

# Coarse-to-fine Task-driven Inpainting for Geoscience Images

Huiming Sun<sup>†</sup>, Jin Ma<sup>†</sup>, Qing Guo, Qin Zou, Shaoyue Song, Yuewei Lin\*, Hongkai Yu\*

**Abstract**—The processing and recognition of geoscience images have wide applications. Most of existing researches focus on understanding the high-quality geoscience images by assuming that all the images are clear. However, in many real-world cases, the geoscience images might contain occlusions during the image acquisition. This problem actually implies the image inpainting problem in computer vision and multimedia. As far as we know, all the existing image inpainting algorithms learn to repair the occluded regions for a better visualization quality, they are excellent for natural images but not good enough for geoscience images, and they never consider the following gescience task when developing inpainting methods. This paper aims to repair the occluded regions for a better geoscience task performance and advanced visualization quality simultaneously, without changing the current deployed deep learning based geoscience models. Because of the complex context of geoscience images, we propose a coarse-to-fine encoder-decoder network with the help of designed coarse-to-fine adversarial context discriminators to reconstruct the occluded image regions. Due to the limited data of geoscience images, we propose a MaskMix based data augmentation method, which augments inpainting masks instead of augmenting original images, to exploit the limited geoscience image data. The experimental results on three public geoscience datasets for remote sensing scene recognition, cross-view geolocation and semantic segmentation tasks respectively show the effectiveness and accuracy of the proposed method. The code is available at: <https://github.com/HMS97/Task-driven-Inpainting>.

**Index Terms**—image inpainting, geoscience images, coarse-to-fine, task-driven

## I. INTRODUCTION

THE geoscience images have various representations, e.g., street-view and aerial-view images. Geoscience image processing is an inter-disciplinary research with wide applications in computer vision and multimedia, such as remote sensing scene recognition [1], [2], cross-view geolocation in urban environments [3], [4], change detection [5], [6], hyper-spectral classification [7]–[9], satellite-view object detection [10], [11], image captioning based remote understanding [12], [13], semantic segmentation [14], [15] and etc.

Huiming Sun, Jin Ma, and Hongkai Yu are with Cleveland State University, Cleveland, OH, 44115, USA. Qing Guo is with the Centre for Frontier AI Research (CFAR) and Agency for Science, Technology and Research (A\*STAR), Singapore, and the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A\*STAR), Singapore. Qin Zou is with Wuhan University, Wuhan, China. Shaoyue Song is with Beijing University of Technology, Beijing, China. Yuewei Lin is with Brookhaven National Laboratory, Upton, NY, 11973, USA. This work was supported by NSF 2215388.

<sup>†</sup> indicates the co-first authors. \* Corresponding authors: Yuewei Lin (e-mail: [ywlin@bnl.gov](mailto:ywlin@bnl.gov)) and Hongkai Yu (e-mail: [h.yu19@csuohio.edu](mailto:h.yu19@csuohio.edu)).

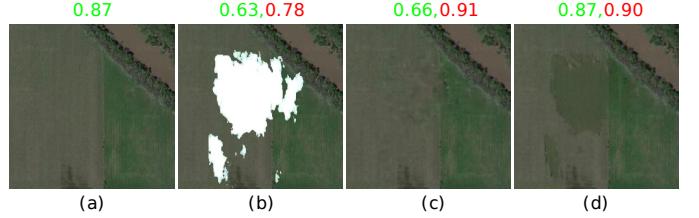


Fig. 1. Illustration of the task-driven inpainting problem for geoscience images, taking the remote sensing scene recognition/classification task as an example: (a) a clean satellite image, (b) occluded image, (c) reconstruction by the image inpainting method CSA [18], (d) reconstruction by the proposed inpainting method. Green and red colored numbers indicate the classification confidence of the correct class by a remote sensing scene recognition model (pre-trained on clean images) and the image quality SSIM of reconstruction respectively.

Most of the existing researches in this area assume that all the obtained geoscience images are clear without occlusions. However, in many real-world cases, the geoscience images might contain occlusions during the image acquisition. For example, some regions of a street-view image might be occluded by a passing pedestrian or car close to the camera [16], and an aerial-view image by UAV (Unmanned Aerial Vehicle) might be partially occluded by a tall tower or a kite, and an aerial-view image by satellite might be occluded by thick cloud(s) to some extents [17]. The occlusion challenge could result in the significant performance drop when using deployed deep learning based geoscience models pre-trained on clean geoscience images. One way to relieve the challenge is to recover the occluded regions by using image inpainting methods. Then, the current deep geoscience models could be still functional without the need of any changes if we could well reconstruct the occluded image regions. This paper proposes a new learnable pre-processing (inpainting) to overcome the occlusion challenge without changing the deployed deep learning based geoscience model.

