J. Korean Soc. Ind. Appl. Math. Vol.27, No.1, 56–74, 2023 https://doi.org/10.12941/jksiam.2023.27.056

MULTI-APERTURE IMAGE PROCESSING USING DEEP LEARNING

GEONHO HWANG[1], CHANG HOON SONG[1], TAE KYUNG LEE[2], HOJUN NA[1], AND MYUNGJOO KANG[1,†]

[1]DEPARTMENT OF MATHEMATICAL SCIENCES, SEOUL NATIONAL UNIVERSITY, SEOUL, 08826, REPUBLIC OF KOREA

[2] INTERDISCIPLINARY PROGRAM IN ARTIFICIAL INTELLIGENCE, SEOUL NATIONAL UNIVERSITY, SEOUL, 08826, REPUBLIC OF KOREA

Email address: {hgh2134, goldbach2, dlxorud1231, lahj91, †mkang}@snu.ac.kr

ABSTRACT. In order to obtain practical and high-quality satellite images containing high-frequency components, a large aperture optical system is required, which has a limitation in that it greatly increases the payload weight. As an attempt to overcome the problem, many multi-aperture optical systems have been proposed, but in many cases, these optical systems do not include high-frequency components in all directions, and making such an high-quality image is an ill-posed problem. In this paper, we use deep learning to overcome the limitation. A deep learning model receives low-quality images as input, estimates the Point Spread Function, PSF, and combines them to output a single high-quality image. We model images obtained from three rectangular apertures arranged in a regular polygon shape. We also propose the Modulation Transfer Function Loss, MTF Loss, which can capture the high-frequency components of the images. We present qualitative and quantitative results obtained through experiments.

1. INTRODUCTION

Remote sensing [1], the process of acquiring images of the ground from an aircraft or satellite, has achieved many accomplishments thanks to the rapid development of optical systems and satellite launch vehicles. Accordingly, there are increasing attempts to acquire various images and information through satellite observation. However, an optical system with high-quality and good optical properties must secure a large aperture and long image distance as a trade-off, which causes an increase in the volume and weight of the payload. This trade-off relationship between quality and weight has been recognized as an insurmountable problem for traditional optical systems.

Received December 30 2022; Revised March 20 2023; Accepted in revised form March 23 2023; Published online March 25 2023.

2020 *Mathematics Subject Classification.* 68U10.

Key words and phrases. Deep learning, Remote Sensing, Multi Image Deblurring, Modulation Transfer Function.

† Corresponding author.

As one of the new attempts to overcome this problem, a multi-aperture optical system [2] has been proposed. In the multi-aperture optical system, an optical path of the light passing through each aperture shares the same space, thereby reducing the volume of the entire optical system. Through this, a multi-aperture optical system having the same volume and mass can acquire a higher quality image by securing a sufficient image distance compared to a single-aperture optical system of the same size. However, multi-aperture optical systems suffer from severe diffraction due to the relatively small aperture than single-aperture systems. The diffraction distorts the features of the image in the frequency domain, which causes a blur in the spatial domain. Therefore, we can obtain only multiple blurred images corresponding to each aperture through the multi-aperture optical system. Fortunately, since each aperture is arranged in a different direction, the images obtained from each aperture have different blur characteristics. Thus, by utilizing and combining frequency information from multiple images in each direction simultaneously, we can expect to acquire a single high-quality image[3].

In this regard, we suggest a novel deep-learning method that combines multiple blurred images and generates a deblurred image. Most existing deep learning deblurring techniques are designed to output a deblurred image by taking a single image as an input. Unlike this, our problem setting aims to obtain better results by using multiple blurred images, which are complementary to each other. We revised the existing deep learning model that took a single image as input and presented a deep learning model that generates high-quality images by receiving multiple images and PSF as inputs. Our model overcomes the physical limitations of the optical systems and acquires high-quality images. Our contributions are as follows:

- We propose a novel task that receives multiple blurred images as input and performs deblurring. Our suggested deep learning architecture handles the raised problem.
- A deep learning network is presented to tackle the proposed task, and the performance of the network is confirmed quantitatively and qualitatively through experiments.
- We suggest an MTF Loss that helps the network to capture the high-frequency components of the image. The loss borrows the concept of Modulation Transfer Function(MTF) used in optics.

2. Related Works

Image Deblurring is the task of getting a high-quality image by recovering the latent sharpness of a blurred image. Deblurring is a highly ill-posed problem, and there were many methods to solve it based on the inverse problem [4]. But those methods failed to handle realistic blur due to insufficient blur modeling and ill-posedness of the inverse problem. Recently, deep neural network (DNN) based methods have been developed to tackle the problem. Those methods overcome such limitations so that they outperform others [5]. DNN-based image deblurring methods could be classified into single-image deblurring and multi-image deblurring.

Single-image deblurring is to recover a latent sharp image from a blurry image. Single-image deblurring methods have been developed and have shown better performance in highly-ill-posed problems like deblurring using known blur kernels, so-called blind image deblurring [5, 6]. In [4], the authors created the RealBlur dataset, which consists of real-world images

to train models and make better results. In [7], a re-blur architecture called DBRBGAN is proposed, which includes two generative adversarial network models, one for blurring and the other for deblurring. In [8], a deblurring model for saturated images exploiting a maximum posterior-based optimization framework is introduced. But, still, single-image deblurring has limitations. First, it lacks information on a latent sharp image. The information from only a single image is usually not enough. Second, in the case of a real image, the kernel and motion of the camera and that of the object would be different, which results in variant images.

However, depending on the kernel and motion, different images include different information. As an instance, [3] shows that if blur kernels have directionality, the sharpness of the image depends on the direction. A sharp image can be attained by combining multiple images which are blurred in different directions. In this way, *multi-image deblurring* can solve the limitations of single-image deblurring by obtaining multiple pieces of information that complement each other image. The authors of [9] used two motion-blurred images with different moving directions to get a finer image. In [10], the authors remove dynamic scene blur by exploiting recurrent neural network to remove dynamic blur and convolutional neural network to restore the sharp images.

