower scale scenario, so they perform poorly

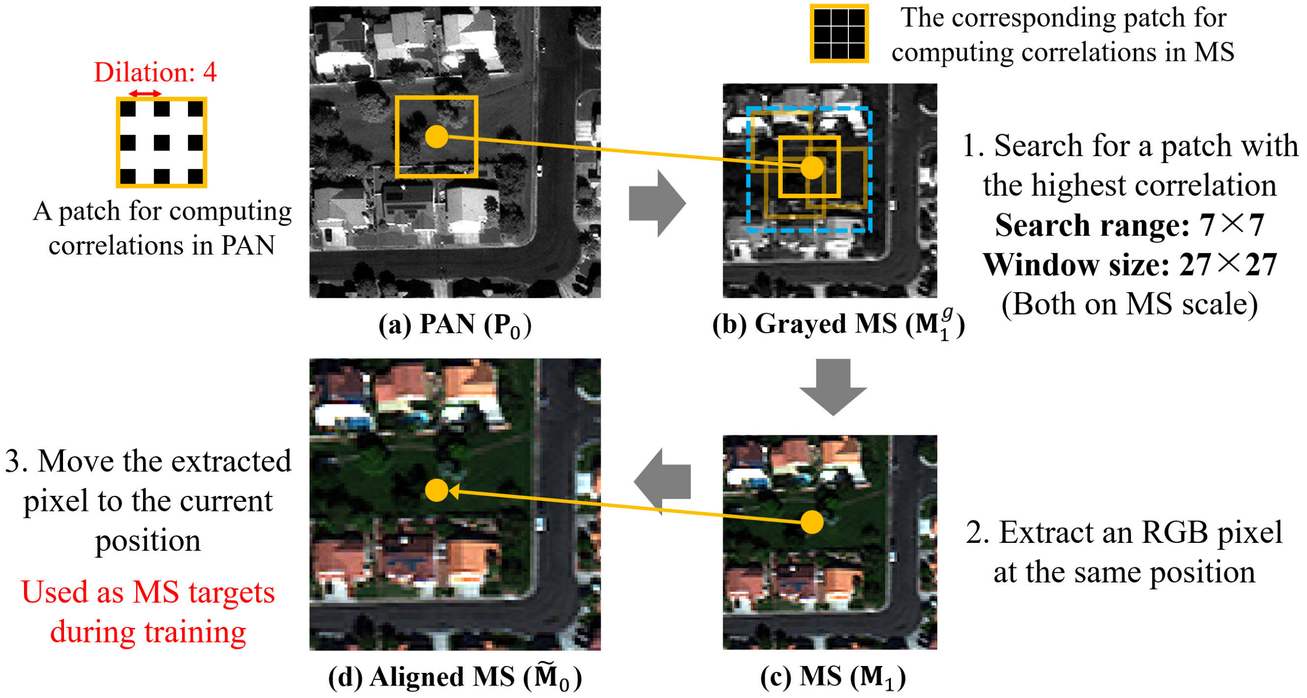

Dilation: 4
A patch for computing correlations in PAN
The corresponding patch for computing correlations in MS
(a) PAN ($\mathbf{P}_0$)
(b) Grayed MS ($\mathbf{M}_1^g$)
1. Search for a patch with the highest correlation
Search range: 7×7
Window size: 27×27
(Both on MS scale)
2. Extract an RGB pixel at the same position
(c) MS ($\mathbf{M}_1$)
(d) Aligned MS ($\tilde{\mathbf{M}}_0$)
3. Move the extracted pixel to the current position
Used as MS targets during training

FIGURE 3. Proposed PAN-MS registration method based on correlation maximization.

when tested in the original scale scenario which is always a realistic case. Since the misalignment between MS and PAN images would be more severe in their original scales, the networks trained in such a lower scale scenario are not able to appropriately handle the PAN and MS input images with larger misalignment.

On the contrary, the proposed unsupervised learning framework can overcome this problem, as our network is trained and tested under the same original scale scenario. The conceptual difference between conventional methods and the proposed framework is depicted in Fig. 2.

Unlike the conventional methods in Fig. 2-(a) for pan-sharpening that are trained under a lower-scale scenario, UPSNet is trained and tested under the same original scale as depicted in Fig. 2-(b). For the training, unlike the lower-scale scenario, the original PAN images are used as targets for a detail loss, and the aligned MS images of the same scale as PAN images are used as targets for a color loss. By doing so, our UPSNet can be trained in the original scale scenario. Here, one of the main points is how to obtain the aligned MS images of the same scale as the PAN and PS images. This will be detailed in the following subsections.

C. REGISTRATION

The conventional pan-sharpening methods that were trained with L1 or L2 loss functions on the misaligned datasets tend to produce the PS images of inferior visual quality, including double edge and spread color artifacts. To remedy this, it is necessary to use aligned datasets for the training of pan-sharpening networks. For the alignment between PAN and MS images, we propose a novel correlation-based PAN-MS registration on the PAN scale, which is done off-line. The resulting MS images have the same size as PAN images and are aligned to the PAN images. It should be noted that the aligned MS images are used as targets in the color loss function during training, not as the input for the network. In doing so, UPSNet internally learns the registration for the misaligned PAN-MS input pairs. That is, the aligned MS image is not required during the test.

Fig. 3 shows the off-line alignment steps. For a given pair of an original PAN image $\mathbf{P}_0$ and a grayed MS image $\mathbf{M}_1^g$, a PAN-sized aligned MS image $\tilde{\mathbf{M}}_0$ is constructed via a correlation-based searching process. For each pixel location of the PAN image, an optimal multi-channel (e.g., RGB) pixel value in the MS image is selected and placed in the corresponding pixel location of the aligned MS image of the PAN scale. The optimal pixel is determined as the center pixel of a searching window that finds the highest correlation value between the PAN and gray MS images is found by searching the grayed MS image within a search region. When the searching window size for the gray MS image is $M \times M$, the corresponding window size for the PAN image is set to $dM \times dM$ where d is a dilation equivalent to the scale difference between PAN and MS images.

The details of the searching process are as follows: First, we obtain a grayed MS image (Fig. 3-(b)) where a searching window of size 27×27 with dilation 1 is applied. The searching window slides in a pixel-wise manner of stride 1 within a pre-defined search region of size 7×7. The corresponding window size applied for the PAN image is of size 27×27 with

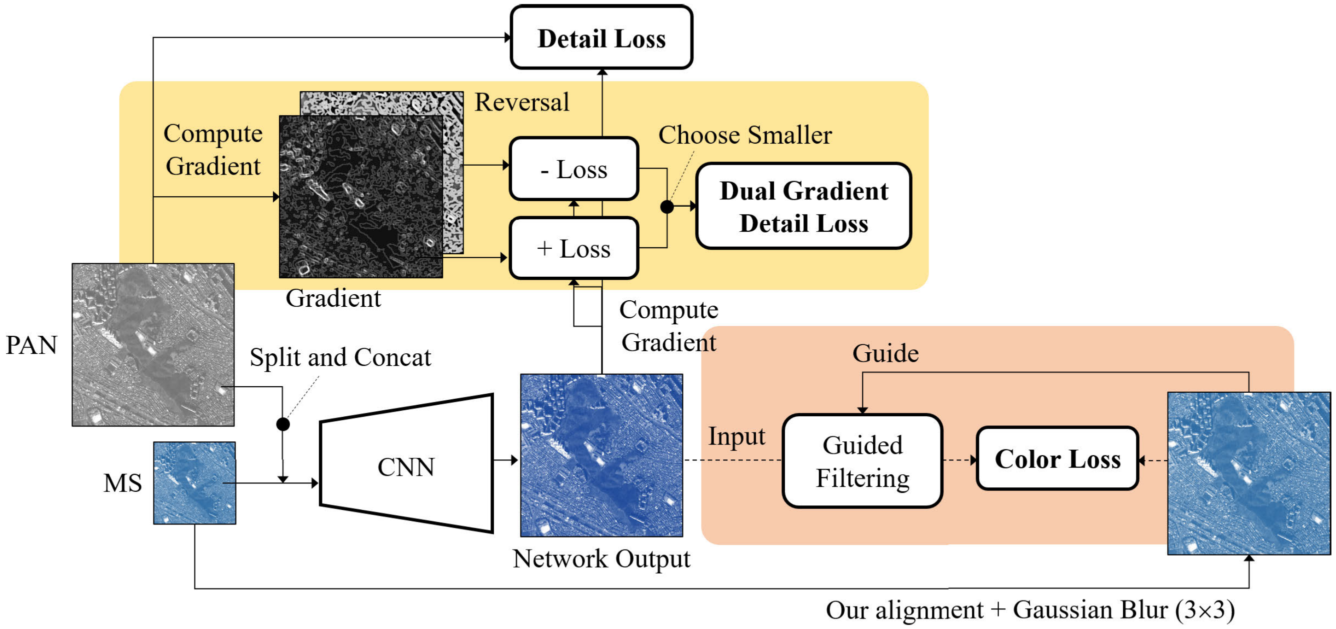


FIGURE 4. Overall training process and loss functions for our UPSNet.

dilation 4 due to the 4 times resolution difference between PAN and MS images. The goal of this registration is to replace all pixels in PAN with the best matching MS pixels, so that we can get MS images that are well aligned to their corresponding PAN images. Therefore, for a current pixel location of the PAN image, we search for the best matching patch with the highest correlation value in a search region of the grayed MS image. The 49 correlation values are calculated by sliding the searching window of size 27 × 27 with stride 1 in the 7 × 7 pixel grid (search region) centered at the current pixel location. When the best match is found, the MS pixel corresponding to the center pixel of the searching window is placed in the corresponding pixel location of the aligned MS image of the PAN scale. 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 as MS targets for the color loss during training.

