S & M 3094

Swin Transformer UNet for Very High Resolution Image Dehazing

Yuxin Bian,[1] Enguang Zhang,[1,2] Jiayan Wang,[1] Rixin Xie,[1] and Shenlu Jiang[1*]

[1]School of Computer Science and Engineering, Faculty of Innovation Technology,
Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau SAR, China
[2]School of Mechanical Engineering, Zhuhai College of Science and Technology, Jinwan, Zhuhai 519040, China

(Received July 29, 2022; accepted September 22, 2022)

Keywords: image dehazing, VHR image processing, deep learning, transformer, UNet

Rapid image acquisition for a region affected by an earthquake is important to manage the rescue operation. The use of an unmanned aerial vehicle (UAV) to rapidly cruise an affected region and obtain very high resolution (VHR) images is highly advantageous. However, haze is a problem for many UAV aerial images, especially when UAVs cross clouds. In this paper, we present a parallel predicting workflow that cooperates with Swin Transformer UNet (ST-UNet) for this task. ST-UNet utilizes the Swin Transformer instead of a convolutional layer (CNN), which greatly enhances the processing speed without accuracy loss. The predicting workflow employs parallel processing and a reasonable data structure to maximize the computing resources for rapid processing. To demonstrate the advantageousness of the proposed workflow, we employed three public remote sensing datasets for evaluation, and the proposed ST-UNet obtained the highest accuracy and speed. Furthermore, the high dehazing performance of ST-UNet was demonstrated using a real post-earthquake scene.

1. Introduction

The rapid and accurate mapping of earthquake-induced infrastructure damage is an important task. Currently, unmanned aerial vehicles (UAVs) are commonly used to accomplish this task as they provide high-fidelity images and flexible imaging of affected regions in a rapid-response mode. The UAV must sometimes cross a cloudy environment in this task, which leads to hazed images. Hazed images have greatly reduced visibility and illumination, leading to low accuracy in detection tasks for post-earthquake evaluation, such as the detection of damaged roads and buildings. Image dehazing technology, which recovers the visibility and illumination of original images, is one of the most recent achievements in artificial intelligence. Image dehazing algorithms can be grouped into three categories. 1) Image enhancement(1,2) is a method with low complexity that enables rapid processing. However, it is difficult for a simple mathematical formula to cover all situations in the real world, making the results highly dependent on the parameter settings. 2) Physical dehazing(3,4) utilizes prior knowledge based on

[*]Corresponding author: e-mail: shenlujiang@must.edu.mo
https://doi.org/10.18494/SAM4059

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

an atmospheric scattering model to estimate the parameters of the dehazing model for an image. However, obtaining the prior knowledge requires a very large amount of data to be collected, and this method cannot cover all situations, leading to frequent failure during processing. 3) Deep learning technology [convolutional neural networks (CNNs),(5) generative adversarial networks (GAN),(6) etc]. are used to generate dehazing systems. In general, high-quality images can be restored using state-of-the-art deep neural networks (DNNs).

Although dehazing methods based on DNNs have very promising accuracy, they mostly process normal images. Several methods have also been proposed for dehazing remote sensing images. However, direct dehazing using existing DNNs on very high resolution (VHR) images has the following issues:

1) The image size is very large for VHR images, making it difficult to input the image to the DNN in a single step.
2) The response time for post-earthquake evaluation should be very short for rapid management, but DNNs generally have a very long processing time.

To resolve the above issues, we propose a workflow for rapid VHR image dehazing. The workflow consists of three steps. 1) Parallel processing technology is employed to efficiently divide VHR images, allocate the resources of the system memory, and maximize the usage rate of the proposed workflow. 2) A shallow-weight transformer DNN is designed to reduce the processing time of VHR earthquake images. 3) Speeding up technology is used to reduce the computing time with very limited accuracy loss.

To demonstrate the feasibility of the workflow, we employ three common datasets to test the proposed method. We examined the processing speed for each VHR image and compared our proposed method with other methods in terms of speed and accuracy. The results illustrate that the proposed workflow has a very high processing speed with promising accuracy compared with other state-of-the-art methods. Finally, the proposed workflow is employed in a real earthquake scene, and the results show the marked enhancement of images due to scene dehazing.

2. Dehazing Using ST-UNet

The attention-based neural network and transformer have been well studied and perform similarly to CNNs. Moreover, the complexity of the transformer is much lower than that of CNNs. Rapid and accurate processing is required to dehaze images obtained after an earthquake. We proposed the use of Swin Transformer UNet (ST-UNet) for this task.