Most of the image deblurring methods focus on casual images in the human eye, like cars, streets, animals, etc. But, there are deblurring methods for satellite images. For example, a method [11] is proposed based on feature alignment, of which a feature alignment module (FAFM) aligns the feature maps according to the adjusted position of each sample in the convolution kernel, and a feature importance selection module (FISM) filters the feature maps to preserve reliable details. In [12], the authors used a local binary pattern prior, which can record the features of images to filter out pixels containing important features in blurry images. There are several satellite image deblurring methods using a single image [11, 12, 13]. But, we propose multi-image deblurring for satellite images. Also, there is a modulation transfer function (MTF) value to measure the quality of satellite images. In another paper [13], MTF is just used to evaluate. But our proposal uses MTF value in the training process.

3. BACKGROUNDS

3.1. **Point Spread Function.** In this section, we introduce some concepts from optics. The purpose of the optical system is to image the information of the target to be observed through the detector. And an ideally operating optical system can divide the target so that each divided area corresponds to one sensor. The spatial size of the target corresponding to one sensor is called the ground sampling distance(GSD). In other words, GSD means the range included in one pixel, so it often represents the performance of an optical system used in the satellite field. However, in the case of an optical system in a practical setting, the light emitted from one point light source is reflected on multiple sensors because of the diffraction. And the kernel describing this is called Point Spread Function (PSF) in optics. In other words, PSF is the impulse response of the optical system, and the response of the optical system to the entire target can be expressed as a linear combination of responses to each point light source by linearity. Expressing this mathematically, the image I of the target T detected by the optical

system is as follows.

$$I = (k * T) + n,$$

where k is the kernel(PSF) of the system, n is the noise, and $*$ denotes the convolution operator.

Thus, if we need to simulate the relationship between the target and the image, we have to calculate the PSF of the optical systems. The PSF of the optical system is affected by various factors, including the shape and the size of the aperture, the distance to the target, the wavelength observed by the detector, etc. Fortunately, there are several extreme cases in which we can easily calculate the effect of the diffraction using these variables. Among these, the diffraction situation when the distance between the target and the aperture is sufficiently far is called Fraunhofer diffraction [14]. Since this assumption is satisfied under the remote sensing situation, we can calculate the PSF under the Fraunhofer diffraction condition of the optical system.

Let $P(\xi, \eta)$ be the aperture's indicator function:

$$P(\xi, \eta) = \begin{cases} 1 & \text{if } (\xi, \eta) \text{ belongs to the aperture} \\ 0 & \text{otherwise} \end{cases} .$$

where (ξ, η) represents spatial coordinates. Then, in the case of Fraunhofer Diffraction, the PSF $\mathcal{P}(x, y)$ is calculated as follows [3]:

$$\mathcal{P}(x, y) = \left| \iint_D P(\xi, \eta) \exp\left[-2i\pi \left(\frac{x}{\lambda l}\xi + \frac{y}{\lambda l}\eta \right) \right] d\xi d\eta \right|^2 , \tag{3.1}$$

where λ is the wavelength of the light observed by the detector, D is the domain of the aperture, and l is the image distance, the distance to the detector from the aperture.

3.2. **Modulation Transfer Function (MTF).** It is often convenient to see the degradation from the perspective of the frequency domain. The Optical Transfer Function(OTF) is the frequency response of the optical system [15]. The OTF observes how the image is degraded in terms of the frequency domain. Because the convolution operator is converted to the product operator in the Fourier transform, it becomes

$$\mathfrak{F}(I) = OTF\left(\mathfrak{F}(T)\right),$$

where OTF is the OTF function, and $\mathfrak{F}$ denotes the Fourier transform. The OTF is the complex-valued function, and the magnitude of OTF is used to evaluate the accuracy of signal separation. In this regard, the Modulation Transfer Function(MTF), the magnitude of the (OTF), is used to depict the performance of the optical system:

$$MTF = |OTF|.$$

The optical system with good optical properties has an MTF value near one, retaining all information along the entire frequency range. In this regard, the MTF curve in the frequency domain

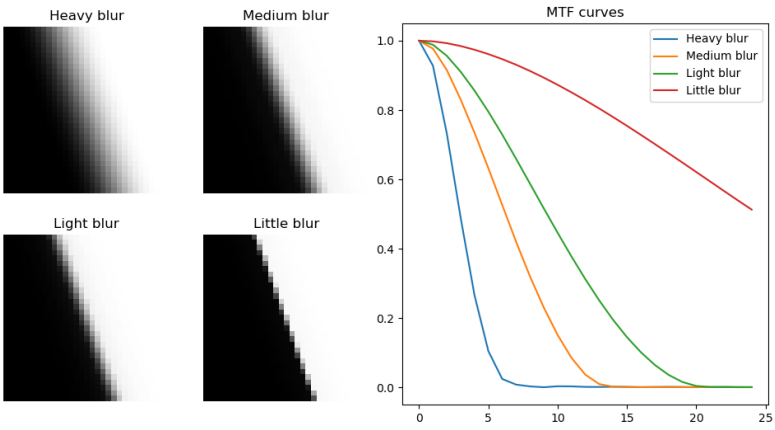

FIGURE 1. Different blurred edges (left) and correspondent MTF curves (right).

is used as the criterion for optical systems. Typically, the MTF value at the Nyquist frequency [16] defined as follows is also used for the evaluation:

$$f = \frac{1000 \frac{\mu m}{mm}}{2 * (\text{pixel size}) \, \mu m}.$$

In the image, MTF is measured at an edge, and the MTF value at a specific frequency means how accurately a signal with the frequency can be divided by the edge. For example, an image of size 100×100 with an MTF value of one at the Nyquist frequency perfectly distinguishes a signal whose frequency is 50; as the MTF value decreases, the signal is blurred. Figure 1 shows the correlation between the blur in the image and the MTF. Hence, an MTF measures the sharpness of an edge in an image.

3.3. **Measuring of the MTF.** Using the MTF, we can evaluate the ability of the optical system to measure how accurately the target is converted into an image. However, this is a value that exists only theoretically, and in order to calculate it, a process of finding out the MTF through an image is required [17]. In order to measure the MTF of the optical system through the image, a target assumed to be an ideal edge and an image observed through the optical system are used. If an ideal edge target exists, the values on both sides of the edge will be exactly bisected. Assuming that the target is degraded by the MTF and the image is created, the MTF can be restored by modeling this process. MTF is calculated through the following process theoretically.

- We detect the one edge obtained from the area which separates a signal from the image.
- Edge Spread Function (ESF): We choose pixels with a pre-defined distance from the perpendicular line to the detected edge above and get the pixel values along the line.