The above correlation-maximization-based registration involves two hyper-parameters: the searching window size (27 × 27) and the search range size (7 × 7) that are set empirically through extensive experiments. The searching window size should be large enough to capture a sufficient amount of local structures for correlation calculation but at the expense of computational complexity. A too small-sized searching window will ignore neighboring pixel correlation, and a too large one may ignore some misaligned pixels in correlation matching because the amount of misalignment gets relatively small. For the search range, a larger search range would be beneficial in handling larger misalignment but also at the expense of computational complexity. The search range size of 7 × 7 is large enough to handle the inherent misalignment between PAN and MS images for our experiments because it can cope with up to 3-pixel misalignment in MS scale that corresponds to a maximum 12-pixel misalignment in PAN scale.

It is worthwhile to mention some other alignment options to perform alignment in the MS scale. In this case, the computation of color loss can have two possible options for matching the scale (resolution), where PS images and MS images have different resolutions. The first option is to downscale the PS images to the MS scale by applying a degradation model, which causes the resulting trained PS networks to yield PS outputs with checkerboard artifacts. The second option is to upscale the aligned MS images (aligned in the MS scale) to have the same resolution as PS images. However, this causes a new misalignment due to the upscaling process, thus leading to the degraded quality of the PS output. The experimental results for these options are provided in Sec. IV-C2.

D. LOSS FUNCTIONS

Previous deep-learning-based methods in supervised learning have applied a degradation model to the input images $\mathbf{P}_0$ and $\mathbf{M}_1$, which yields $\mathbf{P}_1$ and $\mathbf{M}_2$. Then, the network output $\mathbf{S}_1$ is compared to the pseudo ground truth MS images $\mathbf{M}_1$ by using a simple L1 or L2 loss between them. On the other hand, to train the network in an unsupervised manner, we propose two different types of loss functions: First, a detail loss that enforces the network output to have similar details (high-frequency information) with PAN images $\mathbf{P}_0$; Secondly, a color loss that helps the network match the spectral information of the network output $\mathbf{S}_0$ and the aligned PAN-resolution MS image $\widetilde{\mathbf{M}}_0$. More details for the proposed loss functions will be thoroughly explained in the following.

1) DETAIL LOSS

We now define a detail loss that minimizes spatial distortions between network outputs $\mathbf{S}_0$ and PAN inputs $\mathbf{P}_0$. We first obtain grayed PS outputs $\mathbf{S}_0^g$. In general, a vanilla detail loss, which is a simplified version of the spatial loss [6], can be defined as

$$L_d = \sum ||d(\mathbf{S}_0^g) - d(\mathbf{P}_0)||_1^1 \tag{1}$$

where $d(\cdot)$ is a gradient extractor using horizontal and vertical difference (e.g. [1, -1]) operators.

One of the difficulties in pan-sharpening tasks is inherent differences in image signal characteristics between the PAN and MS images. PAN images generally cover a wide range of wavelengths by merging a broad spectrum of visible lights into a single-channel image. Therefore, luminance values in MS images considerably differ from the PAN images. For example, certain objects that appear bright in an MS image (e.g., water) can appear dark in a corresponding PAN image or vice-versa (e.g., trees, grass). When we consider three bands (R, G, B) in MS images separately, the luminance difference between each of the bands and PAN images would be even larger than comparing with the grayscale versions of the MS images.

This inherent luminance difference between PAN and MS images generates not only dissimilar luminance values but also opposite directions of intensity gradients between them, which hinders deep-learning networks from properly learning the task of pan-sharpening. To solve this, we propose a novel loss function, called a dual-gradient detail loss, which is specially designed to handle such opposite gradient directions. This loss is utilized to enforce the PS outputs to have similar edge details with PAN images, together with the vanilla detail loss. Our dual-gradient detail loss is defined as

$$\begin{aligned} L_{dg} = \sum & \min(||d(\mathbf{P}_0)|! -d(\mathbf{S}_0^R)||_1^1, || - d(\mathbf{P}_0) - d(\mathbf{S}_0^R)||_1^1) \\ & + \min(||d(\mathbf{P}_0) - d(\mathbf{S}_0^G)||_1^1, || - d(\mathbf{P}_0) - d(\mathbf{S}_0^G)||_1^1) \\ & + \min(||d(\mathbf{P}_0) - d(\mathbf{S}_0^B)||_1^1, || - d(\mathbf{P}_0)-d(\mathbf{S}_0^B)||_1^1), \end{aligned} \tag{2}$$

where $-d(\mathbf{P}_0)$ is the reversed gradient map of PAN input $d(\mathbf{P}_0)$, and $\mathbf{S}_0^R$, $\mathbf{S}_0^G$ and $\mathbf{S}_0^B$ are R, G, B channels of the network output PS image respectively. The gradient map of the output PS image is compared to both the gradient map and the reverse gradient map of the PAN image. Then, the smaller gradient differences (in absolute value) are chosen to be included in the loss computation. The proposed dual-gradient detail loss enables the network to handle the opposite directions of gradients which frequently occur between PAN and each channel of an MS image. The loss then enforces the PS output to have similar edge details with PAN, while preserving the gradient directions as those of the color channels. This prevents the double edge artifact which happens due to the gradient direction mismatch, resulting in better visual quality.

2) COLOR LOSS

In addition to the two detail loss functions, we propose a guided-filter-based color loss to impose color similarity between the MS input and the network PS output. Here we utilize previously aligned PAN-resolution MS images $\widetilde{\mathbf{M}}_0$ as color targets to avoid any artifact that comes from the misalignment between $\mathbf{P}_0$ and $\mathbf{M}_1$. The previous deep-learning-based methods in supervised learning have used L1 or L2 loss between the network PS output $\mathbf{S}_1$ and the pseudo ground truth MS image $\mathbf{M}_1$, under the assumption that those two have similar high-frequency details and colors.

However in our unsupervised learning setting (original scale scenario), there exists no ground truth, but the network output $\mathbf{S}_0$ is supposed to have high-frequency details learned from the PAN image $\mathbf{P}_0$, where such high-frequency details are not present in the input MS image $\mathbf{M}_1$. So as to ensure that the network produces the PS output $\mathbf{S}_0$ having similar colors as the aligned PAN-resolution MS image $\widetilde{\mathbf{M}}_0$ while not losing the high-frequency information, we first apply a guided filter to the network output $\mathbf{S}_0$ using the previously aligned MS target $\widetilde{\mathbf{M}}_0$ as guidance. Then the resulting guided-filtered PS output $GF(\mathbf{S}_0)$ is compared with the aligned MS target $\widetilde{\mathbf{M}}_0$ using L1 loss. Without the guided-filtering step, this becomes a direct comparison between the network output $\mathbf{S}_0$ and the aligned MS image $\widetilde{\mathbf{M}}_0$, which would result in a substantial loss of the high-frequency details that are learned from the PAN image $\mathbf{P}_0$. Our guided-filter-based color loss is defined as

$$L_c = \sum ||GF(\mathbf{S}_0) - b(\widetilde{\mathbf{M}}_0)||_1^1 \tag{3}$$

where $GF(\mathbf{S}_0)$ is a guided filtering operation on the network PS output $\mathbf{S}_0$ with guidance $\widetilde{\mathbf{M}}_0$, and $b(\cdot)$ is a Gaussian blurring operation with the filter size of 3 with $\sigma = 2/3$. The values are set empirically to apply a mild blur as strong blur often leads to a loss of detail information. The Gaussian blur is applied to reduce the pixel blocking artifact introduced during the alignment and upscaling operation described in Sec. III-C. The proposed guided-filter-based color loss enforces the PS output to have a similar color as that of the MS image, while avoiding the checkerboard artifacts that may come from a down-sampling operation and loss of high-frequency details due to a direct comparison between PS and MS images.

3) TOTAL LOSS

The total loss function to train the network is defined as a weighted sum of the aforementioned loss functions, which is given by

$$L_{total} = L_d + w_{dg}L_{dg} + w_c L_c \tag{4}$$

where w_{dg} and w_c are empirically set to 1 and 2, respectively. Our total loss function is simple yet effective.

E. NORMALIZATION

Throughout the whole training and test processes, the inputs are normalized by the mean and standard deviation values

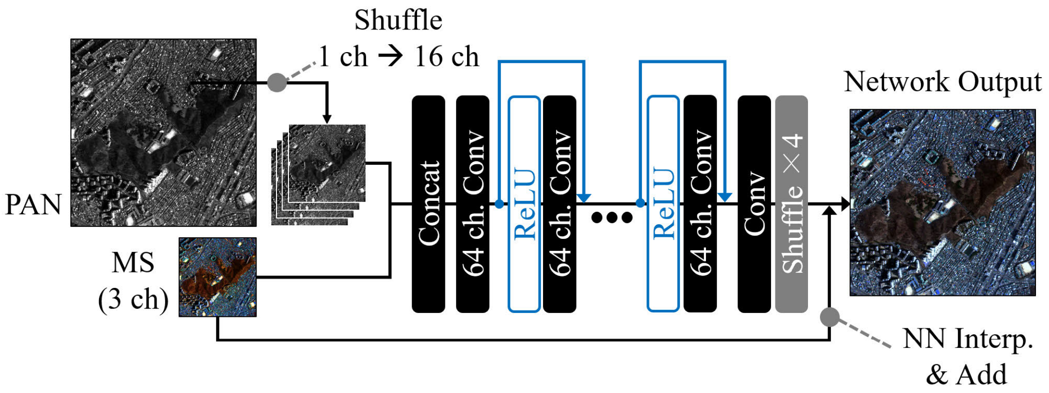


FIGURE 5. Network architecture of our proposed UPSNet.

at each pixel computed within a local patch around the pixel. We have conducted extensive experiments for various types of normalization, such as uniform normalization for all images using the dataset statistics, and global normalization by computing mean and standard deviation values for each image. But local normalization per patch has shown to be the most effective method.

As mentioned earlier, PAN and MS input images are non-stationary, having various pixel intensity distributions depending on geographical features. Also, pixel intensity distributions can be very different according to satellite sensor types. It is time-consuming and costly to train dedicated PS networks for different satellite datasets. Motivated by this, we propose a simple but effective patch-based normalization technique that allows the network trained on the images acquired by a specific satellite to be well generalized for unseen images of other satellites. Applying our normalization helps maintain the color information of the MS input.