2.1 Downsampling

A DNN first downsamples an input image three times to extract features at different scales from the inputting image. We assume that the input of the model is $I \in R^{h, w, c}$, where h is the height of the feature map, w is the width of the feature map, and c is the channel depth. Then, for each scale, a 3 × 3 convolutional layer is employed before downsampling, whose shape is $F_0 \quad R^{h, w, c}$. The initial feature map is downsampled to half its size ($F_1 \in R^{h/2, w/2, c}$) by a 3 × 3 convolutional layer whose padding is 1 and stride is 2. Then, in the second stage of downsampling, we apply the Swin Transformer blocks twice, where F_k is selected as the downsampling input. The downsampled stages are processed k times, and we select F_k as the deepest feature map for the subsequent upsampling. We double the channel depth c during each downsampling, i.e., $= 3 \times 2\ ^{-1}$.

2.2 Upsampling

The decoder is symmetrical to the encoder in its shape and also contains k stages in different scales. The upsampling method employs k–1 Swin Transformer modules and a 3 × 3 convolutional layer. For upsampling, a 3 × 3 transposed convolution is employed to increase the size of the feature map from $F_{k-1} \in R^{h/2,\ w/2,\ c}$ to $F_{k-1} \in R^{h,\ w,\ c/2}$. During the same upsampling, the depth of the feature map is also halved, i.e., it has the same kernel size as the encoder.

2.3 Skipping connection

A skipping connection is employed to connect the downsampling encoder to the upsampling decoder. The encoder $E \in R^{h,\ w,\ c}$ and the decoder $D \in R^{h,\ w,\ c}$ are connected, whose scales are of the same size; then, the DNN generates the concatenation $C^{h,\ w,2*c} = E + D$. A 3 × 3 convolutional filter is employed to return $C^{h,\ w,2*c}$ to $C^{h,\ w,c}$.

2.4 Swin Transformer blocks

The Swin Transformer[7] is extended by a traditional transformer and utilized in several tasks. Because the attention part in the transformer enables long-range information transfer in the model, the transformer is also applied in computer vision.[8] The major challenge is the image size, i.e., the size of an image (especially a VHR image) saved in the memory is very large. Therefore, a high computing cost exists if the brute force strategy used in CNNs is employed.

The Swin Transformer utilized in computer vision builds hierarchical feature maps by integrating image patches in a dense layer, but with low computation complexity with linear calculation by using a self-attention module based on a local window. As shown in Fig. 1, the Swin Transformer block is doubled in size during encoding and decoding, and layer normalization is employed before them. Residual connections are used to integrate pre-information in both processes. Common window-based multi-head self-attention and shifted window multi-head self-attention are also employed in the patches.[9]

3. Proposed Workflow

As mentioned earlier, the dehazing of VHR images has a much higher computing cost than common computer vision tasks because of the large image size, meaning that the model cannot import a VHR image to the DNN in a single time. Therefore, we design a workflow to solve this issue.

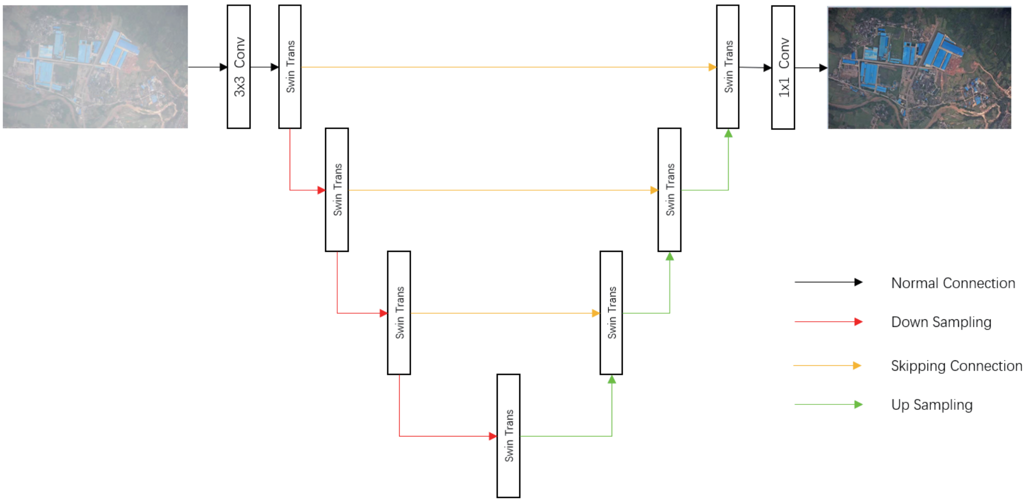

Fig. 1. (Color online) Architecture of ST-UNet.