Then, by smoothing the pixel values, we generate a function that depicts the behavior of the pixel values near the edge.
- Line Spread Function (LSF): We differentiate ESF to obtain LSF.
- MTF: Finally, we generate the MTF curve created by Fourier transforms the LSF.

More detail on the process of calculating MTF for a real image is introduced in section 5.3.

3.4. **MANet.** In [18], the authors proposed a two-phase non-blind deblurring method called a mutual affine network (MANet). The two-phase non-blind deblurring process is divided into two steps. First, the kernel is estimated using a blurred image. Second, the deblurred image is predicted using the kernel estimated in the first step and the blurred image. In mathematical form, the low-resolution blurred image I^{LR} (low-quality image) is generated from the high-resolution image I^{HR} (high-quality image) using the following equation,

$$I^{LR} = (k * I^{HR}) \downarrow_s + n,$$

where $*$ denotes the convolution operator, and $\downarrow_s$ is the downsampling with scale factor s. In other words, if the high-resolution image I^{HR} is in $\mathbb{R}^{C\times H\times W}$, the low-resolution image I^{LR} is rescaled to $\mathbb{R}^{C\times \frac{H}{s}\times \frac{W}{s}}$. The entire operation $(k * I^{HR}) \downarrow_s$ can be interpreted as the single operation KI^{HR} where K denotes the blur kernel from I^{HR} to I^{LR}. MANet tackles the space-variant kernel estimation problem; that is, the kernel K can be different in different pixel points. MANet is trained to take the input I^{LR} and output an estimated kernel K. For the deblurring part, the authors adapted the RRDB-SFT network with RRDB blocks [19] and SFT layers [20]. In our method, the way to make low-quality images would be different. But still use the concepts of kernel estimation and deblurring.

4. METHODS

4.1. **Problem Formulation.** Our goal is to take blurred or low-quality images and convert them into a single deblurred or high-quality image. We focus on the optical systems described in Fig. 2, but our method is not restricted to the specific optical systems. Indeed, our framework can cover an arbitrary multi-aperture optical system consisting of the optics with the apertures that the PSF of the aperture can be calculated. Our optical systems of interest consist of four sub-optical systems, as described in Fig. 2. Each sub-optical system contains an aperture, and since other parameters are usually fixed in our situation, like a satellite image, the aperture mainly determines the property of the optical system. Sub-optical systems fall into two classes, wide field-of-view optical system(WFOV) and narrow field-of-view optical systems(NFOV). The entire optical system consists of one WFOV and three NFOVs. WFOV and NFOV have been delicately designed to operate in a complementary manner. WFOV has a circular aperture, denoted as W in Fig. 2. It has a relatively larger aperture compared to NFOVs. A larger aperture needs a longer image distance to perform all its potential. However, the image distance of WFOV is short for the system to perform at its all potential due to the constraints on the volume of the optical system. Therefore, the WFOV has a much larger GSD than NFOVs. On the other hand, NFOVs have rectangular apertures, denoted as $N1$, $N2$, and $N3$ in Fig. 2. NFOVs have a relatively long image distance, which results in a smaller GSD with respect to that of the

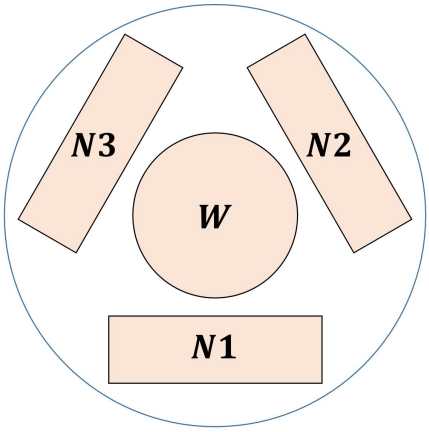

FIGURE 2. Arrangement of apertures in the multi-aperture optical system.

WFOV. However, the small size of the aperture of the NFOV results in high diffusion in the small GSD field. To partially overcome the limit, the three NFOVs are arranged at different angles. A rectangular aperture produces less diffraction along the longer side. This means that the convolution in the longer side direction occurs with smaller support, so the information loss in that direction is smaller. By arranging three NFOVs rotated by 120 degrees, we can efficiently collect information from all directions.

In mathematical formulas, the optical system in Fig. 1 is represented as follows. Let k_0 be the PSF of the aperture of WFOV W and k_1, k_2, and k_3 be PSFs of apertures of NFOV $N1, N2$, and $N3$, respectively. Then, for the target T and $i = 1, 2, 3$ the following equation holds for NFOV images:

$$I_i = (k_i * T) + n_i, \tag{4.1}$$

where n_i is the Gaussian noise. For the WFOV image, the following equation holds:

$$I_0 = k_0 * (T \downarrow_s) + n_0,$$

where $\downarrow_s$ is the downscaling with scale factor s and n_0 is the Gaussian noise. Scale factor s could be any number, so we set s to be five in the remaining. Note that the WFOV image I_0 has a size smaller than the NFOV images due to the larger GSD of WFOV. Given the blurred images I_0, I_1, I_2, I_3, we want to restore the original target T.

4.2. **Calculation of the PSF.** For the training of the deep learning model, we need pairs of the original image and blurred images. Blurred images can be generated from the original image using the PSF as Eq. (4.1), and we can calculate the PSF corresponding to its aperture using Eq. (3.1). Let NFOV has a rectangular aperture. In the case of a rectangular aperture with a length of w_x on the x-axis and w_y on the y-axis, the PSF is calculated as follows using the $sinc$ function.

$$p_{sinc}(x, y) = sinc^2\left(\frac{xw_x}{\lambda l}\right) sinc^2\left(\frac{yw_y}{\lambda l}\right),$$

where $sinc(x) := \frac{sin(x)}{x}$, λ is the wavelength, and l is the image distance. And for a circular aperture with a diameter of D, the PSF is calculated as follows:

$$p_{Bessel}(x, y) = \left(\frac{2J_1(\rho)}{\rho} \right)^2,$$

where J_1 is a Bessel function of the first kind $\rho = D\pi r/\lambda l$, and $r = \sqrt{x^2 + y^2}$.

Also, there is an additional effect that occurs in multi-aperture optical systems: the alignment error. Because the sub-optical systems could not be perfectly aligned, images observed by different optical systems are subject to different degrees of translation effects. Thus, the actual image generated by the optical system with the PSF k_i becomes

$$L_i(k_i * T + n_i),$$

where L_i is a translation operator.