Our proposed normalization downscales the PAN and aligned MS images to the MS scale, and computes the mean and variance values in a local window of size 9 × 9 over downscaled images for complexity reduction. Then upscaled mean and variance maps are used to normalize the PAN and MS input images. Denormalization is applied to the network output to yield the final PS result images. The values that are used for the denormalization are the upscaled mean and variance map of the MS input that were used for normalization. Local window size should be large enough to capture the regional characteristics of geographical features, as the goal of local normalization is a generalization to unseen datasets. However, the computational complexity quadratically increases as the window size goes bigger. Through a set of experiments, we found that a local window of size 9 × 9 shows good generalizability without being computationally too expensive. Our normalization technique can be easily adopted to any PS networks. This allows the network to maintain the low-frequency color information of the MS input, having a similar effect as the residual connection.

F. NETWORK ARCHITECTURE AND TRAINING DETAILS

Our network, UPSNet, comprises of 28 residual blocks, each of which has one leaky ReLU (negative slope = 0.1), one convolution layer, and one identity mapping. In total, our network has 30 convolution layers with about 1M filter parameters. To reduce the computational complexity, a single channel PAN input is de-shuffled and transformed into an MS-sized 16-channel input, which is an opposite operation of the sub-pixel convolution layer [35]. The de-shuffled PAN image is then concatenated with the 3-channel MS input. Therefore, the MS-sized 19-channel data is fed into the first convolution layer of UPSNet. The last convolution layer generates 48-channel feature maps, which is then converted to a PAN-sized 3-channel (if MS is RGB) residual output via a shuffle layer. Finally, a nearest-neighbor interpolated MS image is added to the residual output to generate the final pan-sharpened image. Fig. 5 illustrates the network structure of our UPSNet.

IV. EVALUATION

A. EXPERIMENT SETTING

1) DATASETS

We evaluate the performance of UPSNet using two different remote sensing image datasets that are captured with the WorldView-3 (WV3) and KOMPSAT-3A (K3A) satellite sensors. The WorldView-3 satellite provides 0.31m PAN resolution and 1.24m MS resolution. The KOMPSAT-3A satellite provides 0.55m PAN resolution and 2.2m MS resolution. Both sensors have a resolution ratio equal to 4 between PAN and MS images. Randomly cropped patch pairs of PAN-MS images were used for training of the networks, where various data augmentations were conducted on the fly. Each cropped MS image patches have a size of 32 × 32, while the corresponding PAN image patches have a size of 128 × 128. As mentioned earlier, the training of our UPSNet is done in the original scale scenario, while the other deep-learning-based PS methods under comparison are trained in a lower scale scenario according to their original settings in their papers. For testing, 100 PAN-MS image pairs that were unseen during training were randomly selected.

2) TRAINING

We trained UPSNet using the ADAMW [23] optimization technique with the initial learning rate of 10^{-4} and the weight decay of 10^{-7}. For training other deep-learning-based PS methods, we followed the training details provided in their original papers. We employed the uniform weight initialization technique in [14] for training. All the networks were implemented using TensorFlow [1], and were trained and tested on NVIDIA TITAN$^{\text{TM}}$ RTX GPU. Our network is trained for 10^6 iterations, where the learning rate was lowered by a factor of 10 after 5×10^5 iterations. The mini-batch size was set to 2. Training of UPSNet takes about 10 hours, and it takes 0.237 seconds for testing an image of size 648×648 (PAN) on average.

B. RESULTS AND DISCUSSIONS

1) QUANTITATIVE COMPARISON

a: PS METHODS FOR COMPARISON

We compare our UPSNet with seven non-deep-learning PS methods including Brovey transform [9], affinity PS [37], guided-filtering-based PS [39], intensity-hue-saturation (IHS) PS [5], principal component analysis (PCA) PS [34], P+XS PS [3] and variational PS [8], and five deep-learning-based PS methods including PNN [28], PanNet [43] and DSen2 [19], and their variants trained with S3 loss [6], called PanNet-S3 and DSen2-S3, respectively. UPSNet trained without the registration learning (UPSNet w/o align) is also evaluated for comparison, which is trained with bicubic interpolated original MS image for guided-filter-based color loss instead of the aligned MS image.

b: LOWER-SCALE VALIDATIONS

Due to the unavailability of ground-truth pan-sharpened images, we evaluate the performances of UPSNet and other PS methods under two different settings: lower-scale and full-scale (original-scale) validations. We use the full-reference metrics under the lower-scale validation following the Wald's protocol [40]. For this, the downscaled versions of PAN and MS images are fed as input to all the methods under comparison, and the resulting output PS images of lower-scale are compared with their corresponding pseudo-ground-truth original MS images. Four different metrics are used for the lower-scale validations: (i) spatial correlation coefficient (SCC) [47]; (ii) erreur relative globale adimensionnelle de synthèse (ERGAS) [24]; (iii) Q index [41]; and (iv) peak signal-to-noise ratio (PSNR).

c: FULL-SCALE VALIDATIONS

For the full-scale validation, SCC is also measured between original PAN inputs and grayscale versions of PS output images. The SCC values measured at full-scale indicate how much a pan-sharpening method can maintain the sharpness of the input PAN images in the PS output images. We also measure the quality-with-no-reference (QNR) [2] which is a no-reference metric for pan-sharpening, and another no-reference metric called a joint quality measure (JQM) [30] metric, which is known to better coincide with the perceived visual quality of PS output images than QNR.

d: MISALIGNMENT ISSUE BETWEEN PAN AND MS IMAGES

In general, the PAN and MS images are misaligned due to inevitable acquisition time difference and mosaicked sensor arrays. However, none of the above seven metrics for lower- and full-scale validations considers the inherent misalignment between PAN and MS images. On one hand, UPSNet is designed to correct the inherent misalignment between them by aligning the color (MS) of an object with the objects' details (PAN). So, it can produce output PS images that have very well aligned colors and shapes of objects. In this case, it is important to note that directly measuring the spectral distortion of the PS output with respect to the color of the original MS input is meaningless for the aligned PS output. This is because the colors of the PS output generated by UPSNet are moved (aligned) to match the shapes (details). Therefore, in addition to such conventional direct measures with respect to the original MS inputs, we also measure the distortions with respect to the aligned MS images created by the alignment method in Section III-C for fair and meaningful comparison.

e: ANALYSIS FOR EXPERIMENTAL RESULTS

Tables 1 and 2 show the average metric scores for 100 randomly chosen test image pairs from the WorldView-3 dataset measured with respect to the original MS input without alignment and with the aligned MS image, respectively. ↑ and ↓ indicate that the higher the better, and the lower the better performance, respectively for each metric. In Table 1, UPSNet (w/o align) outperforms all other methods for all lower-scale validations and JQM when measured with the original MS input. When measuring full-scale SCC metric for original scale validation, the SCC values in Table 1 are the same as those in Table 2. This is because MS images are not used in measuring the SCC metric, as mentioned earlier. As shown in Tables 1 and 2, UPSNet performs the best in terms of SCC. DSen2 shows the highest QNR value in Table 2, however it shows poor perceived visual quality, which will be later discussed in Sec. IV-B3. In Table 2, UPSNet outperforms all other PS methods in terms of all quality metrics except QNR, and UPSNet (w/o align) achieves the highest value of QNR.

2) QUALITATIVE COMPARISON

Fig. 6 and 7 show visual comparisons for our UPSNet against the previous state-of-the-art methods. It is clearly shown in Fig. 6-(p) and 7-(p) that the PS output image from UPSNet well preserves the high-frequency details of PAN inputs and the color information as similar as possible with MS inputs, also having minimal distortions. The effectiveness of registration (alignment) learning by our UPSNet can be clearly seen around the pool area in Fig. 6-(h). Since the pool is located at a slightly up-right position in the MS image (Fig. 6-(b)) compared to the PAN image (Fig. 6-(a)), most of the previous

TABLE 1. Quantitative comparison (measured with original MS input without alignment).

	Lower-scale				Full-scale		
	SCC ↑	ERGAS ↓	Q ↑	PSNR ↑	SCC ↑	QNR ↑	JQM ↑
Brovey [9]	0.836	7.190	0.726	26.098	0.952	0.720	0.855
Affinity [37]	0.791	3.993	0.662	32.207	0.747	0.739	0.762
GF [39]	0.801	4.005	0.656	32.175	0.778	0.741	0.756
IHS [5]	0.836	3.877	0.722	32.676	0.958	0.790	0.855
PCA [34]	0.817	5.060	0.579	30.41	0.899	0.768	0.832
P+XS [3]	0.774	4.459	0.677	30.366	0.861	0.834	0.857
Variational [8]	0.781	3.638	0.768	33.415	0.665	0.825	0.819
PNN [28]	0.694	5.694	0.476	29.594	0.697	0.711	0.733
PanNet [43]	0.802	4.048	0.737	32.371	0.811	0.850	0.833
PanNet-S3 [6]	0.838	3.816	0.758	32.767	0.954	0.782	0.888
DSen2 [19]	0.820	3.765	0.753	32.931	0.837	**0.883**	0.853
DSen2-S3 [6]	0.844	5.330	0.739	29.39	0.959	0.759	0.875
UPSNet	**0.845**	3.536	0.786	33.582	**0.960**	0.783	0.887
UPSNet (w/o align)	**0.845**	**3.523**	**0.787**	**33.603**	0.937	0.827	**0.907**

TABLE 2. Quantitative comparison (measured with aligned MS image created by the alignment method in Section III-C).