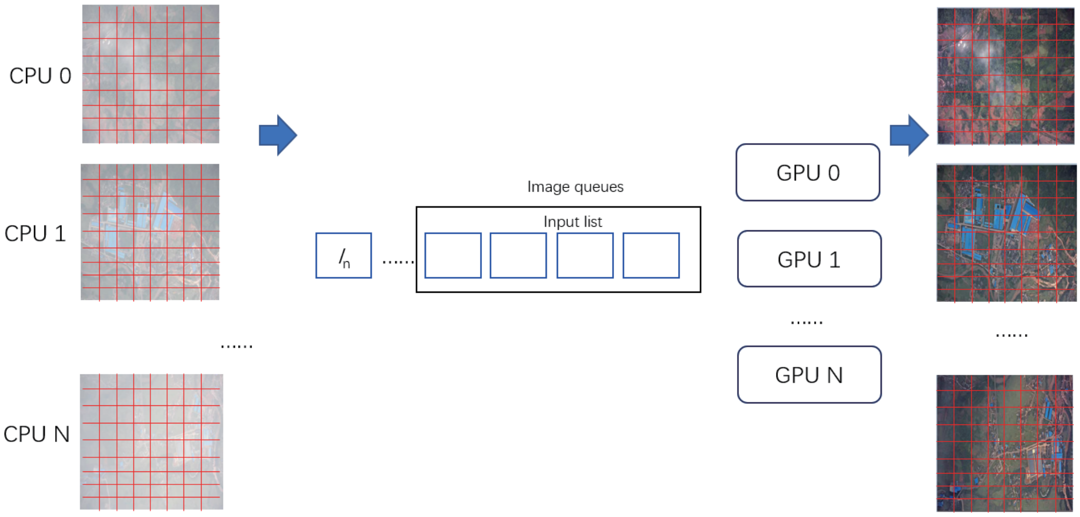

Fig. 2. (Color online) Workflow of predicting optimizing method.

As shown in Fig. 2, rather than importing a whole image into the model at one time, we first divide a whole image into small patches with unique image sub-IDs. As the prediction by DNNs requires much less GPU computation than training, parallel processing by increasing the batch size is applied to maximize the rate of utilization of computer components. Before the image is divided into small patches, an image decoding procedure is essential. However, decoding a VHR image is also time-consuming. As a result, we employ the multi-CPU to decode various VHR images in parallel at the same time and save them in the RAM. Then, a dividing module is employed to divide the decoded VHR images to small patches. On each patch, a unique ID is saved. Afterward, ST-UNet is employed to dehaze each patch, and the result is directly saved on the original image by replacing the initial pixels. Finally, the dehazed image is outputted to be saved. Moreover, tensorRTX is employed to increase the speed of the DNNs.

4. Experiment

4.1 Training details

An Intel I7 12700k is employed as the CPU with 64 GB DDR4 RAM, and two Titan RTX GPUs are employed as the platform. PyTorch is employed for the training and testing, and the two Titan RTX GPUs are used to train the model in the Ubuntu 18.04.1 LTS system. The initial learning rate is set to 0.001, and the Adam optimizer is used to accelerate the training. We train the network for 100 epochs with a batch size of 16.

4.2 Indicators used for quantitative evaluation

To quantitatively evaluate the dehazing effectiveness of different methods, we use the two most common evaluation indicators for the dehazing task, i.e., the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). PSNR indicates the error between the corresponding pixels, which is an error-sensitive image quality evaluation index. In general, a higher PSNR value means higher similarity between the dehazing results and the ground truth, and less image distortion. SSIM measures the image similarity from the three aspects of luminance, contrast, and structure. A higher SSIM value means that the results after dehazing are closer to the ground truth.

4.3 Establishment of dataset

To train the proposed ST-UNet for the dehazing task, it is necessary to generate a dataset. We generate hazed images using a non-uniform haze-adding algorithm,[9] and we employ two typical aerial view and haze-free datasets as basic datasets for training, i.e., AID30[10] and RSSCN8.[11]

AID30 is a remote sensing dataset open sourced by Wuhan University, which is used for image recognition. The images in this dataset have 0.5–0.8 m resolution and 600 × 600 image size. In total, 660 haze-free images are employed to synthesize the non-uniform haze-added training set. The RSSCN8 dataset is a 400 × 400 aerial view image published by Wuhan University. We select 164 haze-free images to serve as the validation set. To test the proposed images, we select 22 AID30 images and 18 Google Earth images.

