A Correlation Context-Driven Method for Sea Fog Detection in Meteorological Satellite Imagery

Yixiang Huang, Ming Wu, Jun Guo, Chuang Zhang, *Member, IEEE*, and Mengqiu Xu

***Abstract*— Sea fog detection is a challenging and essential issue in satellite remote sensing. Although conventional threshold methods and deep learning methods can achieve pixel-level classification, it is difficult to distinguish ambiguous boundaries and thin structures from the background. Considering the correlations between neighbor pixels and the affinities between superpixels, a correlation context-driven method for sea fog detection is proposed in this letter, which mainly consists of a two-stage superpixel-based fully convolutional network (SFCNet), named SFCNet. A fully connected Conditional Random Field (CRF) is utilized to model the dependencies between pixels. To alleviate the problem of high cloud occlusion, an attentive Generative Adversarial Network (GAN) is implemented for image enhancement by exploiting contextual information. Experimental results demonstrate that our proposed method achieves 91.65% mIoU and obtains more refined segmentation results, performing well in detecting fogs in small, broken bits and weak contrast thin structures, as well as detects more obscured parts.**

***Index Terms*— Deep learning, satellite imagery, sea fog detection, superpixel.**

I. Introduction

SEA fog is a common and disastrous weather phenomenon for marine transportation and navigation. It reduces horizontal visibility to less than 1 km or even much lower, thereby greatly threatening the safety of maritime activities and seriously affecting the operation of fishery industry. Hence, sea fog detection is a highly demanding and significant task, which can provide accurate and real-time weather forecast guidance.

In this work, sea fog detection is considered as a segmentation task, which requires pixel-level classification. Based on remote sensing data of meteorological satellites, traditional sea fog detection methods mainly use hand-crafted features and empirical thresholds to identify whether a pixel is fog or nonfog [1], [2]. The threshold-based methods are stable and simple, but not flexible enough to deal with complex situations by the combination of thresholds. Besides, the structures and textures of sea fog are neglected. Wang *et al.* [3] adopted a texture filter for 3×3 pixels, but the extracted features are rough.

Manuscript received March 31, 2021; accepted July 2, 2021. Date of publication July 26, 2021; date of current version December 15, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFF0303300 and Grant 2019YFF0303302 and in part by the Ministry of Education (MoE)-China Mobile Communications Group Company Ltd., (CMCC) Artificial Intelligence Project under Grant MCM20190701. *(Corresponding author: Ming Wu.)*

The authors are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: wuming@bupt.edu.cn).

Digital Object Identifier 10.1109/LGRS.2021.3095731

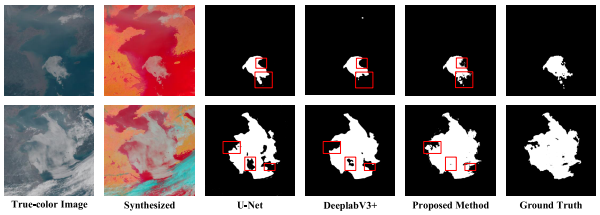

Fig. 1. Results of segmentation methods on remote sensing images with ambiguous boundaries and thin structures, which remains a hard case in sea fog detection. Note that these segmentation results by U-Net and DeeplabV3+ have failed in detecting fogs in small, broken bits, and weak contrast thin structures.

In recent years, deep learning has been gradually applied to the field of geoscience and remote sensing [4], [5]. The deep learning methods can automatically extract features of different levels and learn from context. Many convolutional neural network (CNN) structures for semantic segmentation, such as U-Net [6] and Deeplab V3+ [7], show impressive gains in remote sensing image interpretation. For example, Chunyang *et al.* [8] employed the U-Net model for sea fog detection in MODIS multispectral images. However, U-Net is a basic network and has no specific design for the characteristics of sea fog. Most of the existing methods cannot work well on ambiguous boundaries and thin structures (see Fig. 1), which still remains a hard case. Thin fog is semitransparent and has a weak contrast with the background, so it can be easily missed.

Despite getting progress, the aforementioned methods are still not reliable enough due to their operations limited at the pixel level. Therefore, superpixel-based methods are proposed. In computer vision, superpixels over-segment an image into perceptually meaningful subregions, containing adjacent pixels that are similar in appearance. Xie *et al.* [9] improved the traditional superpixel segmentation method simple linear iterative clustering (SLIC) [10] and designed a superpixel-level cloud detection framework using CNN. Liu *et al.* [11] built a superpixel-level remote sensing database and proposed hierarchical fusion CNN (HFCNN) as a classifier to take full advantage of low-level features like color and texture. These superpixel-based methods used superpixels as basic units and achieved excellent results compared with pixel-level methods. Nevertheless, global information may be lost from remote sensing images, and a single superpixel patch contains insufficient semantic features.