As far as we know, the inpainting problem that focuses on geoscience images has not been systemically studied before. The existing image inpainting methods [16], [18]–[23] learn to recover the occluded regions for a better visualization quality, which are excellent for natural images but not good enough for geoscience images because they ignore the geoscience related tasks and fail to consider the domain speciality. For example, as shown in Fig. 1, the reconstruction visualization quality (SSIM) metric itself cannot fully represent the advanced geoscience task performance, since a higher SSIM score might not necessarily achieve better classification confidence in the remote sensing scene recognition/classification task.

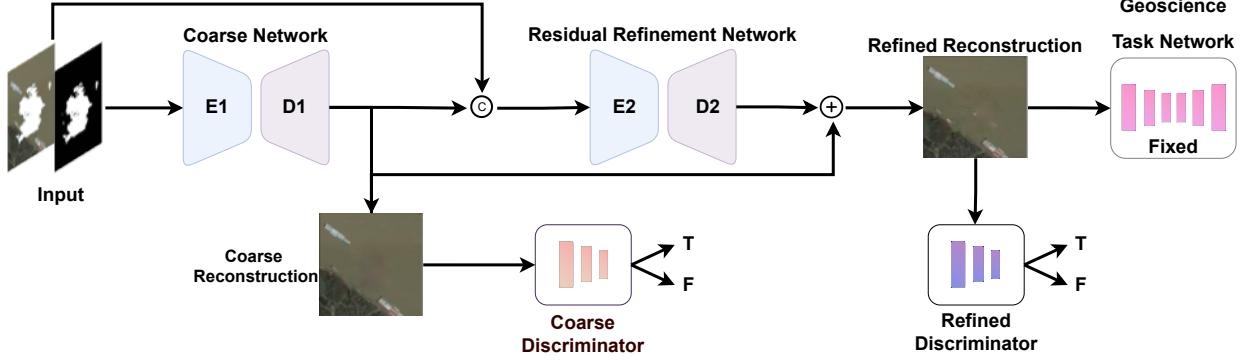


Fig. 2. Overview of the proposed image inpainting network for geoscience images. From the Coarse Network to Residual Refinement Network with two adversarial context discriminators, we learn to reconstruct occluded regions in a coarse-to-fine way. Note that  $\odot$  is concatenation and  $\oplus$  is element-wise summation.

Therefore, we introduce the task-driven inpainting problem for geoscience images in this paper.

There are several domain specialities for the task-driven image inpainting problem of geoscience images. First, the reconstruction objective is to largely improve the geoscience task performance with relatively high image quality, without changing the existing deep learning based geoscience task network pretrained on clean images. Second, the context of geoscience image is more complex than nature image without prior knowledge. Taking a face image as an example, if one eye is occluded, the deep learning based image inpainting model still knows the occluded region should be an eye there, because we have the prior knowledge of human face. Third, the dataset of geoscience images is typically much smaller than the regular nature images. For example, the Place2 [24] dataset that is frequently used for image inpainting has 10 million nature images. However, the geoscience images are relatively expensive to collect, so that many geoscience image datasets [1], [12], [25] only have thousands of geoscience images.

We have made efforts in this paper to solve the challenges of the above domain specialities. In this paper, we design a deep learning based image inpainting framework to embed the geoscience task network so that the reconstructed images could align with the geoscience task network. The geoscience task network can be replaced accordingly so as to fit different geoscience tasks, making the designed framework very flexible. Due to the above mentioned challenges, it might be difficult to simply learn a reliable deep learning based model within one stage to deal with the complex context of geoscience images, so we propose a coarse-to-fine encoder-decoder network to reconstruct the occluded regions with coarse-to-fine adversarial context discriminators. Due to the limited data of geoscience images, we design a MaskMix based data augmentation method to improve model robustness and overcome unforeseen corruptions by mixing different augmented random occlusion masks during the model training. We expect that the inter-disciplinary research proposed in this paper could particularly benefit the geoscience image processing and recognition. In summary, the contributions of this paper are as the following.