4.3. **Deep Learning Architecture.** Although we model the PSF for each aperture, there may be unknown randomnesses like noise, translation, or distortion according to the environment where the image is captured. Especially, the translation or distortion heavily affects the quality of predictions of a deblurring model. Thankfully, almost all conditions, including image distance and focal length, are almost equal in a satellite image; there is little distortion. In addition, we use two-phase deblurring as if the kernel is unknown with some randomness of aperture parameters in training as an augmentation. The kernel estimation phase makes the deep learning model more robust to some noise in capturing environment or even in the misalignment of multiple images, as explained in the following paragraph.

We adapt and revise the MANet [18] as the kernel estimation model and use the RRDB-SFT as the deblurring model. However, there is a problem with using naive MANet as the kernel estimation model, which takes a single image as the input. For images that have different translation errors, the pixel in the same position of each image does not point to the same position in the target. Also, the error can be variant in every detection, which makes consistent correction impossible. For multi-image deblurring, multiple images need to be concatenated and used as input. Even subpixel align errors can lead to a fatal error in this deblurring process.

To overcome this problem, we note that the translation of either kernel k_i or image T yields the same translation of the final image $k_i * T$; for a translation operator $L_v I(x) \coloneqq I(x - v)$, we have

$$L_v \left(k_i * T\right)(x) = \left(L_v k_i\right) * T = k_i * \left(L_v T\right) = I(x - v).$$

Thus, we revise the MANet to predict the kernel with the translation error. On the other hand, because the translation is the relative value, we need the template to conduct the translation. Therefore we paired the WFOV image and an NFOV image to estimate the relative position of the NFOV image. It is implemented as the following. For the WFOV image I_0 and an NFOV image I_i, we first resize the WFOV image I_0 to the size of the NFOV image.

$$\widetilde{I_0} = \left(I_0\right) \uparrow_s .$$

Then, we concatenate the resized WFOV image $\widetilde{I_0}$ and the NFOV image I_i along the channel.

$$D_i = \widetilde{I_0} \oplus I_i.$$

The original spatially invariant version of the MANet can be considered the function

$$\text{MANet} : \mathbb{R}^{c\times W\times H} \to \mathbb{R}^{k\times k}.$$

In our version, we both estimate the kernel of the WFOV image and the NFOV image using the concatenation of the WFOV image and the NFOV image. We properly adjust the number of input and output channels in MANet as follows.

$$\text{MANet} : \mathbb{R}^{2c\times W\times H} \to \mathbb{R}^{2\times k\times k}.$$

The first channel $\mathbb{R}^{k\times k}$ is the estimated kernel of the WFOV image, which is centered at the origin. The second channel $\mathbb{R}^{2\times k\times k}$ is the estimated kernel of the NFOV image. The kernel of the NFOV image may not be centered at the origin to correct the translation error between the WFOV and NFOV images. Then, by applying MANet three times to each pair of the WFOV and NFOV images, we have

$$k_0 \oplus k_i = \text{MANet}(D_i) = \text{MANet}(\widetilde{I_0} \oplus I_i),$$

for $i = 1, 2, 3$ where k_0 is the kernel of the WFOV image and k_1, k_2, k_3 are the kernel of NFOV images. Because the WFOV kernel k_0 is redundantly computed three times, we take the pixel-wise average of all predicted WFOV kernels. Although averaging multiple predictions can be significantly different from the original PSF, every model for each i predicts a very similar WFOV kernel. Indeed, as we use the same WFOV images as reference for each NFOV, all models output the same WFOV kernel provided they are well trained. Figure 4 shows the input images and NFOV and WFOV kernels from the images, and Fig. 5 shows the variance of WFOV kernels.

Now the remaining is to reconstruct the deblurred image using four images and four estimated kernels. The original RRDB-SFT model consist of two parts. The RRDB part extracts the feature information from the kernel and the blurred image.

$$F = RRDB(I, K).$$

Then, the original image is restored using the residual connections in the feature and the original image through convolutional layers $Conv_1, Conv_2$.

$$\widetilde{T} = Conv_1(F) + Conv_2(I).$$

In the multi-image case, we modified this process in the following manner. First, extract features from each pair of the kernel and the blurred NFOV image.

$$F_i = RRDB(I_i, k_i),$$

for $i = 1, 2, 3$. Then, using the

$$\widetilde{T} = Conv_1(F_1 \oplus F_2 \oplus \cdots \oplus F_n) + Conv_2(\widetilde{I_0}),$$

the estimated target image $\widetilde{T}$ is compared to the original target image T in the training process.

FIGURE 3. Small patch P (red box) and unique edge E (blue line), and normal line to the edge (red line) in P.

5. EXPERIMENTS

5.1. **Datasets.** We use two datasets to evaluate our model, the WHU-Aerial dataset [21] and the Urban 3D dataset [22]. The WHU-Aerial dataset is the dataset consisting of satellite images. It contains 8,192 patches of the image, each 512×512 in size. All of the images have the same GSD of 0.3m. Similarly, USSOCOM Urban 3D Challenge Benchmark Dataset, or the Urban 3D dataset, consists of 896 satellite images. Each image is 2048×2048 in size, and the GSD is 0.5m. We use the Urban 3D as the training dataset and the WHU-Aerial dataset as the test dataset.

5.2. **Training Data Preprocessing.** As an augmentation and preprocessing, we resize Urban 3D images to have 1024×1024 size and GSD 1m. Then we crop a 128×128 patch from the resized image and train a model on the patch.

For each sample T, we generate four blurred images; three of them, I_1, I_2, and I_3, are the convolution of rotated p_{sinc} and T with rotation degree 0, $\frac{2}{3}\pi$, $\frac{4}{3}\pi$, and the other, I_0, is the convolution of p_{Bessel} and T. We add Gaussian noise to all blurred images, and our PSF estimation model predicts kernels k_0, k_1, k_2, and k_3 from I_0, I_1, I_2, and I_3. The following deblurring model recovers I from I_0, I_1, I_2, I_3 and k_0, k_1, k_2, and k_3.