Note that sea fog exhibits self-similarity and spatial continuous distribution. Human generally identifies sea fog roughly by analyzing both local and global features of remote sensing images, rather than pixel by pixel. We argue that

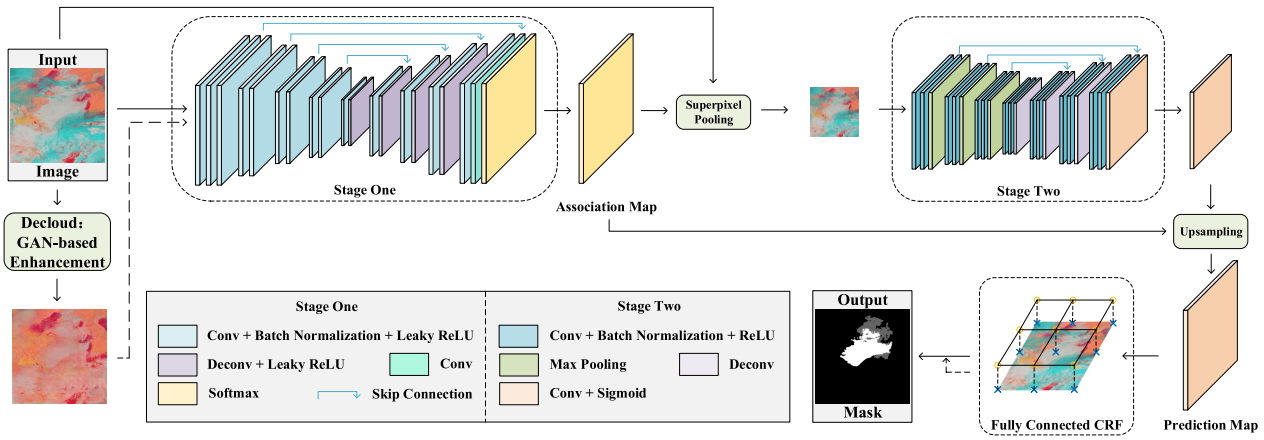

Fig. 2. Overview of the proposed method. The decloud model represents the part of GAN-based image enhancement. First, it takes a composite image as input and outputs an enhanced one after removing high clouds, both of which are fed into SFCNet, respectively. Then, SFCNet predicts a pixel-to-superpixel association map. Based on this, superpixel feature maps are computed and used for superpixel-level dense prediction. Subsequently, the pixel-level prediction map is upsampled with the association map. It is followed by a fully connected CRF for optimization to get more refined segmentation results. Finally, the prediction mask is obtained by both the results of sea fog detection before/after high cloud removal.

correlation matters; the intrinsic pixel-wise relationship and superpixel-wise association can be exploited, which is conductive to our task. Hence, we propose a correlation context-driven method for sea fog detection in satellite imagery. First, a pixel-to-superpixel association map which assigns each pixel to its surrounding grids is obtained via a simple, yet effective fully convolutional network, following up on [12]. Then, superpixel feature maps are computed and fed into the downstream segmentation network for superpixel-level dense prediction. To strengthen the pixel-wise semantic association and capture contextual information, a fully connected conditional random field (CRF) is used to model the dependencies between pixels based on visual features and position features.

Generally, sea fog and all kinds of clouds are different in forms and cloud heights. High clouds may occlude some parts of sea fog, which hinders the expression of local context and destroys the spatial coherence. Therefore, we additionally employ an enhancement model based on an attentive generative adversarial network (GAN) [13] to remove high clouds for more precise prediction of sea fog.

In summary, the main contributions of this letter are as follows.

1) We propose a correlation context-driven sea fog detection method by exploiting pixel-superpixel relationships and pixel-wise dependencies. Experimental results show that our model obtains better detection effects for sea fog in small, broken bits and weak contrast thin structures, preserving distinct boundaries and fine details. Compared with the existing methods, ours yields more refined segmentation results.
2) In order to alleviate the problem of cloud occlusion and enhance contextual information, we employ a GAN-based model for high cloud removal. An attention mechanism is introduced to focus on high cloud area and its surrounding structures for better restoration. With high cloud removed, the obscured part of sea fog can be predicted.