- To the best of our knowledge, this paper is the first deep learning work for the task-driven inpainting problem of geoscience images. The reconstruction goal is to largely improve the geoscience task performance with relatively high image quality without changing the existing pretrained deep geoscience task model, however all the existing inpainting methods only focus on improving image quality.
- Due to the complex context in geoscience images, this paper proposes a deep coarse-to-fine encoder-decoder network to reconstruct the occluded image regions in two stages, with the help of designed coarse-to-fine adversarial context discriminators.
- Due to the limited training data in geoscience images, this paper proposes a MaskMix based data augmentation method, which augments inpainting masks instead of augmenting original images, to improve model robustness and overcome unforeseen corruptions.

In the following of this paper, Section II reviews the related work. Section III explains the proposed method. Experiment setting and results are described in Section IV, followed by a conclusion in Section V.

## II. RELATED WORK

### A. Geoscience image processing

Image processing and recognition have wide applications in geoscience and remote sensing. The Google Earth aerial-view images taken by satellites are desirable for remote sensing scene classification [1], [2]. The street-view Google Street images can be used to retrieve the aerial-view UAV or satellite images for the cross-view geolocation in urban environments [3], [4]. Different objects are possible to be detected by some CNN based methods in the aerial-view Google Earth color images [10], [11]. Given an aerial-view Google Earth color image, image captioning can be utilized for remote sensing understanding [12], [13]. Most of the existing researches assume that all the images are clear and high-quality. However, the geoscience images might contain occlusions during the image acquisition in many real-world cases, which result in significant difficulties for geoscience

studies. For example, the street-view image might be partially blocked by a passing pedestrian or car close to the camera; the aerial-view image might be occluded by some thick clouds or flying objects. This paper focuses on the processing for the occluded geoscience images.

### B. Inpainting

The inpainting problem aims to repair the occluded image regions [16], [20]–[23], [26]–[32]. Different methods have been designed for image inpainting. Maintaining both temporal coherence and semantic structural coherence is useful for the structure-guided image sequence inpainting [20]. Pyramidal attention mechanism and dynamic normalization to ensure feature integrity and consistency is beneficial for insufficient feature representation and inaccurate regularization during inpainting [33]. Both deep features and the structural prior information from a reference image could be utilized for reference-guided inpainting [34]. Deep generative model can be used to address the challenges posed by traditional convolution operations for inpainting [35]. Roughly, the image inpainting methods can be divided into two classes based on the input style: regular and irregular holes. The image inpainting with regular holes means that the input is a rectangle-shape hole or multiple rectangle-shape holes, like the problems in [16], [26]. The image inpainting with irregular holes means that the input shape is irregular, like the problems in [27]–[29]. The context-aware information is very important to repair the occluded regions [16] no matter for the regular or irregular input. This context-aware information could be enhanced in multiple ways, such as learning by the reconstruction and adversarial losses [16], the residual aggregation [29], contextual attention CNN layers [26], and partial convolutions [27], [20], [33], [35]. There are three main differences for our work: 1) Other inpainting methods deal with common image inpainting while our method is specifically designed for geoscience images. 2) Other inpainting methods focus only on enhancing image quality, however our approach focuses more on improving the geoscience task performance. 3) Our method is an effective learnable preprocessing for the already deployed deep learning based geoscience model, which does not need to be changed after deployment, to solve the occlusion problem.

### C. Inpainting for geoscience images

Existing inpainting methods for geoscience images can be divided into two classes. The first class belongs to non-deep learning based methods, which apply the traditional image processing based methods [36], [37] to repair the contaminated region. Due to the advanced performance of deep learning, the second class leverages the CNN based methods for geoscience image inpainting [38]–[42] to remove the occlusions.

The proposed task-driven inpainting problem for geoscience images has several differences compared to regular image inpainting on nature images: 1). The final goal is to improve the geoscience related task performance with the advanced visualization quality. 2). The context of geoscience image is more complex than nature image without prior knowledge.

For example, toward the inpainting of face images, we can somewhat infer the corresponding occluded region because we have the prior knowledge of face. 3). The cost of collecting geoscience images is higher than that of collecting nature images, so the geoscience image dataset is typically much smaller than the regular nature image dataset. The existing geoscience image inpainting works are almost all focused on improving the repaired visualization quality, by ignoring the geoscience related task performance. In addition, the existing geoscience image inpainting methods ignore the above mentioned second and third special difficulties for geoscience images. Due to the complex contexts and limited training data, it might be challenging to learn a reliable deep learning model within one stage, so we propose to reconstruct the occluded region by a coarse-to-fine adversarial encoder-decoder structure and the assistance of a MaskMix based data augmentation during model training.

## III. PROPOSED METHOD

In this section, we explain the proposed network of task-driven image inpainting for geoscience images in details. The particular network structure is explained in Section III-A, while the loss functions for network training are presented in Section III-B, and the MaskMix based data augmentation is shown in Section III-C.