	Lower-scale				Full-scale		
	SCC ↑	ERGAS ↓	Q ↑	PSNR ↑	SCC ↑	QNR ↑	JQM ↑
Brovey [9]	0.932	6.448	0.792	27.14	0.952	0.779	0.885
Affinity [37]	0.873	3.103	0.729	34.253	0.747	0.677	0.759
GF [39]	0.886	3.078	0.726	34.293	0.778	0.681	0.762
IHS [5]	0.928	2.808	0.813	35.419	0.958	0.852	0.883
PCA [34]	0.908	4.311	0.669	31.829	0.899	0.744	0.845
P+XS [3]	0.858	3.585	0.729	32.066	0.861	0.851	0.863
Variational [8]	0.839	2.931	0.831	35.379	0.665	0.755	0.759
PNN [28]	0.762	5.110	0.538	30.612	0.697	0.649	0.752
PanNet [43]	0.877	3.316	0.800	34.109	0.811	0.806	0.802
PanNet-S3 [6]	0.933	2.725	0.833	35.450	0.954	0.843	0.916
DSen2 [19]	0.900	2.933	0.826	35.045	0.837	0.852	0.827
DSen2-S3 [6]	0.942	4.391	0.821	31.211	0.959	0.807	0.909
UPSNet	**0.943**	**2.293**	**0.871**	**37.064**	**0.960**	0.848	**0.929**
UPSNet (w/o align)	0.942	2.354	0.868	36.849	0.937	**0.867**	0.889

SOTA PS methods show artifacts (color of the water is placed at slightly up-right position compared to the shape of the pool) due to this misalignment, but the output PS image of UPSNet shows no such artifacts. Also, UPSNet produces the most similar color with the original MS images, especially the color of the water in the pool (Fig. 6-(h)). The effectiveness of the registration learning is even more emphasized in Fig. 7-(h). As can be seen in Fig. 7-(a) and (b), the color of the orange roof in the MS image is placed slightly upward compared to the shape of that in the PAN image. UPSNet is the only method that is able to fuse the colors of the orange roof from the MS image with their appropriate shapes in the corresponding PAN image. More visual comparisons are provided in Figs. 13 and 14.

3) CONSIDERATIONS FOR NO-REFERENCE METRICS: QNR AND JQM

In this paper, we have utilized two full-scale no-reference metrics, QNR and JQM. However, several previous works have pointed out the drawbacks and unexpected properties of QNR [16], [30], [40], 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 it can be expected that the values of the QNR metric are not well agreed with the observed visual quality.

We have intensively investigated this discrepancy between QNR metric and subjective quality for PS output. Figs. 8 and 9 show visual comparison on PS outputs obtained by various pan-sharpening methods. As shown, it is important to note that, although the PS output images of PNN, PanNet and Dsen2 relatively exhibit higher QNR scores than those of PanNet-S3, DSen2-S3 and UPSNet, their perceived visual qualities are much worse, showing severe ghost artifacts in Fig. 8 and the misalignment between colors and shapes (details) in Fig. 9. It is also worthwhile to point out that the PS output of UPSNet in Fig. 9 shows the best visual quality but has the lowest QNR value.

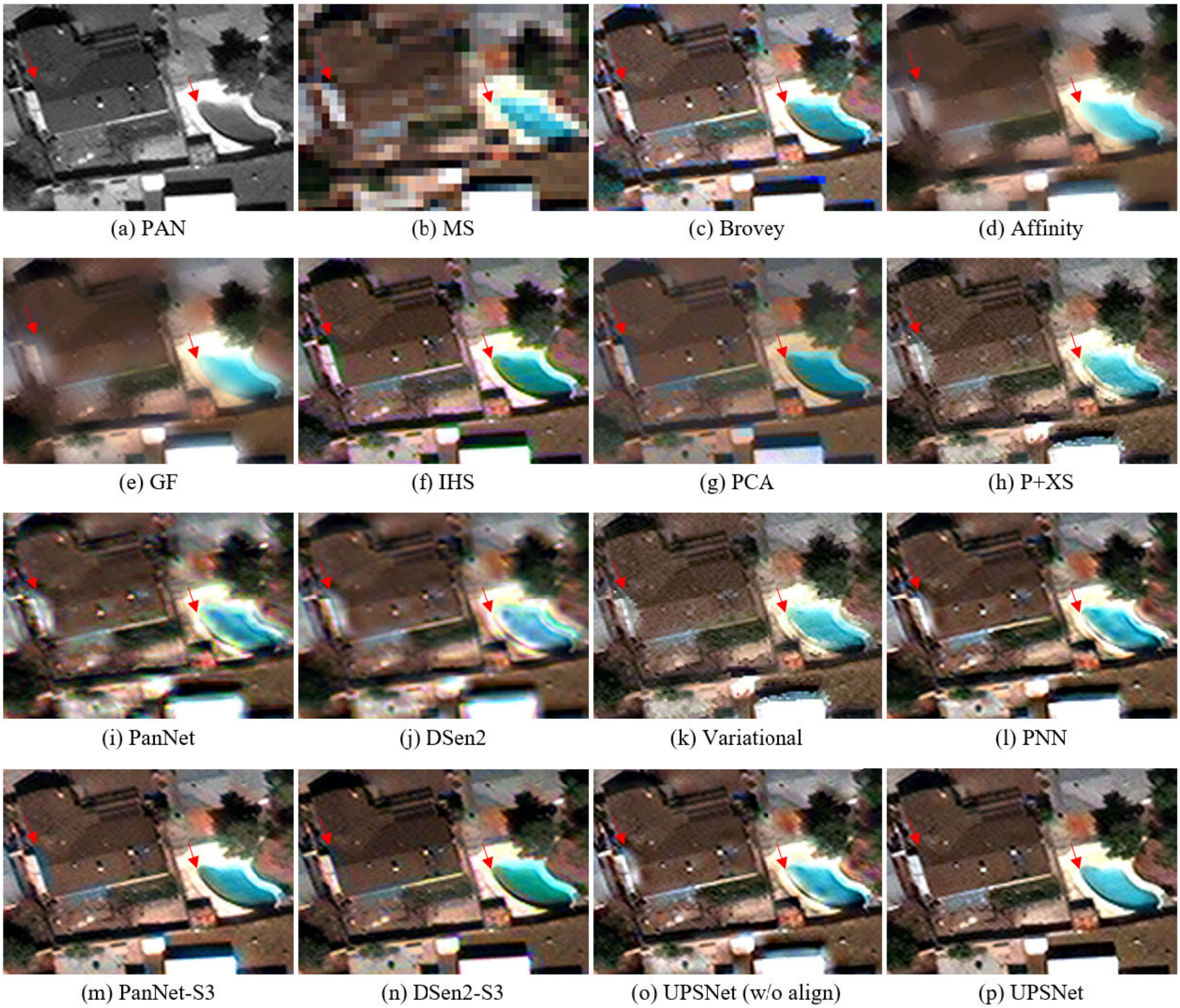


FIGURE 6. Result images for pan-sharpening using various methods and our UPSNet.

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 [32]. As shown in Figs. 8 and 9, 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. As opposed to the QNR metric, PNN, PanNet and Dsen2 relatively exhibit lower JQM scores than those of PanNet-S3, DSen2-S3 and UPSNet in Figs. 8 and 9. In both figures, PS outputs from our UPSNet yield the highest JQM scores, coinciding with the perceived visual quality. The visual qualities of the PS outputs produced by DSen2-S3 and PanNet-S3 are ranked the second and the third in terms of JQM values, which are very reasonably ranked in agreement with the perceived visual qualities.

The discrepancy between QNR and perceived visual quality comes from the fact that QNR does not directly reflect the spectral and spatial distortions in its calculation form [2]. The spectral distortion term (D_λ) of QNR indirectly obtains the spectral distortion index by taking the difference between inter-band similarity measures of the MS and PS images. Similarly, the spatial distortion term (D_S) of QNR is measured indirectly 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. On the other hand, JQM [30] directly measures both the spectral distortion between MS and downscaled PS images, and the spatial distortion between PAN and fused PS images. The JQM was argued that it is better agreed with perceived visual quality than QNR [30]. Throughout our intensive experiments, we also have found that the JQM is better correlated with perceived visual quality for various PS output images, as shown in Figs. 8 and 9.