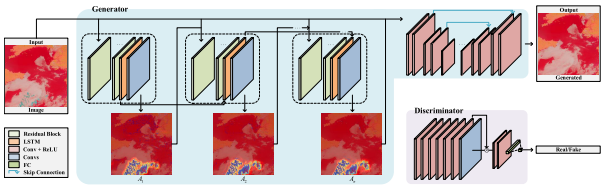

Fig. 3. Detailed components of the GAN model. Each block in the recurrent network composes of residual blocks for feature extraction, a convolutional long short-term memory (LSTM) unit, and convolutional layers for producing attention maps. The attention map is a 2-D matrix ranging from 0 to 1, where the greater the value, the more the attention it indicates. At each time step, the current attention map is concatenated with the input image and then fed into the next block. In the training phase, the initial value is set to 0.5.

II. Methodology

The whole framework can be roughly divided into two parts. In Section II-A, we introduce the first part which focuses on the problem of high cloud occlusion. An attentive GAN model is trained for high cloud removal and outputs an enhanced image as auxiliary data for further sea fog detection. In Section II-B, we introduce the second part, its goal is to achieve superpixel-based sea fog detection by superpixel-based fully convolutional network (SFCNet), and utilize a fully connected CRF to get more refined segmentation results. A brief overview of our proposed method is shown in Fig. 2.

A. GAN-Based Enhancement

The high-level thin clouds in remote sensing images are translucent and flocculent to some degree. We notice that humans could roughly "see" through the clouds by synthesizing contextual information, in cases where the clouds do not completely obstruct the scene. Some parts of the thin clouds, especially at the periphery and semitransparent regions, convey some information about the background, which can be revealed and exploited. Thus, we utilize a GAN for high cloud removal. The overall architecture of GAN is shown in Fig. 3.

It consists of two networks: a generative network (G) and a discriminative network (D), which compete with each other in a Minimax game [14]. G attempts to generate an image

as real as possible and free from thin clouds. D attempts to distinguish between real and generated images. The generative adversarial loss can be formulated as follows:

$$\min_G \max_D \begin{matrix} E_{R\sim p_{\text{clean}}}[\log(D(R))] \\ +E_{I\sim p_{\text{cloud}}}[\log(1-D(G(I)))] \end{matrix} \quad (1)$$

where R is a clean natural sample. I is a clouded sample, which is the input of G. After several rounds of the Minimax game between G and D, the distribution of $G(I)$ will be similar to p_{clean} and D will be unable to distinguish between $G(I)$ and R. Our goal is to obtain a "realistic" cloud-free image generated by G.

1) Generative Network: To focus on high cloud area and its surrounding structures, we introduce the attention mechanism in [13]. Consequently, G consists of two subnetworks: a recurrent network and an autoencoder. The recurrent network aims to find regions in the input image that need attention and learns the attention map progressively. The loss function is defined as the mean squared error (MSE) between the generated attention map A and the binary mask M. The autoencoder attempts to generate a cloud-free image, which is the output of the whole network. It takes the concatenation of the input image and the final attention map from the recurrent network as input. To capture more contextual information from different scales, we adopt the multiscale loss. Overall, the generative loss is expressed as

$$L_G = L_A(\{A\}, M) + L_M(\{S\}, \{T\}) + \theta log(1 - D(O)) \quad (2)$$

where T is the corresponding ground truth of S with the same scale. O is the output of G. We give a weight $\theta = 0.01$ to the GAN loss for proportional balance.

2) Discriminative Network: The discriminative network is a CNN classifier to verify whether the input image is real or fake. In order to guide the discriminator to focus on the attention area, we add an interior convolutional layer to output a mask and multiply it with previously extracted features. We define a loss function based on the mask and the attention map as follows:

$$L_I(O, R, A_n) = L_{\text{MSE}}(D_I(O), A_n) + L_{\text{MSE}}(D_I(R), 0) \quad (3)$$

where A_n is the final attention map and $\mathbf{0}$ is a map with full 0 values. Overall, the discriminative loss is written as

$$\begin{aligned} L_D(O, R, A_n) = &-\log(D(R)) - \log(1 - D(O)) \\ &+L_I(O, R, A_n). \end{aligned} \quad (4)$$

B. SFCNet

1) Superpixel Segmentation: Given an image of size $H \times W$, the first step is to partition it into regular grids of size $h \times w$ as initial superpixels (i.e., seeds) and then learn a mapping which assigns each pixel p to one of the seeds s. In practice, it is not cost-effective to compute all pixel–superpixel pairs. Given a pixel p, a set of 3×3 surrounding grid cells S_p are considered, instead. Note that the constraint of search scope contributes to the compactness for spatial coherence in the local region. Therefore, the mapping can be expressed as a tensor $A \in \mathbf{R}^{H\times W\times |S_p|}$, where $|S_p| = 9$. Here, let $a_s(p)$ be the probability that p is assigned to $s \in S_p$, such that $\sum_{s\in S_p} a_s(p) = 1$.