4.4 Quantitative evaluation

Dehazing VHR images generally involves addressing heavy haze, non-uniform haze, and non-uniform illumination. We employ 64 challenging images to address these challenges, and we compare ST-UNet with other state-of-the art dehazing DNNs, i.e., DehazeNet,[5] DEFADE,[12] SID,[13] NLD, [14] CAP,[15] PFF-Net,[16] W-U-Net,[17] and FCTF-Net.[18] We set the input image size to 512 × 512 for all DNNs as shown in Fig. 3.

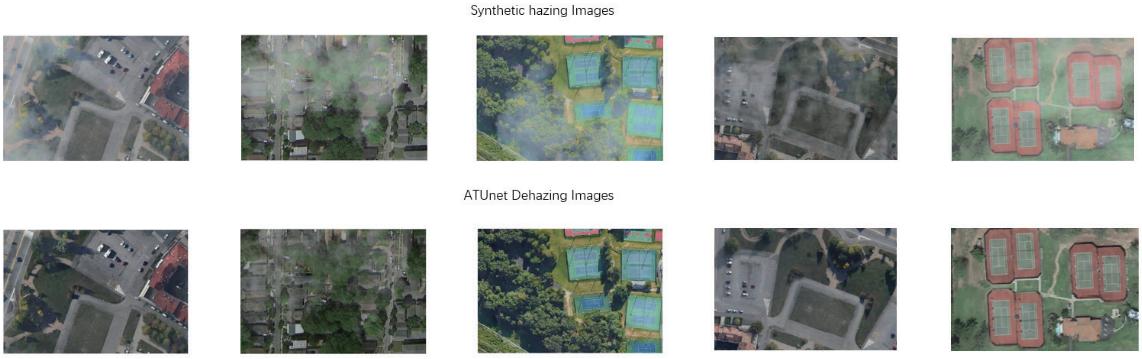

Fig. 3. (Color online) Visualization on testing dataset.

Table 1 shows a comparison of the results for the proposed ST-UNet with eight other state-of-the-art methods. DEFADE, CAP, SID, and NLD perform considerably worse than the other DNN-based methods for all indicators, i.e., speed, PSNR, and SSIM. These methods require over 0.5 s on average to process each image because they generally employ a CPU core for calculation. In addition, their PSNR (avg. 18) and SSIM (avg. 0.75) values are lower than those for the DNN-based methods. Therefore, they are difficult to employ in dehazing earthquake scenes because of their low speed and accuracy. In contrast, the DNN-based methods obtain very promising accuracies. For example, DehazeNet has PSNR and SSIM values of 19.1753 and 0.7886, respectively, and a processing time of only 0.12 s/image. PFF-Net, W-U-Net, and FCTF-Net have an average PSNR of about 23 and an SSIM of 0.88, which are much higher than the values for the traditional methods. However, they still require a very long time for processing (over 0.2 s/image), making them difficult to deploy in a real-time system. ST-UNet has very high values of PSNR (25.8508) and SSIM (0.9248) and has an average processing time of only 0.08 s/image. Moreover, when using half-float (OP), although the accuracy decreases slightly (PSNR of 25.4239 and SSIM of 0.9162), the average processing time is halved to 0.04 s/image, enabling very fast real-time post-earthquake image dehazing.

Having demonstrated the high performance of the proposed DNN, we next consider the relationship between the image size and the processing speed and accuracy. Table 2 shows the results of processing images of different sizes. When the image size used in the DNN is reduced

Table 1
Quantitative evaluation.

Method	Speed (s/image)	PSNR	SSIM
DehazeNet	0.12	19.1753	0.7886
DEFADE	0.45	17.2354	0.7521
SID	0.64	17.1688	0.7684
NLD	0.56	15.3500	0.7324
CAP	0.71	17.4218	0.7460
PFF-Net	0.28	21.7652	0.8468
W-U-Net	0.24	23.1535	0.8720
FCTF-Net	0.38	25.4378	0.8924
ST-UNet	0.08	25.8508	0.9248
ST-UNet (OP)	0.04	25.4239	0.9162

Table 2
Image size vs processing speed and accuracy.