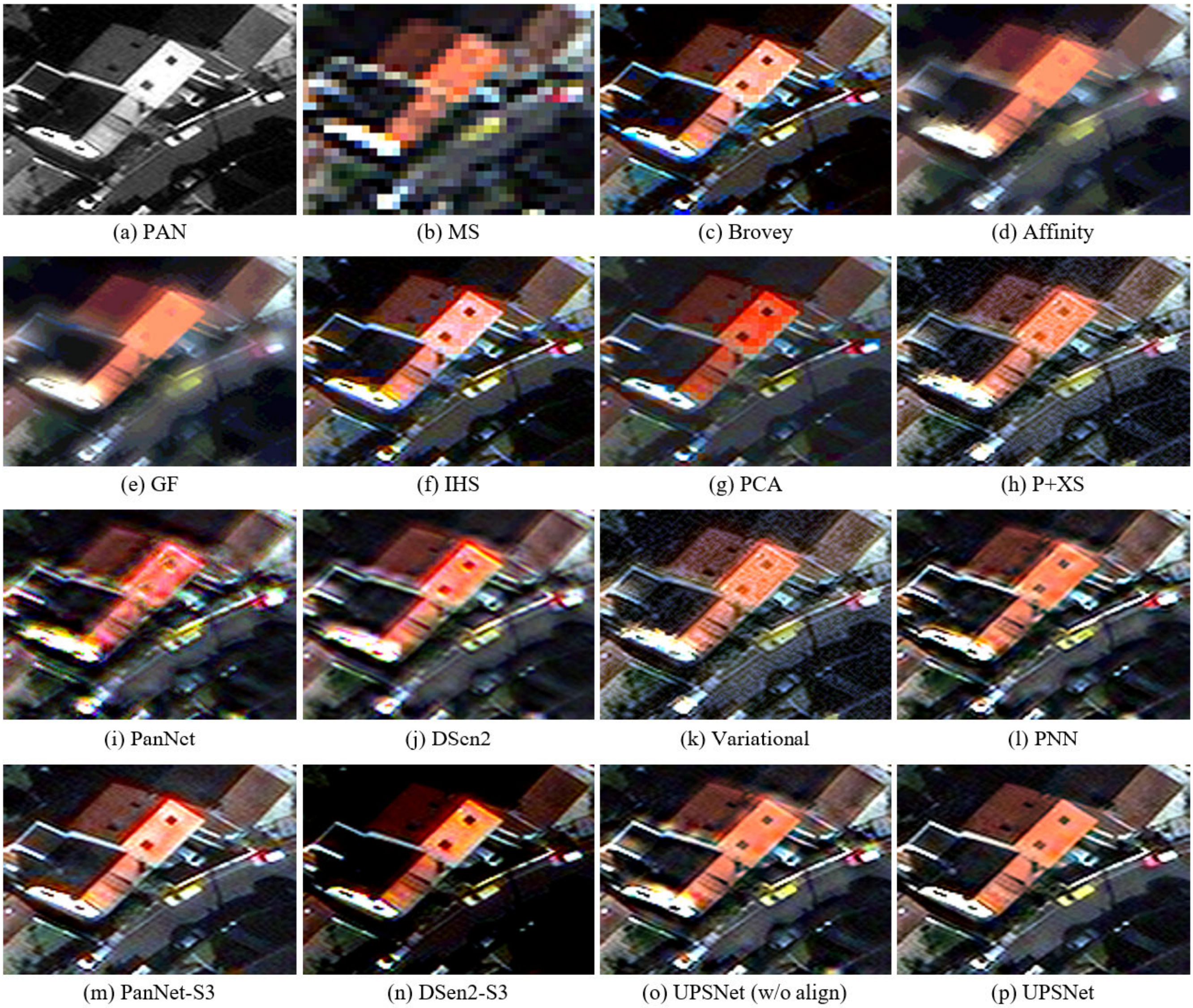

(a) PAN (b) MS (c) Brovey (d) Affinity
(e) GF (f) IHS (g) PCA (h) P+XS
(i) PanNet (j) DSen2 (k) Variational (l) PNN
(m) PanNet-S3 (n) DSen2-S3 (o) UPSNet (w/o align) (p) UPSNet

FIGURE 7. Result images for pan-sharpening using various methods and our UPSNet.

C. ABLATION STUDIES

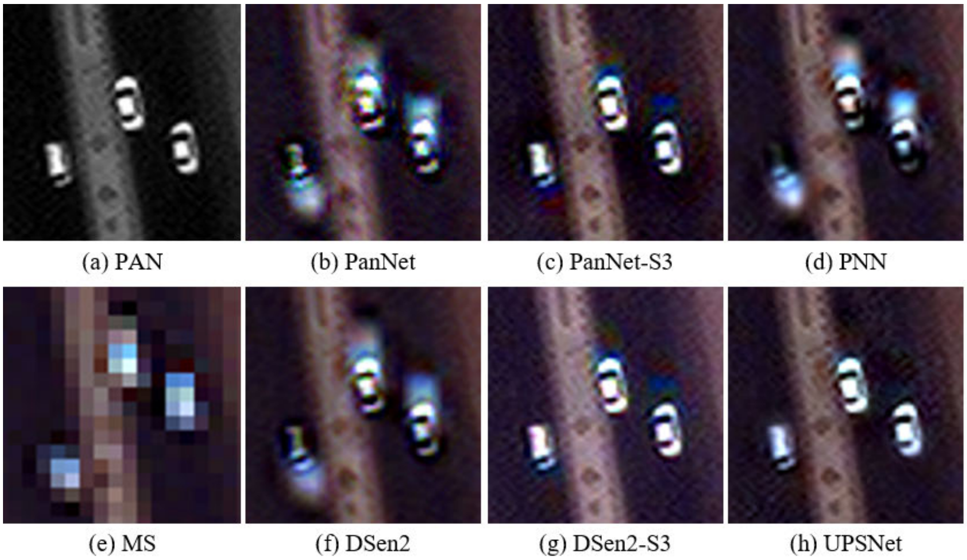
Ablation studies have been conducted in a few different settings to show the effectiveness of key aspects of our proposed UPSNet. Throughout the experiments, only one component has been changed, and others remained the same. Evaluation of different models has been conducted under full-scale, using original MS and PAN input as inputs for the network. We measure two different criteria for measuring the performance of output PS images: high-frequency detail similarity with PAN images (SCC) and color similarity with MS images (ERGAS). ERGAS is measured between aligned MS images and PS output images. We denote this as ERGAS-A.

1) LEARNING FRAMEWORK

First, we provide ablation study results on learning framework including unsupervised learning, training in original scales, and alignment. Experiment conditions are as follows.

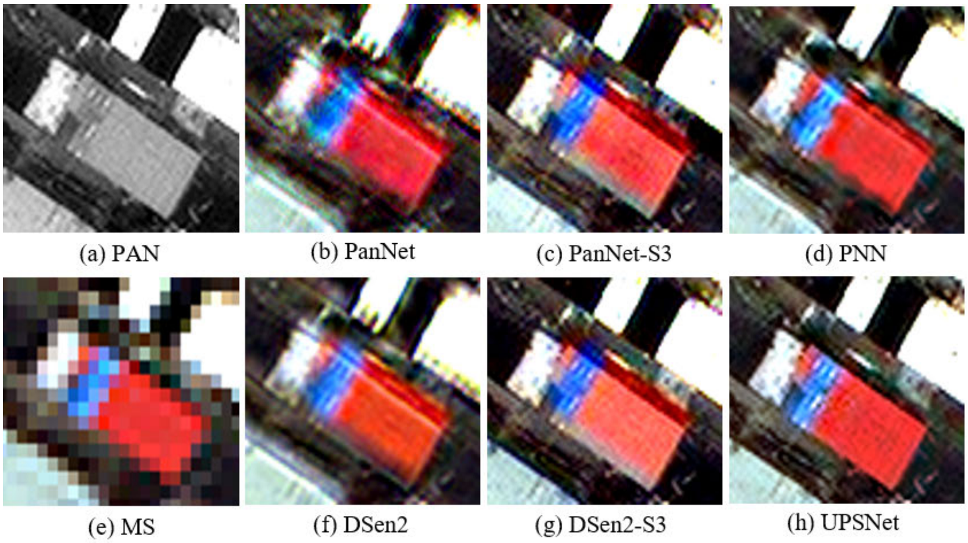
Condition 1 is for training on lower scales using our unsupervised framework and testing on original scales. *Condition* 2 is for training without alignment, using the bicubic interpolated original MS image as a target for the color loss. In *Condition* 3, we train UPSNet in a supervised manner, similarly to PanNet [43] and DSen2 [19], where each training pair of PAN and MS images is downscaled by a scale of 4, and the original MS input is used as a pseudo ground truth. The network for *Condition* 3 is regularized by L1 loss between output PS images and original MS inputs to have similar settings as PanNet [43] and DSen2 [19].

As shown in Table. 3, all conditions entail substantial performance drops in terms of all metrics. Fig. 10 shows the visual comparison for *Conditions* 1, 2, and 3. Due to the scale mismatch between training and testing, and the absence of alignment between MS and PAN images, it is clear that the results in Fig. 10-(b), (c), and (d) suffer from misaligned



	QNR ↑	JQM ↑
PNN [30]	0.918	0.735
PanNet [46]	0.909	0.774
DSen2 [21]	**0.924**	0.797
PanNet-S3 [7]	0.869	0.920
DSen2-S3 [7]	0.811	0.934
UPSNet	0.895	**0.939**

FIGURE 8. Visual comparison of various pan-sharpening methods including their QNR and JQM values.



	QNR ↑	JQM ↑
PNN [30]	**0.885**	0.773
PanNet [46]	0.770	0.722
DSen2 [21]	0.846	0.758
PanNet-S3 [7]	0.755	0.887
DSen2-S3 [7]	0.740	0.886
UPSNet	0.670	**0.919**

FIGURE 9. Visual comparison of various pan-sharpening methods including their QNR and JQM values.

TABLE 3. Performance of UPSNet under different settings of learning at original scales, at lower scales, without alignment, and in an supervised manner.

Methods	Conditions	ERGAS-A↓	SCC↑
ours	-	**1.746**	**0.960**
lower scale	1	2.266	0.944
w/o alignment	2	2.476	0.931
supervised	3	2.718	0.538

colors, especially on the areas pointed by the red arrows. As can be seen in Table. 3, UPSNet trained in a supervised manner has shown a substantial amount of performance drop, especially in terms of SCC. Fig. 10-(d) clearly shows that supervised training in lower scales causes inferior visual quality, also showing artifacts in the homogeneous region.

2) REGISTRATION SCALE

In Sec. III-C, we have discussed other possible alignment options to perform the registration step in the MS scale. The aligned MS, the output of the registration step, is only used as a target for the proposed guided-filter-based color loss and has the same size as the PAN image, as explained in Sec. III-C. Then, the PS output images from UPSNet and their corresponding aligned MS images are compared by the guided-filter-based color loss function without any scale conversion. However, when the registration is performed in the MS scale, aligned MS images would have the same size as input MS images. Therefore there exists a scale mismatch between the PS images and the corresponding aligned MS images. In this case, the computation of color