2) Superpixel Pooling: Let $f(p)$ represent the feature vector of p. The position coordinates of each pixel p can be written as $P = [x, y]^T$. According to the association map A, we calculate the center of each superpixel s, $c_s = (f_s, l_s)$, where f_s and l_s are the feature vector and the location vector of each superpixel s, respectively

$$f_s = \frac{\sum_{p:s\in S_p} f(p)\cdot a_s(p)}{\sum_{p:s\in S_p} a_s(p)}, \quad l_s = \frac{\sum_{p:s\in S_p} P\cdot a_s(p)}{\sum_{p:s\in S_p} a_s(p)}. \quad (5)$$

The feature vector and the location vector of each pixel p can be reconstructed by

$$f'(p) = \sum_{s\in S_p} f_s \cdot a_s(p), \quad p' = \sum_{s\in S_p} l_s \cdot a_s(p). \quad (6)$$

$l2$-norm is used as the distance measure and the superpixel loss is defined as follows:

$$L_{\text{superpixel}} = \sum_p ||f(p) - f'(p)||_2 + k||p - p'||_2 \quad (7)$$

where the former term tends to group similar pixels together and the latter one facilitates spatial compactness. k is a balancing coefficient and set to 0.002 by default.

3) Dense Prediction: Subsequently, the computed superpixel feature maps are fed into a downstream segmentation network for superpixel-level dense prediction. Here, we adopt the focal loss

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise.} \end{cases} \quad (8)$$

$$L_{\text{focal}}(p_t) = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (9)$$

where α and γ are set to 0.2 and 2, respectively, to balance the positive and negative samples, as well as put more focus on hard, misclassified examples. Consequently, the overall loss function becomes

$$L = L_{\text{focal}}(p_t - \hat{p}_t) + \lambda L_{\text{superpixel}}(A) \quad (10)$$

where λ is a weight to balance the two terms. The predicted probability map is upsampled via the association map to get the final pixel-level prediction.

4) Network Architecture: SFCNet consists of two-stage fully convolutional networks: the former for superpixel segmentation and the latter for dense prediction. In the first stage, the network is a standard encoder–decoder architecture with skip connections. The encoder automatically extracts features and produces high-level embeddings via convolutional layers. Then the decoder upsamples the feature maps concatenated with the corresponding size ones from the encoder via deconvolution. All convolutional layers are followed by batch normalization and leaky rectified linear unit (ReLU) activation except the last prediction layer, where softmax is used. In the second stage, we adopt a similar U-Net structure as segmentation network, since it combines high-level information (for semantic comprehension) and low-level information (for precise localization).