5.3. **MTF in the real image.** In the ideal case, MTF is computed with an image patch obtained from a clean signal, for example, a checkerboard. In practice, however, since an image does not include such a checkerboard, a patch with a unique edge is used instead. Suppose we have such patch P and unique edge E in P (Figure 3). Then a line spread signal is obtained from P along the line normal to E. The line spread signal is too noisy to be used for measuring MTF. Hence we regress the signal by a parameterized curve: convolution of a Bessel kernel and clean

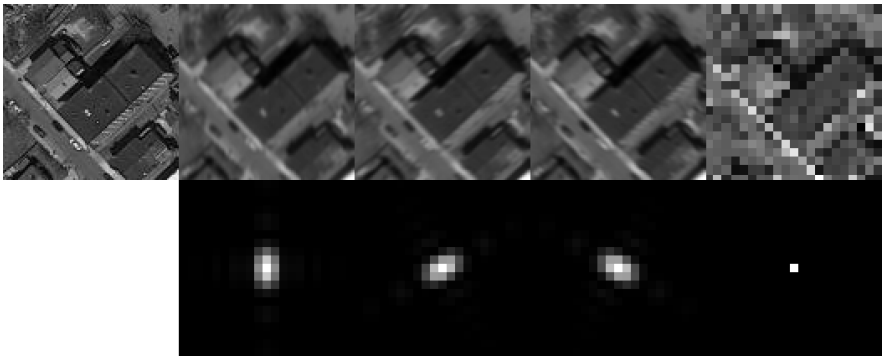

FIGURE 4. Estimation of the kernel(PSF) of each NFOV and WFOV image. The first row indicates the input image, and the second shows the estimated kernels. Each column means GT, NFOV1, NFOV2, NFOV3, and the average of three predicted three WFOV kernels.

signal. In other words, we generate the line spread function by fitting points using the following function

$$a + b\left(\frac{2J_1(cx+d)}{cx+d}\right)^2,$$

where a, b, c, and d are the four free paraemters that can be used in the fitting. The optimization process is conducted using the Newton method. An edge spread function, ESF, is defined by the optimized curve. A line spread function, LSF, is defined by differentiating ESF. Finally, the MTF curve is defined by the Fourier transform of the LSF. Although the MTF is actually a curve, we will focus on the specific frequency, the Nyquist frequency. We often call the value at the Nyquist frequency MTF, as there is no room for confusion. A function MTF is a curve, and a real number MTF is the value of the curve.

5.4. **MTF Loss.** Note that we have removed some bad patches after non-random cropping. In the process where a patch is determined to be removed or not, a smaller patch and its edge direction have been obtained. Using the same smaller patch and edge direction, we can regress the MTF curve at the output of our model. Then we define the *MTF Loss* by the L_2 distance between two MTF curves, one from the patch of the target image and the other from the output of our model. In this paper, MTF is measured in six directions: 0°, 30°, 60°, 90°, 120°, and 150°, and the sum of the MTF loss in each direction is calculated. All the values can be calculated in a differential manner. Thus the difference can be used as the loss function during the training process.

5.5. **Training Setting.** The rectangular aperture of the NFOV has 110mm$\times 70$mm in size, and the image distance is set to $2,750$mm. The circular aperture of the WFOV has a diameter of 140mm, and the image distance is set to 550mm. In common for both types of optical systems, the wavelength 550nm is used for calculation, and the pixel size of detectors is $3.76\mu m$. For

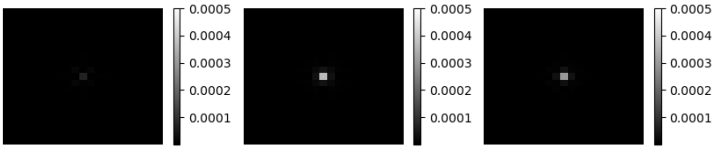

FIGURE 5. The absolute difference between a WFOV kernel predicted with each NFOV and the average of all WFOV predictions. The corresponding NFOV is NFOV1, NFOV2, and NFOV3 from the left.

the robustness of the trained model, each length of the aperture is randomly sampled within a certain range during the training process to give randomness to the data. We sample uniformly from 0.8 times to 1.2 times the length of apertures. In the test process, the values are fixed and evaluated. Also, we translate the image and the corresponding kernel by l pixels. l is sampled from $[-2, 2]$ uniformly. The Gaussian noise has the variance sampled from the uniform distribution $[0, 20]$. We use the L_1 loss and the MTF loss together as follows:

$$L_1 + \lambda(\text{MTF Loss}),$$

where $\lambda = 1$ is in our setting. Detailed training hyperparameters are in the appendix.

5.6. **Results and Discussion.** We compare three methods. First, we conducted the single image deblurring using the MANet. And next, we constructed the image from multiple images. Finally, we compare the multi-image deblurring with the MTF loss. Each model is trained independently on the same dataset.

Method	PSNR	SSIM
MANet	24.11	0.682
Ours	25.32	0.744
Ours(with MTF Loss)	**25.35**	**0.751**

We can see that our method achieves much better PSNR and SSIM scores than the single image deblurring method using MANet. Also, the model with the MTF Loss has the highest score among the methods. Moreover, there is a big difference in the qualitative result. Figure 6 shows the reconstructed images by three methods. From the left, they are ground truth image, deblurred image using single image deblurring using MANet, deblurred image using multi-image deblurring, and deblurred image using multi-image deblurring with the MTF Loss. Our multi-image deblurring method shows much better results than the existing single-image deblurring method. In addition, when using the MTF Loss, it can be observed that detailed straight-line patterns or small objects are much better restored than when MTF Loss is not used. This shows that MTF Loss has a strong effect on restoring the high-frequency part of the image.