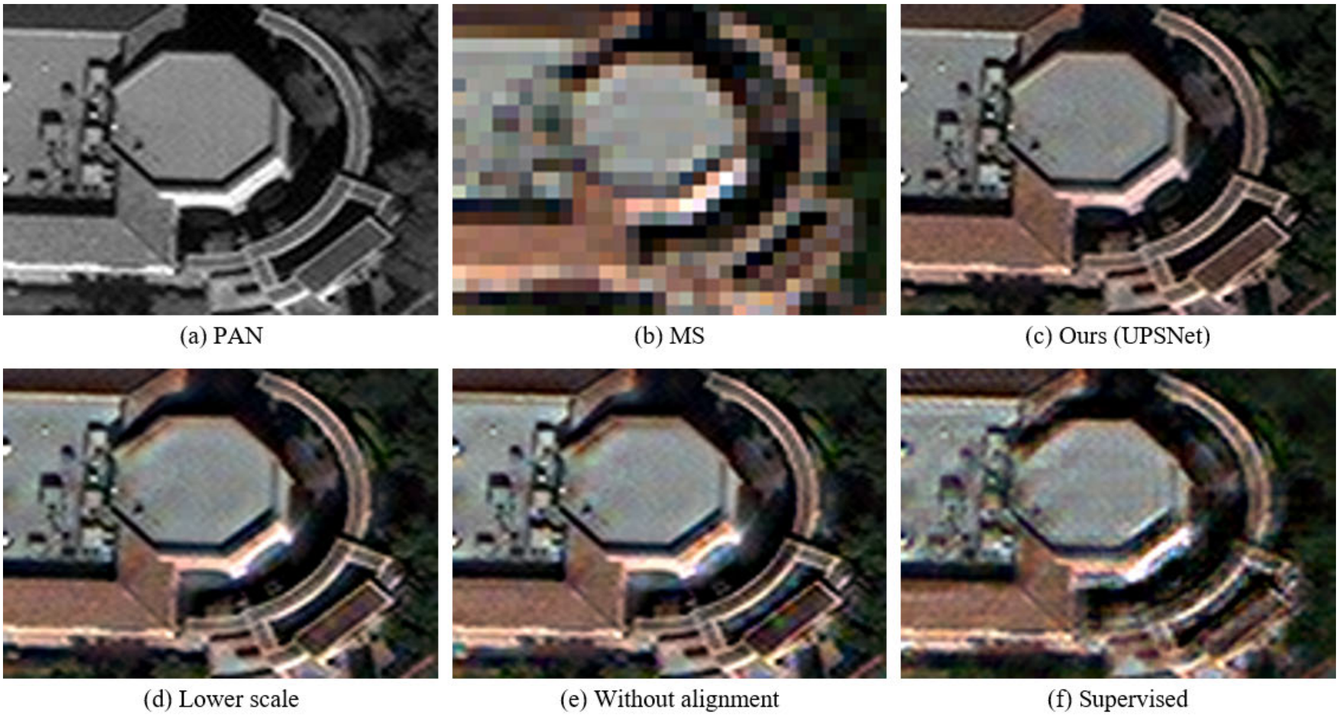

(a) PAN
(b) MS
(c) Ours (UPSNet)
(d) Lower scale
(e) Without alignment
(f) Supervised

FIGURE 10. Visual comparison for ablation study.

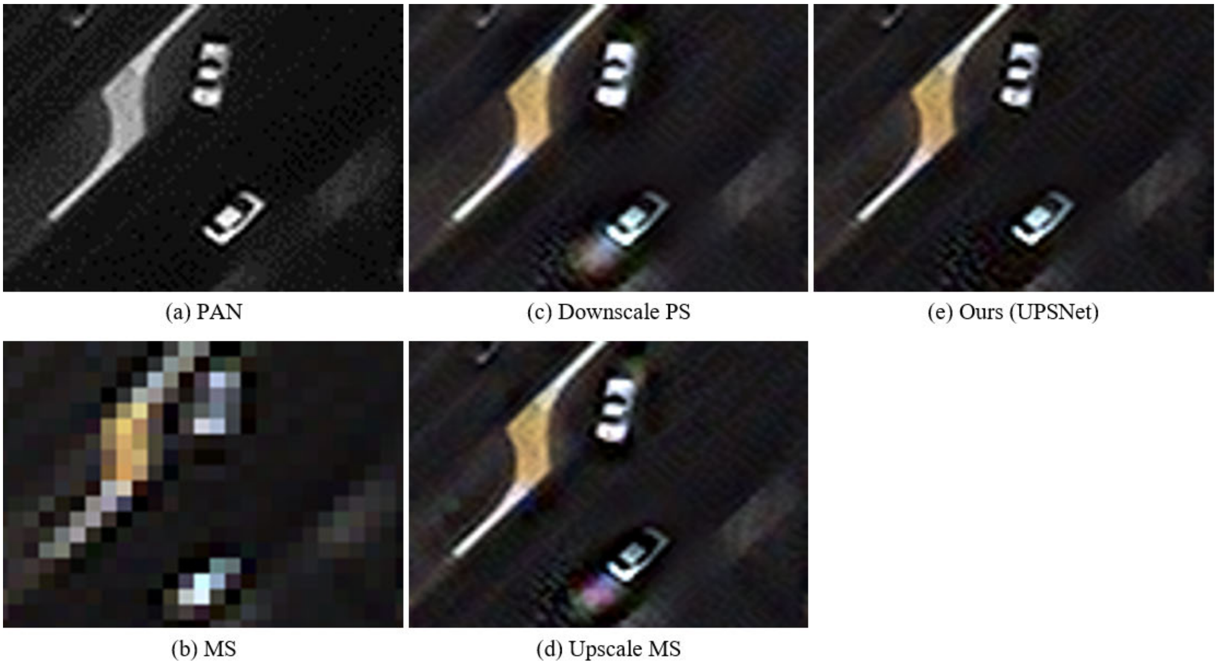

(a) PAN
(c) Downscale PS
(e) Ours (UPSNet)
(b) MS
(d) Upscale MS

FIGURE 11. Qualitative comparison between our proposed UPSNet and its variants trained with registration in MS scale for a cropped region of an image 'AOI_2_Vegas_Roads_Test_public_img161.tif' in WorldView-3 dataset.

loss can have two possible options for matching the scale (resolution).

The first option is to downscale the PS images to the MS scale by applying a degradation model, and the second option is to upscale the aligned MS images (aligned in the MS scale) to have the same resolution as their corresponding PS images. Since both options require scale conversion, a new type of misalignment is introduced inevitably during the scale matching process.

Table 4 provides the quantitative experiment results for UPSNet and its variants trained under two options mentioned above. The values of ERGAS-A and SCC metrics

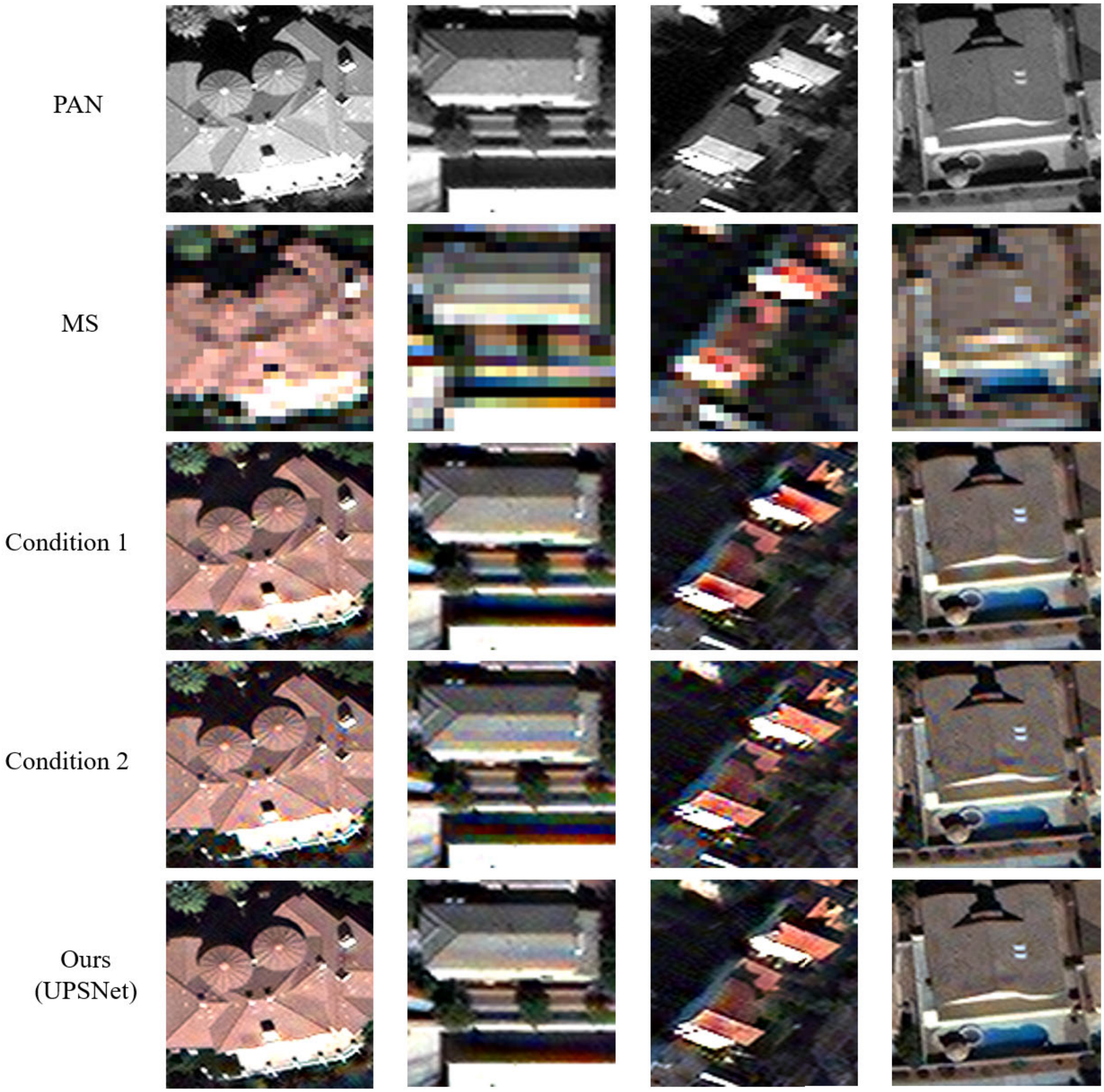


FIGURE 12. Qualitative comparison for ablation study on loss functions.

are lowered in the two options. Figs. 11 shows the artifacts introduced by the scale conversion. UPSNet can effectively handle the misalignment between the PAN and MS images, especially on the moving cars, but variants of UPSNet that included scale conversion (Fig. 11-(c), (d)) failed because they could not properly learn to handle the misalignment. The overall experiment results show that registration in the PAN scale yields the best pan-sharpening performance in both quantitative and qualitative perspectives.

3) LOSS FUNCTIONS

In this section, we discuss the effectiveness of the proposed loss functions. Two loss functions have been newly proposed to train our UPSNet: a guided-filter-based color loss (L_c) between network outputs and our aligned MS targets; and a dual-gradient detail loss (L_{dg}) between network outputs and PAN inputs.