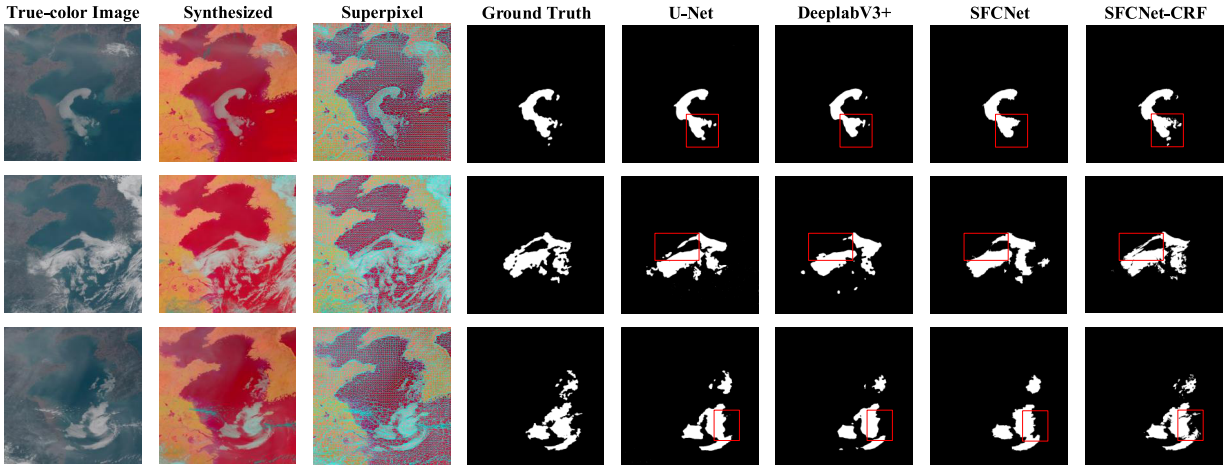

Fig. 4. Results of sea fog detection by all methods. From left to right: true-color image, synthetic image, superpixel image, ground truth, U-Net, DeeplabV3+, SFCNet, and SFCNet with CRF.

5) CRF Refinement: A fully connected CRF is used to model the relationships between pixels for further refining pixel-level prediction. In the CRF model, the Gibbs energy is the objective function

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i,j} \psi_p(x_i, x_j) \tag{11}$$

where $\psi_u(x_i)$ is the unary potential energy of the pixel i belonging to the category x_i. $\psi_p(x_i, x_j)$ is the pairwise potential energy of the pixels i and j belonging to the categories x_i and x_j, respectively. The latter term encourages similar pixels to be predicted as the same category, while pixels with sharp contrast are prone to be assigned different labels. Overall, the combined energy is minimized by Mean-field inference [15] to iteratively optimize the results. As thus, our model achieves more precise segmentation.

III. Experiments

A. Data

We choose Himawari-8 satellite to obtain remote sensing data due to its new payload called Advanced Himawari Imager (AHI) for dedicated meteorological missions. It has the advantages of high radiometric, spatial, spectral, and temporal resolution, which can provide comprehensive and detailed data support for our research. We focus on parts of the Yellow Sea and Bohai Sea, with the longitude ranging from $117.50°E$ to $128.76°E$ and the latitude ranging from $29.74°N$ to $41.00°N$. In this letter, Himawari-8 Standard Data (HSD8) is used as experimental data, including 193 daytime sea fog samples from 2017 to 2019. The dataset has 155 training images and 38 test images of size 1024×1024. HSD8 contains 16 observation bands distributed in visible, near-infrared, and infrared channels. According to the detection category and physical meaning of each band, we select three bands which are closely related to the recognition of sea fog, namely B03, B04, and B14, to synthesize nature-color images for further experiments. Compared with the true-color (RGB) images, sea fog in the synthesized images can be better distinguished from cloud groups visually, as shown in Fig. 4.

To train the image enhancement model, we first manually extract high cloud covers in the above images, then randomly composite them with clean background samples in red, green, blue, alpha (RGBA) color space, preserving the value of transparency with an extended alpha channel. Consequently, the composited images (i.e., input) and the corresponding backgrounds (i.e., ground truth) make up 5625 pairs of images for GAN, from which the binary masks for attention can be derived by contrast.

B. Implement Details

1) High Cloud Removal: In the GAN model, the generative network and the discriminative network are trained in an adversarial manner. We use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.001. The model is trained for 200 iterations with batch size 4. The learning rate for both is set to 1×10^5. In the recurrent network, the number of iterations is set to 4, considering a tradeoff between the quality of attention map and memory allocation.

2) Sea Fog Detection: During the training phase, we first pretrain the superpixel network and then update it with the rest of network simultaneously. We apply some augmentation techniques: randomly flip and rotate the images by 90° as well as crop them to the sizes between three-quarters and the whole and then resize to the original size. The input size of the network is 1024×1024 and we predict superpixels with 16×16 grid size to perform $16\times$ up/down sampling. Adam is adopted as the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The model is trained for 200 iterations with batch size 4. The initial learning rate is set to 5×10^5 and reduced by half after 150 and 180 epochs, respectively. In the post-processing, the Mean-field inference is iterated ten times.