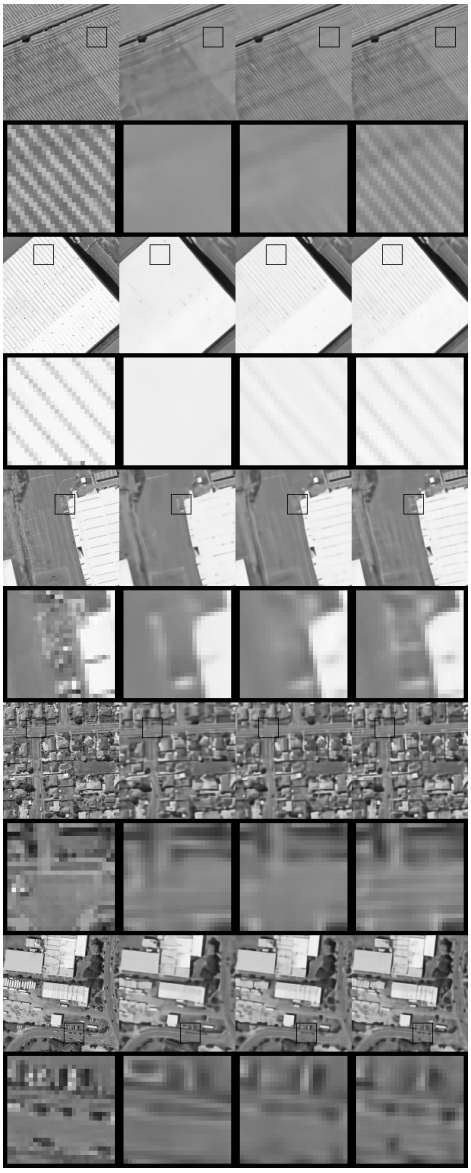

FIGURE 6. Ground truth image, deblurred image using single image deblurring using MANet, deblurred image using multi-image deblurring, and deblurred image using multi-image deblurring with the MTF Loss, from left to right. For each image, a black bounding box is enlarged and displayed below it.

6. Conclusion

In this paper, we propose a novel multi-image deblurring method based on a multi-aperture optical system. We adapt the concept of the MTF from the optics, which is used to select the training image and help the training process. Our method provides a new perspective on understanding the multi-image deblurring method. We hope that studies that obtain better results by combining optical knowledge with deep learning can continue.

Acknowledgement

This work has been supported by the Challengeable Future Defense Technology Research and Development Program through ADD[No. 915020201], the NRF grant[2012R1A2C3010887] and the MSIT/IITP[No. 2021-0-01343].

References

[1] Campbell, J. & Wynne, R. *Introduction to remote sensing*, Guilford Press, 2011.

[2] Green, P., Sun, W., Matusik, W. & Durand, F. *Multi-aperture photography*, Association for Computing Machinery, Proceedings of ACM SIGGRAPH, San Diego, California, USA 2007.

[3] Lv, G., Xu, H., Feng, H., Xu, Z., Zhou, H., Li, Q. & Chen, Y. *A Full-Aperture Image Synthesis Method for the Rotating Rectangular Aperture System Using Fourier Spectrum Restoration*, Photonics, **8** (2021), 1–17.

[4] Rim, J., Lee, H., Won, J. & Cho, S. *Real-world blur dataset for learning and benchmarking deblurring algorithms*, Lecture Notes in Computer Science, Springer, Proceedings of ECCV, Glasgow, UK 2020.

[5] Zhang, K., Ren, W., Luo, W., Lai, W., Stenger, B., Yang, M. & Li, H. Deep image deblurring: A survey. *International Journal Of Computer Vision*. **130** (2022), 2103-2130.

[6] Bai, Y., Jia, H., Jiang, M., Liu, X., Xie, X. & Gao, W. *Single-image blind deblurring using multi-scale latent structure prior*, IEEE Transactions On Circuits And Systems For Video Technology, **30** (2019), 2033-2045.

[7] Zhang, K., Luo, W., Zhong, Y., Ma, L., Liu, W. & Li, H. *Adversarial spatio-temporal learning for video deblurring*, IEEE Transactions On Image Processing, **28** (2018), 291-301.

[8] Chen, L., Zhang, J., Lin, S., Fang, F. & Ren, J. *Blind deblurring for saturated images*, Proceedings Of Conference On Computer Vision And Pattern Recognition, 2021.

[9] Wang, P., Sun, J., Li, H., Chen, X., Zhu, Y. & Zhang, Y. *Multi-image Deblurring Using Complement*, Lecture Notes in Computer Science, Springer, Proceedings of International Conference On Intelligent Science And Big Data Engineering, Dalian, China, 2017.

[10] Zhang, J., Pan, J., Wang, D., Zhou, S., Wei, X., Zhao, F., Liu, J. & Ren, J. *Deep Dynamic Scene Deblurring from Optical Flow*, IEEE Transactions On Circuits And Systems For Video Technology, **32** (2021), 8250 - 8260

[11] Zhu, B., Lv, Q., Yang, Y., Sui, X., Zhang, Y., Tang, Y. & Tan, Z. *Blind Deblurring of Remote-Sensing Single Images Based on Feature Alignment*, Sensors, **22** (2022), 7894

[12] Zhang, Z., Zheng, L., Piao, Y., Tao, S., Xu, W., Gao, T. & Wu, X. *Blind Remote Sensing Image Deblurring Using Local Binary Pattern Prior*, Remote Sensing, **14** (2022), 1276

[13] Kim, H., Seo, D., Jung, J., Cha, D. & Lee, D. *Blind motion deblurring for satellite image using convolutional neural network*, IEEE, Proceedings of Digital Image Computing: Techniques And Applications(DICTA), Perth, Australia 2019.

[14] Born, M. & Wolf, E. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*, Elsevier, 2013

[15] Williams, C. & Becklund, O. *Introduction to the optical transfer function*, SPIE Press, 2002.

[16] Condon, J. & Ransom, S. *Essential radio astronomy*, Princeton University Press, 2016.

[17] Viallefont-Robinet, F., Helder, D., Fraisse, R., Newbury, A., Bergh, F., Lee, D. & Saunier, S. *Comparison of MTF measurements using edge method: towards reference data set*, Optics Express, **26** (2018), 33625-33648

[18] Liang, J., Sun, G., Zhang, K., Van Gool, L. & Timofte, R. *Mutual affine network for spatially variant kernel estimation in blind image super-resolution*, Proceedings Of The IEEE/CVF International Conference On Computer Vision, Montreal, QC, Canada 2021.

[19] Wang, X., Yu, K., Dong, C. & Loy, C. *Recovering realistic texture in image super-resolution by deep spatial feature transform*, Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition, Salt Lake City, UT 2018.

[20] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y. & *Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks*, Proceedings Of The European Conference On Computer Vision (ECCV) Workshops, Munich, Germany 2018.

[21] Ji, S., Wei, S. & Lu, M. *Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set*, IEEE Transactions On Geoscience And Remote Sensing, **57** (2018), 574-586.

[22] Goldberg, H., Brown, M. & Wang, S. *A benchmark for building footprint classification using orthorectified rgb imagery and digital surface models from commercial satellites*, IEEE, Proceedings Of IEEE Applied Imagery Pattern Recognition Workshop(AIPR), Washington, DC, USA 2017.

APPENDIX A. TRAINING HYPERPARAMETERS

We use the following hyperparameters for the training.