Ablation studies have been conducted under two different conditions to show the effectiveness of the proposed loss functions. *Condition* 1 is training the network without the

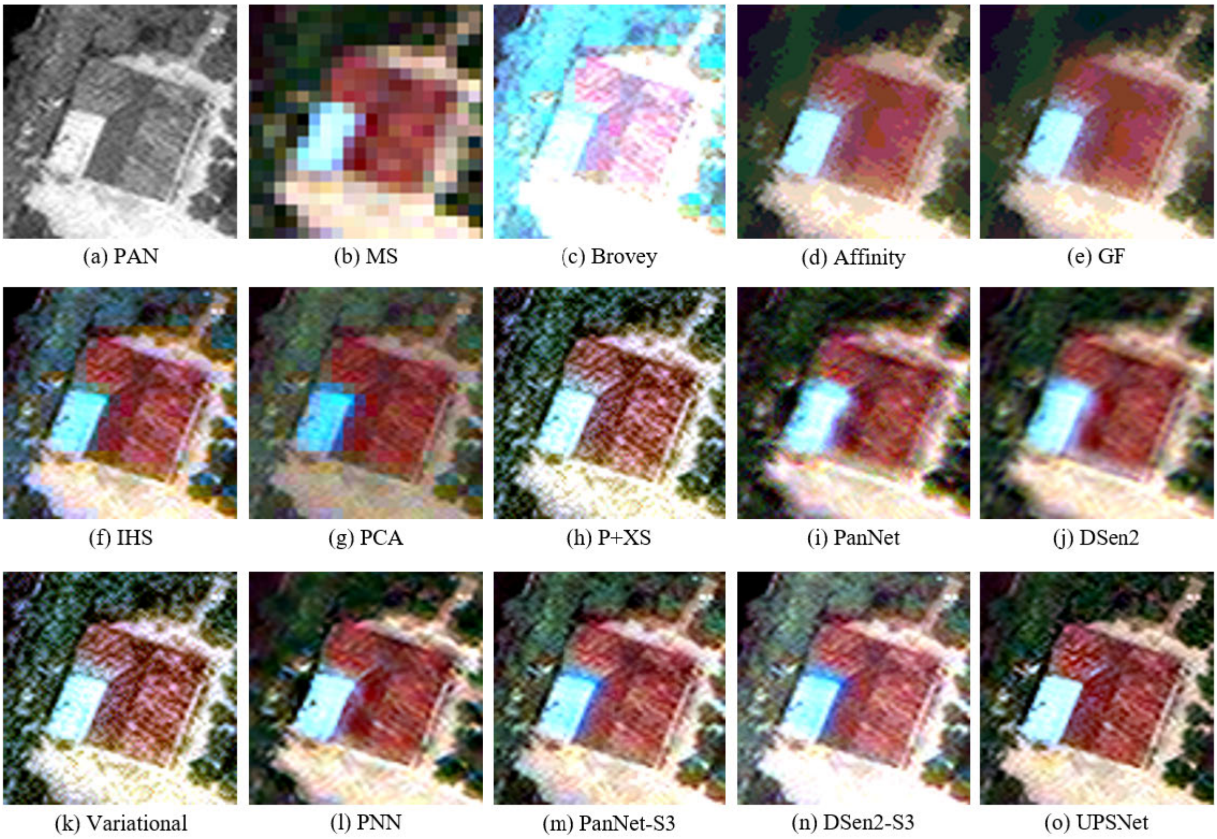

(a) PAN (b) MS (c) Brovey (d) Affinity (e) GF
(f) IHS (g) PCA (h) P+XS (i) PanNet (j) DSen2
(k) Variational (l) PNN (m) PanNet-S3 (n) DSen2-S3 (o) UPSNet

FIGURE 13. Qualitative comparison between our proposed UPSNet and other SOTA methods for a cropped region of an image 'AOI_3_Shanghai_Bldg_Test_public_img2434.tif' in the WorldView-3 dataset.

TABLE 4. Ablation study on registration scale.

Methods	Conditions	ERGAS-A↓	SCC↑
UPSNet (Ours)	-	**1.746**	**0.960**
Downscale PS	1	2.470	0.955
Upscale MS	2	1.968	0.948

TABLE 5. Ablation study on loss functions.

Conditions	Loss function	ERGAS-A↓	SCC↑
UPSNet (Ours)	$L_d + L_{dg} + 2 \times L_c$	**1.746**	**0.960**
1	$2 \times L_d + 2 \times L_c$	1.816	0.939
2	$L_d + L_{dg} + 2 \times L_{gb}$	2.016	0.954

TABLE 6. Evaluation of PS networks (Train: WorldView-3, Test: WorldView-3).

Methods	Trained on	Tested on	ERGAS-A↓	SCC↑
PanNet [43]	WV3	WV3	2.6856	0.7754
PanNet_S3 [6]	WV3	WV3	2.2329	0.9507
DSen2 [19]	WV3	WV3	2.3648	0.8543
DSen2_S3 [6]	WV3	WV3	4.4001	0.9588
UPSNet (Ours)	WV3	WV3	**1.7459**	**0.9597**

TABLE 7. Evaluation of PS networks (Train: K3A, Test: WorldView-3).

Methods	Trained on	Tested on	ERGAS-A↓	SCC↑
PanNet [43]	K3A	WV3	2.7914	0.5386
PanNet_S3 [6]	K3A	WV3	4.5152	0.9478
DSen2 [19]	K3A	WV3	2.8412	0.5543
DSen2_S3 [6]	K3A	WV3	9.9416	0.9195
UPSNet (Ours)	K3A	WV3	**1.826**	**0.958**

dual-gradient detail loss. *Condition* 2 is applying Gaussian blur kernel instead of the guided-filter used for the color loss. We denote the Gaussian blur kernel-based color loss as L_{gb}. The parameters of a Gaussian blur kernel are adequately adjusted so that the PS images after applying the Gaussian blur kernel to have similar visual quality with the corresponding aligned MS images.

Table 5 shows the average metric scores of ERGAS-A and SCC. The performance drops are observed for both *Condition* 1 and *Condition* 2, showing that the proposed loss functions are essential for training our UPSNet. Fig. 12 shows the visual comparison regarding the ablation study on loss functions. *Condition* 1 seems to produce reasonable visual quality, but it tends to disturbingly enhance the local contrast, introducing some artifacts in the output PS images. *Condition* 2 introduce rainbow-like artifacts in all images. Both quantitative and qualitative experiments show the effectiveness of our proposed loss functions in training UPSNet.

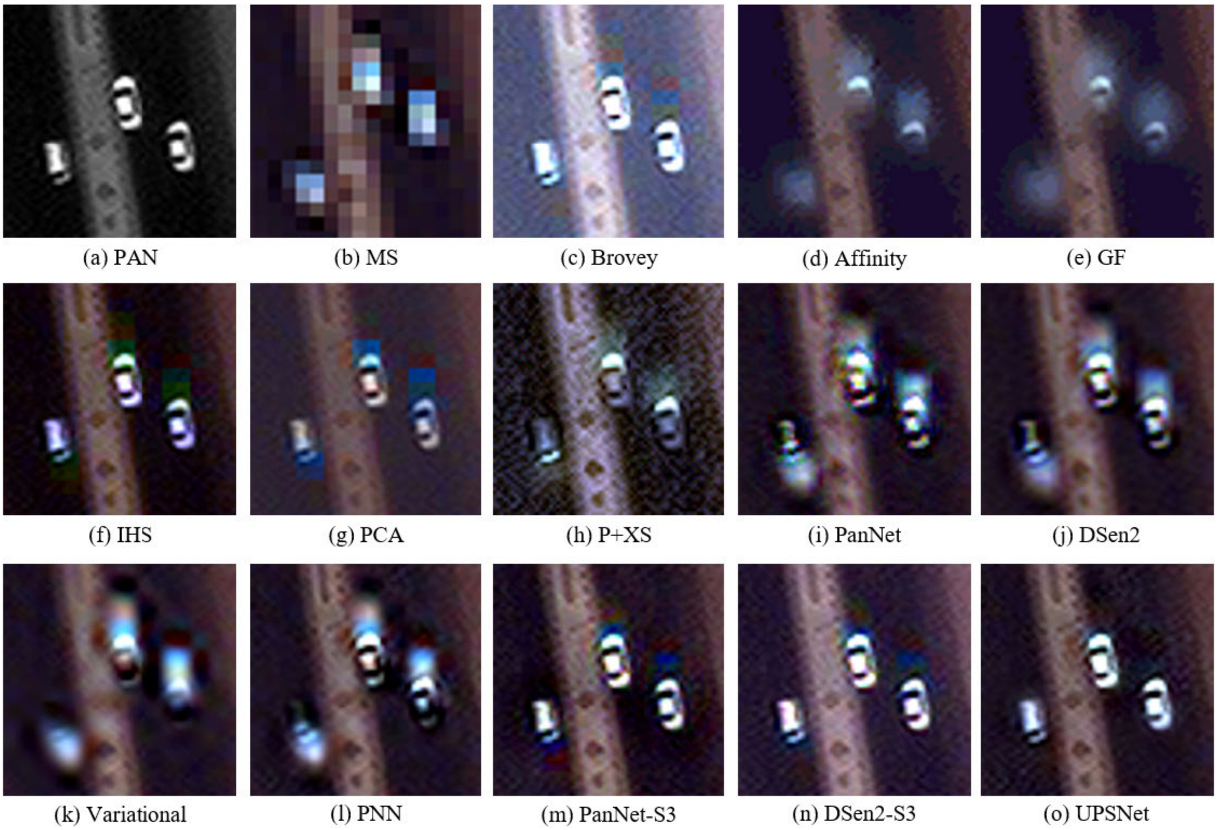


FIGURE 14. Qualitative comparison between our proposed UPSNet and other SOTA methods for a cropped region of an image 'AOI_2_Khartoum_Bldg_Test_public_img1522.tif' in the WorldView-3 dataset.