C. Result Analysis

To validate the SFCNet, we compare it with two pixel-level segmentation methods both quantitatively and qualitatively on the test set. For a fair comparison, all the methods were trained in the same manner. The accuracy of semantic segmentation is measured by the metric, mean intersection over the union (mIoU). As shown in the table, our proposed method shows

TABLE I
QUANTITATIVE RESULTS OF DIFFERENT METHODS

Methods	mIoU(%)
U-Net [6]	82.73
DeeplabV3+ [7]	86.76
SFCNet	87.72
SFCNet-CRF	90.20
SFCNet-CRF-Decloud	**91.65**

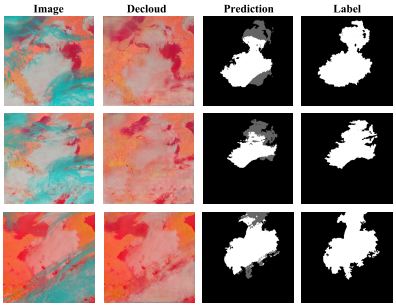

Fig. 5. Results of high cloud removal. The blue area is the high cloud that we want to remove, which hinders the expression of local context and destroys the spatial coherence of sea fog. The white/gray areas are sea fog detected before/after high cloud removal, respectively. Note that the gray areas are the extra detected parts.

better performance compared with U-Net and DeeplabV3+. SFCNet achieves nearly five points higher mIoU than that of U-Net, which verifies the effectiveness of the superpixel-based method. By applying CRF, mIoU is improved by 2.48%. After high cloud removal, mIoU is improved by 1.45%, which demonstrates the effectiveness of GAN-based enhancement.

Fig. 4 shows the results of sea fog detection. It is obvious that most of sea fog can be detected by all methods. However, U-Net and DeeplabV3+ cannot work well on some ambiguous boundaries and thin structures, while our proposed method performs better in detecting fogs in small, broken bits and weak contrast thin fog areas. We also see that fully connected CRF can modify and restore some misjudged points or segments, thus yielding more refined segmentation results in preserving object boundaries and fine details.

Fig. 5 shows the results of high cloud removal and further sea fog detection. It can be seen that high cloud removal visualizes the obscured part of sea fog to be detected by exploiting contextual information, which validates the decloud effect on contextual expression. In this way, the high cloud-free images can be used as auxiliary data that makes sea fog detection more intelligent.

IV. CONCLUSION

In this letter, an automatic and effective correlation context-driven method for sea fog detection in satellite imagery has been proposed. Quantitative and qualitative results show that the proposed method achieves the best mIoU of 91.65% and obtains more refined segmentation results compared with other methods. From the visual effect, our proposed method performs well in detecting fogs in small, broken bits and weak contrast thin structures. Additionally, experimental results demonstrate that the GAN-based image enhancement model can implement high cloud removal, which helps to further detect the obscured part of sea fog. In future work, we will apply the proposed method to sea fog detection in remote sensing images from different satellites to find new enhancements.

REFERENCES

[1] J. Lulu and W. Ming, "Application of fog monitoring with FY-3A data," *Remote Sens. Technol. Appl.*, vol. 26, no. 4, pp. 489–495, 2011.
[2] X. Wu and S. Li, "Automatic sea fog detection over Chinese adjacent oceans using Terra/MODIS data," *Int. J. Remote Sens.*, vol. 35, no. 21, pp. 7430–7457, Nov. 2014.
[3] Y. Wang, S. Gao, G. Fu, J. Sun, and S. Zhang, "Assimilating MTSAT-derived humidity in nowcasting sea fog over the Yellow Sea," *Weather Forecasting*, vol. 29, no. 2, pp. 205–225, Apr. 2014.
[4] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 197–212, Apr. 2019.
[5] N. Davari, G. Akbarizadeh, and E. Mashhour, "Intelligent diagnosis of incipient fault in power distribution lines based on corona detection in UV-visible videos," *IEEE Trans. Power Del.*, early access, Dec. 21, 2020, doi: 10.1109/TPWRD.2020.3046161.
[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention* Cham, Switzerland: Springer, 2015, pp. 234–241.
[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, Sep. 2018, pp. 833–851.
[8] Z. Chunyang, W. Jianhua, L. Shanwei, S. Hui, and X. Yanfang, "Sea fog detection using U-Net deep learning model based on MODIS data," in *Proc. 10th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Sep. 2019, pp. 1–5.
[9] F. Xie, M. Shi, Z. Shi, J. Yin, and D. Zhao, "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3631–3640, Aug. 2017.
[10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[11] H. Liu, D. Zeng, and Q. Tian, "Super-pixel cloud detection using hierarchical fusion CNN," in *Proc. IEEE 4th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2018, pp. 1–6.
[12] F. Yang, Q. Sun, H. Jin, and Z. Zhou, "Superpixel segmentation with fully convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13961–13970.
[13] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2482–2491.
[14] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[15] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. 25th Annu. Conf. Neural Inf. Process. Syst.*, 2011, pp. 109–117.