Method	Speed (s/image)	PSNR	SSIM
ST-UNet (128 × 128)	0.007	15.1722	0.6924
ST-UNet (256 × 256)	0.02	19.6522	0.7823
ST-UNet (512 × 512)	0.04	25.4239	0.9162
ST-UNet (1024 × 1024)	0.16	24.4457	0.9062

from 512 × 512 to 128 × 128 and 256 × 256, PSNR significantly decreases to 15.1722 and 19.6522 and SSIM significantly decreases to 0.6924 and 0.7823, respectively. In contrast, when the size of the input image is increased to 1024 × 1024, the processing speed greatly decreases to 0.16 s/image and PSNR and SSIM also decrease to 24.4457 and 0.9062, respectively. Therefore, 512 × 512 is the most reasonable image size for this task.

4.5 Dehazing in real earthquake scene

Finally, we demonstrate the feasibility of the proposed ST-UNet for real post-earthquake scenes. We select 62 VHR images (5240 × 3840) obtained after the Ya’An Earthquake and process them with ST-UNet to calculate the average speed and evaluate the visualization results.

The visualization results are shown in Fig. 4. Significantly, the proposed ST-UNet obtains very promising dehazing results, even when images are taken in dense cloud. All the buildings

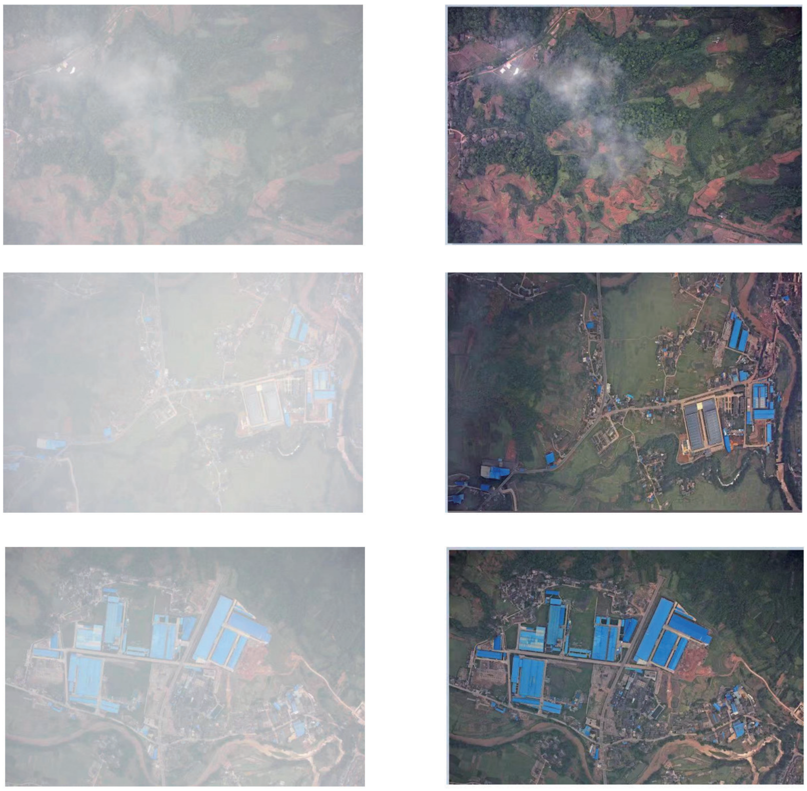

Fig. 4. (Color online) Visualization for real earthquake scenes.

and roads can be clearly recognized in a suitable color without abnormal illumination. Moreover, the processing speed in ST-UNet using the workflow is high (0.97 s/image), making post-earthquake evaluation very fast. Therefore, we consider that the proposed ST-UNet and workflow can be used in real-world tasks with outstanding processing speed and accuracy.

5. Conclusions

In this paper, we propose a dehazing workflow for VHR image processing that enables rapid real-time post-earthquake scene evaluation. The proposed ST-UNet achieves very high processing accuracy with significantly higher precision and processing speed than other state-of-the-art methods. We also explore the potential application of ST-UNet in real earthquake scene evaluation, and obtain high-quality and clear dehazed images with a processing time of only 0.97 s per VHR image. The proposed workflow takes advantage of the fast and robust dehazing process; therefore, will investigate processing in post-earthquake scenes covering larger areas in the near future.