- Optimizer: Adam
- Learning Rate: 5×10^{-4}
- β_1 : 0.9
- β_2 : 0.99
- Learning Rate Scheduling: Cosine Annealing Restart
- Iteration: $235,000$
- Batch Size: 4

APPENDIX B. MODEL ARCHITECTURE

In this section, we describe the detailed architectures of the networks.

B.1. **MANet.**

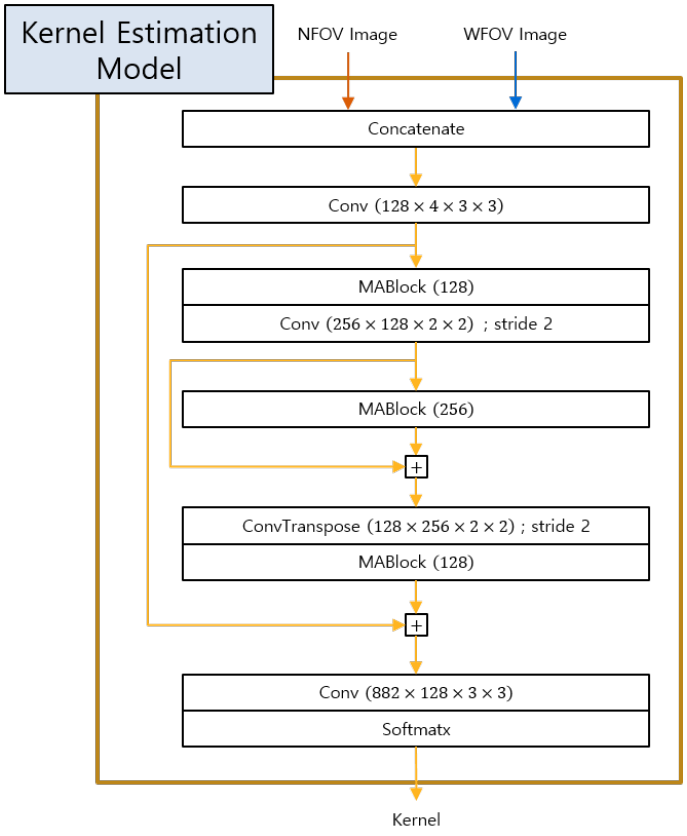

FIGURE 7. Entire MANet model structure used in the kernel estimation.

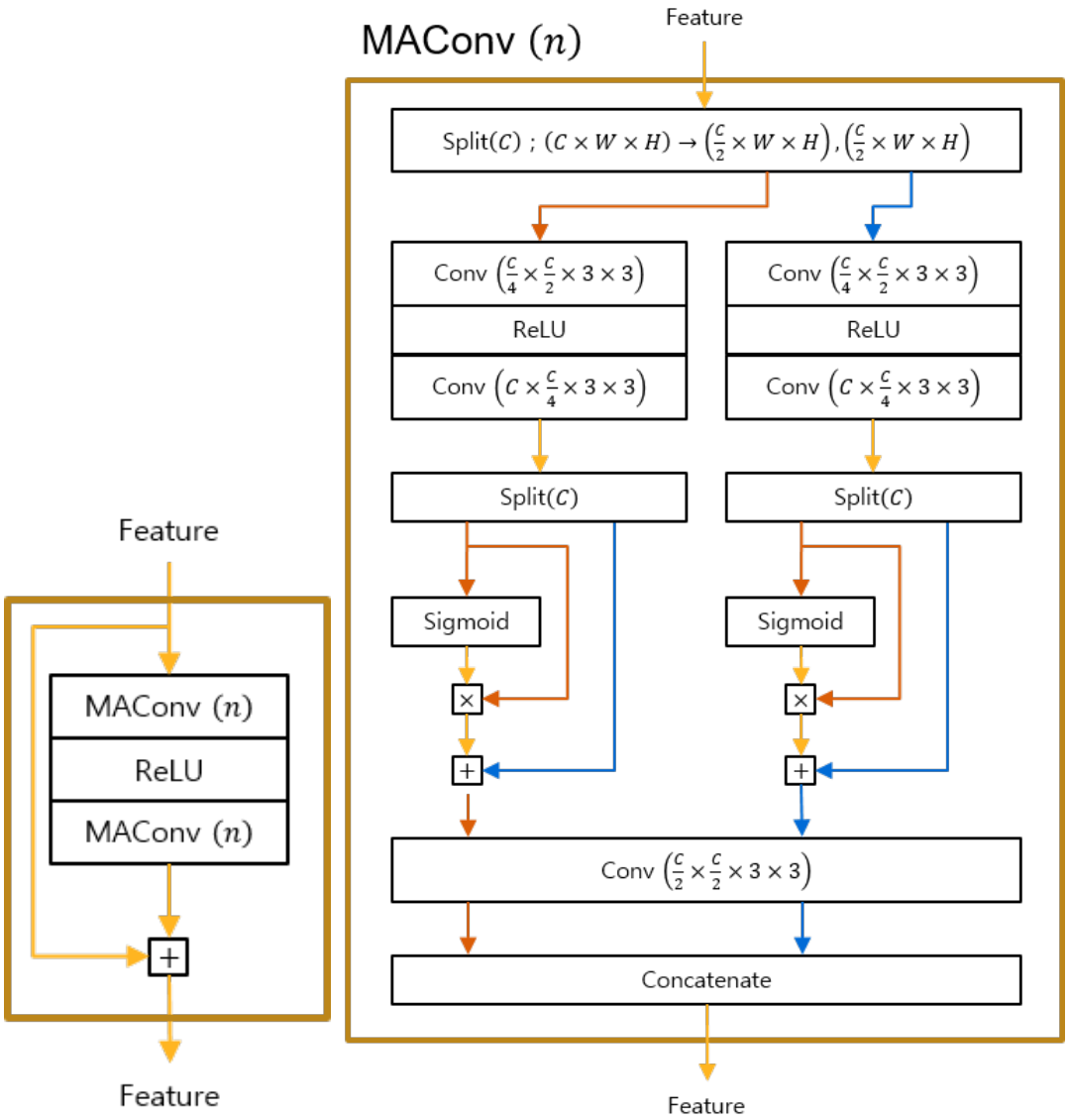

FIGURE 8. MABlock submodule and MAConv submodule.