TABLE 8. Evaluation of PS networks (Train: K3A, Test: K3A).

Methods	Trained on	Tested on	ERGAS-A↓	SCC↑
PanNet [43]	K3A	K3A	1.077	0.545
PanNet_S3 [6]	K3A	K3A	1.498	**0.937**
DSen2 [19]	K3A	K3A	1.031	0.561
DSen2_S3 [6]	K3A	K3A	4.663	0.899
UPSNet (Ours)	K3A	K3A	**0.561**	0.921

TABLE 9. Evaluation of PS networks (Train: WorldView-3, Test: K3A).

Methods	Trained on	Tested on	ERGAS -A↓	SCC↑
PanNet [43]	WV3	K3A	1.022	0.6595
PanNet_S3 [6]	WV3	K3A	0.97	0.912
DSen2 [19]	WV3	K3A	3.407	0.783
DSen2_S3 [6]	WV3	K3A	12.54	**0.944**
UPSNet (Ours)	WV3	K3A	**0.642**	0.929

4) CROSS-DATASET EXPERIMENT

Cross-dataset experiments have been conducted to show the generalization capability of our UPSNet. Each pan-sharpening network is trained and tested in four different settings using the datasets acquired from two different satellites, KOMPSAT-3A (K3A) and WorldView-3, as described in Tables. 6, 7, 8, and 9. The upward and downward arrows ↑↓ indicate that higher and lower values imply better performance, respectively. The best and second-best results are highlighted in bold and underline, respectively. It can be seen that UPSNet is showing a good generalization capability while other methods show performance drop when tested on the dataset that is different from the training dataset.

V. CONCLUSION

In this work, we propose an effective unsupervised learning framework with registration learning for pan-sharpening, called UPSNet. To resolve a misalignment between PAN and MS, we first propose a simple PAN-MS registration based on correlations to obtain an aligned MS target of PAN-resolution from each misaligned PAN-MS input pair. The aligned MS target is then used to enforce the network to learn how to handle the misalignment between PAN and MS images by giving it as a target for the color loss. It should be noted that the registration for training is no longer required in testing. Additionally, we designed two loss functions for the training of our network: a guided-filter-based color loss between the network's PS outputs and our aligned MS targets; and a dual-gradient detail loss between the network's PS outputs and PAN inputs. Intensive experimental results show that our UPSNet can generate pan-sharpened images with remarkable improvements in terms of color similarity and texture details compared to the state-of-the-art pan-sharpening methods.

REFERENCES

[1] M. Abadi, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th Symp. Oper. Syst. Des. Implement.*, 2016, pp. 265–283.

[2] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008.

[3] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for P+XS image fusion," *Int. J. Comput. Vis.*, vol. 69, no. 1, pp. 43–58, Aug. 2006.

[4] M. Bosch, C. M. Gifford, and P. A. Rodriguez, "Super-resolution for overhead imagery using DenseNets and adversarial learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1414–1422.

[5] W. J. Carper, T. M. Lillesand, and R. W. Kiefer, "The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data," *Photogramm. Eng. Remote Sens.*, vol. 56, no. 4, pp. 459–467, Apr. 1990.

[6] J.-S. Choi, Y. Kim, and M. Kim, "S3: A spectral-spatial structure loss for pan-sharpening networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 5, pp. 829–833, May 2020.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[8] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, p. 10.

[9] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. I. decorrelation and HSI contrast stretches," *Remote Sens. Environ.*, vol. 20, no. 3, pp. 209–235, Dec. 1986.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[11] X. He, L. Condat, J. M. Bioucas-Dias, J. Chanussot, and J. Xia, "A new pansharpening method based on spatial and spectral sparsity priors," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4160–4174, Sep. 2014.

[12] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1037–1041, May 2015.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[14] 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 *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[15] K. Jiang, Z. Wang, P. Yi, and J. Jiang, "A progressively enhanced network for video satellite imagery superresolution," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1630–1634, Nov. 2018.

[16] M. V. Joshi and K. P. Upla, *Multi-Resolution Image Fusion Remote Sensing*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[17] X. Kang, S. Li, and J. A. Benediktsson, "Pansharpening with matting model," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5088–5099, Aug. 2014.

[18] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[19] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 305–319, Dec. 2018.

[20] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[21] L. Liebel and M. Körner, "Single-image super resolution for multispectral remote sensing data using convolutional neural networks," *ISPRS Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. XLI–B3, pp. 883–890, Jun. 2016.

[22] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*. [Online]. Available: http://arxiv.org/abs/1711.05101

[24] L. Wald, *Data Fusion: Definitions and Architectures—Fusion of Images of Different Spatial Resolutions*. Paris, France: Les Presses de l'École des Mines, 2002.

[25] S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via unsupervised convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4295–4310, 2020.

[26] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Inf. Fusion*, vol. 62, pp. 110–120, Oct. 2020.

[27] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[28] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.

[29] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpening algorithm based on total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 318–322, Jan. 2014.

[30] G. Palubinskas, "Joint quality measure for evaluation of pansharpening accuracy," *Remote Sens.*, vol. 7, no. 7, pp. 9292–9310, Jul. 2015.

[31] Z. Pan, J. Yu, H. Huang, S. Hu, A. Zhang, H. Ma, and W. Sun, "Super-resolution based on compressive sensing and structural self-similarity for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4864–4876, Sep. 2013.

[32] Y. Qu, R. K. Baghbaderani, H. Qi, and C. Kwan, "Unsupervised pansharpening based on self-attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 23, 2020, doi: 10.1109/TGRS.2020.3009207.

[33] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.

[34] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1323–1335, May 2008.

[35] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[36] J.-L. Starck, J. Fadili, and F. Murtagh, "The undecimated wavelet decomposition and its reconstruction," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 297–309, Feb. 2007.

[37] S. Tierney, J. Gao, and Y. Guo, "Affinity pansharpening and image fusion," in *Proc. Int. Conf. Digit. Image Comput.*, Nov. 2014, pp. 1–8.

[38] C. Tuna, G. Unal, and E. Sertel, "Single-frame super resolution of remote-sensing images by convolutional neural networks," *Int. J. Remote Sens.*, vol. 39, no. 8, pp. 2463–2479, Apr. 2018.

[39] K. P. Upla, S. Joshi, M. V. Joshi, and P. P. Gajjar, "Multiresolution image fusion using edge-preserving filters," *J. Appl. Remote Sens.*, vol. 9, no. 1, Jul. 2015, Art. no. 096025.

[40] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[41] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[42] Q. Xu, B. Li, Y. Zhang, and L. Ding, "High-fidelity component substitution pansharpening by the fitting of substitution data," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7380–7392, Nov. 2014.

[43] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5449–5457.

[44] W. Yao, Z. Zeng, C. Lian, and H. Tang, "Pixel-wise regression using U-net and its application on pansharpening," *Neurocomputing*, vol. 312, pp. 364–371, Oct. 2018.

[45] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5549–5563, Aug. 2019.

[46] C. Zhou, J. Zhang, J. Liu, C. Zhang, R. Fei, and S. Xu, "PercepPan: Towards unsupervised pan-sharpening based on perceptual loss," *Remote Sens.*, vol. 12, no. 14, p. 2318, Jul. 2020.

[47] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, Jan. 1998.

SOOMIN SEO received the B.S. and M.S. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2017 and 2019, respectively, where she is currently pursuing the Ph.D. degree. Her research interests include low-level computer vision, such as pan-sharpening, super-resolution, image restoration, image quality assessment, and deep learning.

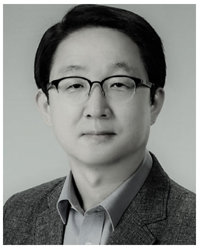
JAE-SEOK CHOI received the B.S. degree in electrical engineering from Hanyang University, Seoul, South Korea, in 2014, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2016 and 2020, respectively. He is currently a Staff Researcher with the Samsung Advanced Institute of Technology (SAIT), Suwon, South Korea. His research interests include image processing and deep learning.

JAEHYUP LEE received the B.S. and M.S. degrees from the School of Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree. His research interests include pan-sharpening, object detection, deep learning, and image processing.

HYUN-HO KIM received the B.S. degree in electrical engineering from Kyunghee University, Suwon, South Korea, in 2016, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2018. He is currently with the Korea Aerospace Research Institute. His research interests include image processing, pattern recognition, and deep/machine learning.

DOOCHUN SEO received the Ph.D. degree in civil engineering from Gyeongsang National University, South Korea, in 2002. He has been with the Korea Aerospace Research Institute since 2002, and has also been a Senior Researcher with the Satellite Cal/Val Department since 2005. His primary research interests and back ground include satellite photogrammetry, sensor modeling, DEM, Ortho-image generation, automated feature extraction from imagery, and geometric calibration of high-resolution satellite image data.

JAEHEON JEONG received the M.S. degree in information communications engineering from Chungnam National University, South Korea, in 2013. He has been with the Korea Aerospace Research Institute since 2014, and has also been a Senior Researcher with the Satellite Cal/Val Department since 2014. His primary research interests and back ground include image processing, computer vision, high performance computing, and automated feature extraction of high-resolution satellite image data.



MUNCHURL KIM (Senior Member, IEEE) received the B.E. degree in electronics from Kyungpook National University, Daegu, South Korea, in 1989, and the M.E. and Ph.D. degrees in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 1992 and 1996, respectively. He joined the Electronics and Telecommunications Research Institute, Daejeon, South Korea, as a Senior Research Staff Member, where he led the Realistic Broadcasting Media Research Team. In 2001, he joined the School of Engineering, Information, and Communications University (ICU), Daejeon, as an Assistant Professor. Since 2009, he has been with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, where he is currently a Full Professor. He was involved in scalable video coding and high-efficiency video coding in JCT-VC standardization activities of ITU-T VCEG and ISO/IEC MPEG. His current research interests include deep learning for image restoration and visual quality enhancement, deep video compression, perceptual video coding, visual quality assessments, computational photography, machine learning, and pattern recognition.