References

1 C. O. Ancuti and C. Ancuti: IEEE Trans. Image Process. **22** (2013) 3271. https://doi.org/10.1109/TIP.2013.2262284

2 C. O. Ancuti, C. Ancuti, and P. Bekaert: Proc. 17th IEEE Int. Conf. Image Process (2010) 3541–3544. https://doi.org/10.1109/ICIP.2010.5651263

3 Y. Liu, H. Li, and M. Wang: IEEE Access **5** (2017) 8890. https://doi.org/10.1109/ACCESS.2017.2710305

4 M. Ju, C. Ding, D. Zhang, and Y. J. Guo: IEEE Trans. Circuits Syst. Video Technol. **29** (2017) 2349. https://doi.org/10.1109/TCSVT.2018.2869594

5 B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao: IEEE Trans. Image Proc. **25** (2016) 5187. https://doi.org/10.1109/TIP.2016.2598681

6 Y. Dong, Y. Liu, H. Zhang, S. Chen, and Y. Qiao: Proc. AAAI Conf. Arti. Intel. (2020) 10729–10736. https://doi.org/10.1609/aaai.v34i07.6701

7 Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo: Proc. IEEE/CVF Int. Conf. Computer Vision (2021) 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

8 Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu: Proc. IEEE/CVF Int. Conf. Computer Vision and Pattern Recognition (2022) 3202–3211. https://doi.org/10.48550/arXiv.2106.13230

9 B. Jiang, G. Chen, J. Wang, H. Ma, L. Wang, Y. Wang, and X. Chen: Remote Sens. **13** (2021) 4443. https://doi.org/10.3390/rs13214443

10 G. S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu: IEEE Trans. Geosci. Remote Sens. **55** (2017) 3965. https://doi.org/10.1109/TGRS.2017.2685945

11 Z. Qin, L. Ni, Z. Tong, and W. Qian: IEEE Geosci. Remote Sens. **55** (2015) 2321. https://doi.org/10.1109/LGRS.2015.2475299

12 L. K. Choi, J. You, and A. C. Bovik: IEEE Trans. Image Process. **12** (2015) 2321. https://doi.org/10.1109/TIP.2015.2456502

13 J. Long, Z. Shi, W. Tang, and C. Zhang: IEEE Geosci. Remote Sens. Lett. **11** (2014) 59. https://doi.org/10.1109/LGRS.2013.2245857

14 D. Berman, T. Treibitz, and S. Avidan: Proc. IEEE/CVF Int. Conf. Computer Vision and Pattern Recognition (2016) 1674–1682. https://doi.org/10.1109/CVPR.2016.185

15 Q. Zhu, J. Mai, and L. Shao: IEEE Trans. Image Process. **24** (2015) 3522. https://doi.org/10.1109/TIP.2015.2446191

16 K. Mei, A. Jiang, J. Li, and M. Wang: Proc. Asian Conf. Computer Vision (2018) 203–215. https://doi.org/10.1007/978-3-030-20887-5_13

17 D. Hong, C. Jocelyn, Y. Naoto, H. Uta, H. Wieke, and X. Zhu: IGARSS 2019 IEEE Int. Geosci. Rem. Sens. Sym. (2019) 373–376. https://doi.org/10.1109/IGARSS.2019.8899865

18 Y. Li and X. Chen: IEEE Geosci. Remote Sens. Lett. **18** (2020) 1751. https://doi.org/10.1109/LGRS.2020.3006533

About the Authors

Yuxin Bian is a senior student in the School of Computer Science at Macau University of Science and Technology and is expected to graduate in 2023. His research interests are computer vision, remote sensing, and deep learning.

Enguang Zhang received his master's degree in mechanical engineering from South China University of Technology in 2013. Since September 2021, he has been studying toward his Ph.D. degree in the School of Computer Science and Engineering, Macau University of Science and Technology. Currently, he is an associate professor in the School of Mechanical Engineering, Zhuhai College of Science and Technology. His main research interests are machine vision and deep learning.

Jiayan Wang is a senior in the School of Computer Science, Macau University of Science and Technology. His research interests are biological sciences, remote sensing, and deep learning.

Rixin Xie is a senior student in the School of Computer Science at Macau University of Science and Technology and is expected to graduate in 2023. His main research interests are computer vision and deep learning.

Shenlu Jiang received his Ph.D. degree in electronic and electrical engineering from the School of Electrical and Electronics Engineering, Sungkyunkwan University, Seoul, South Korea, in 2020. From October 2018 to September 2020, he was a research assistant in the Department of LSGI in Hong Kong Polytechnic University and the Chinese University of Hong Kong. From October 2020 to April 2021, he was a postdoctoral researcher in INRIA. Currently, he is an assistant professor in the School of Computer Science and Engineering, Macau University of Science and Technology. His research interests are robot vision, remote sensing, and deep learning.