B.2. **RRDB-SFT.**

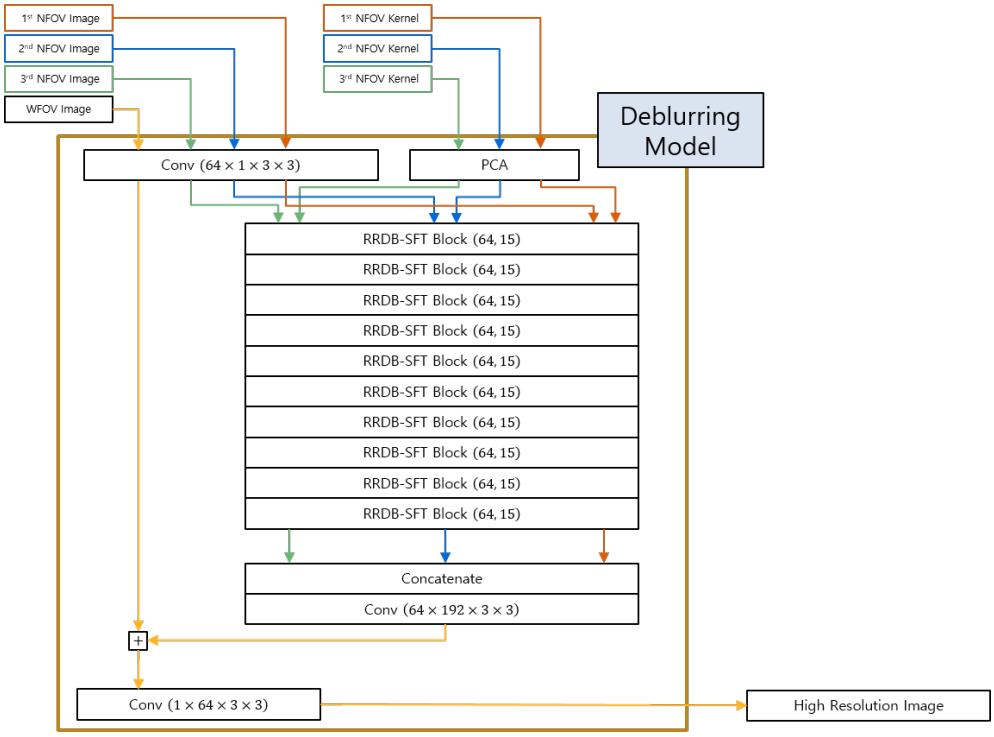

FIGURE 9. Entire RRDB-SFT model structure used in the deblurring.

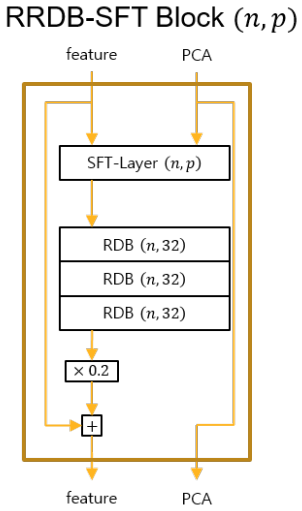

FIGURE 10. RRDB-SFT.

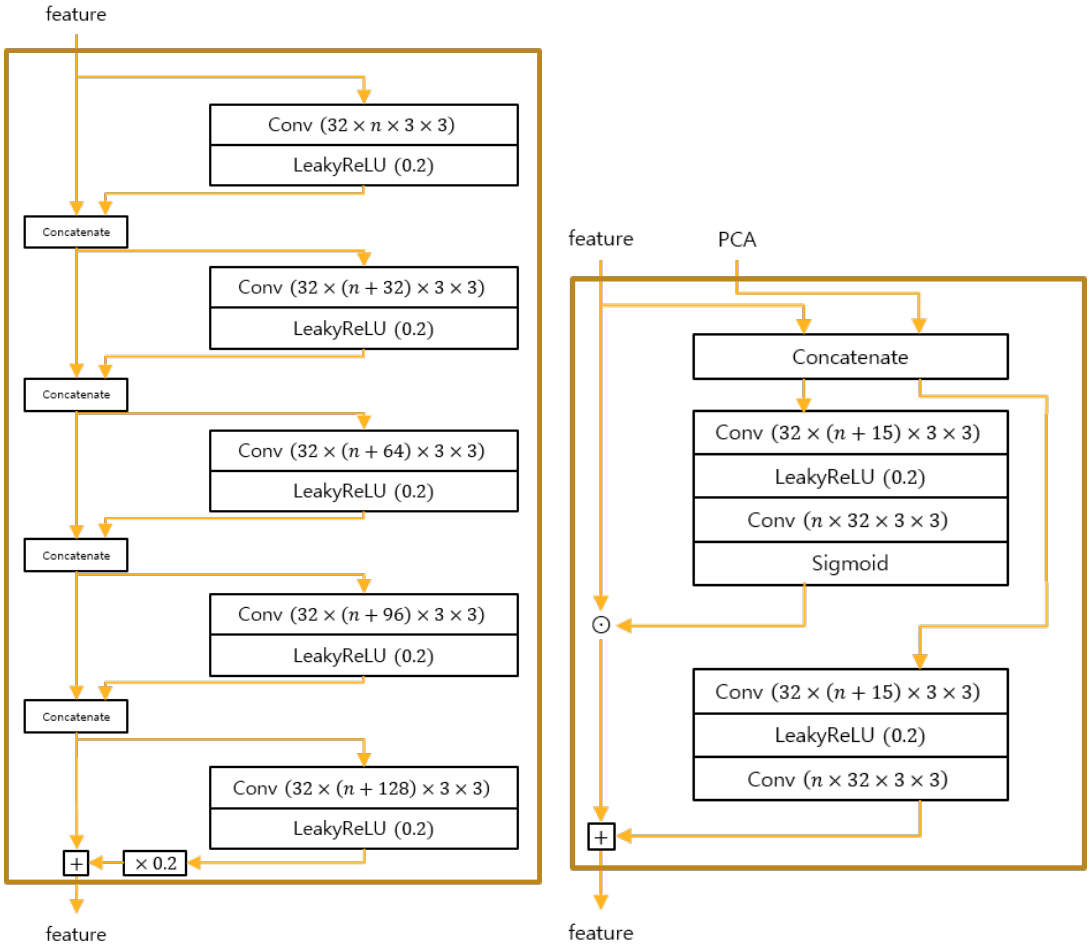

FIGURE 11. RDB submodule and SFT submodule.