### A. Network Overview

The overall network structure is shown in Fig. 2. It is an end-to-end deep encoder-decoder based Generative Adversarial Network (GAN). Generally speaking, encoder-decoder based GAN methods have been widely used in many computer vision tasks with excellent reconstruction and recognition capabilities [43], [44]. As we discussed above, reconstructing the occluded image regions is not easy due to the complex context and limited training data of geoscience images. It is hard to accomplish this task in one stage, so we learn to reconstruct the occluded image regions in a coarse-to-fine manner. In particular, it contains two Encoder-Decoder sub-networks: Coarse Network and Residual Refinement Network. The “coarse-to-fine” spirit is similar to some related computer vision work [45].

**Encoder-Decoder Structure:** The Encoder-Decoder design has been widely used in different computer vision tasks [46]. Our proposed Coarse Network and Residual Refinement Network use the same Encoder-Decoder structure. Our backbone Encoder-Decoder structure is modified from [47]. The Encoder layers are from the ResNet-34 [48], and the Decoder is symmetric to the Encoder in network layers. Six skip connections are built between the symmetric layers of the Encoder and Decoder for a better deep feature fusion and perception. The Encoder implements the feature down-sampling and then the Decoder up-sample the features back to the same size of input image. The input and output of the encoder-decoder structure are  $\mathbb{R}^{H \times W \times 4}$  and  $\mathbb{R}^{H \times W \times 3}$ , respectively. The detailed organization of the used Encoder-Decoder structure is shown in Fig. 3.

**Coarse Network:** The target of Coarse Network is to predict a coarse reconstruction map  $R_{coarse}$  directly from the input occluded image and the occlusion mask  $M$ . It follows the above mentioned encoder-decoder structure to reconstruct the occluded region.

**Residual Refinement Network:** Sometimes, it might be hard for a deep neural network to directly learn the prediction well fitting the ground truth, which is because of the gradient vanishing problem in the backpropagation of the deep neural network. The ResNet work [48] shows that the residuals (by skip connection) could be learned to overcome the gradient vanishing problem in the backpropagation of the deep neural network. The objective of Refinement Network is to predict a refined reconstruction map  $R_{refined}$  from the coarse reconstruction map  $R_{coarse}$  given the occlusion mask  $M$ . Inspired by [47], [49], the refinement network is described as a residual block that refines the coarse reconstruction map, then the problem is transferred to learn the residuals  $E_{residual}$  between the reconstructed maps and the ground truth by the following equation:

$$R_{refined} = R_{coarse} + E_{residual}, \quad (1)$$

where  $R_{refined} = g(R_{coarse}|M)$ , and  $g$  indicates the nonlinear mapping function learned by the Refinement Network. In the network structure, we build a skip connect from the output of Coarse Network ( $R_{coarse}$ ) to the end of the Refinement Network to implement this residual refinement.

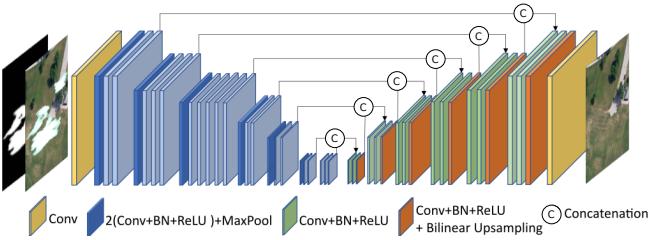


Fig. 3. Architecture of the Encoder-Decoder structure. It includes encoder, decoder and skip connections between the symmetric layers.

**Coarse-to-fine Discriminators:** To enhance the context understanding of the occluded region, we deploy a Coarse Discriminator and a Refinement Discriminator for the Coarse Network and Refinement Network, respectively. The Coarse Discriminator is designed to distinguish the coarse reconstruction map  $R_{coarse}$  as real or fake. Furthermore, the Refinement Discriminator is to judge the refined reconstruction map  $R_{refined}$  as real or fake. By introducing an adversarial learning between the encoder-decoder networks (generator) and the discriminators, our generator could produce the reconstructed maps close to the real ground-truth images so as to fool the discriminators in a coarse-to-fine way. Specifically, in order to make the discriminators pay more attention on occluded regions, we process the reconstructed maps to change the input of discriminator,  $D_{input}$ , by the following equation:

$$D_{input} = R \odot M + I \odot (1 - M), \quad (2)$$

where  $R$  is a reconstructed map,  $I$  is the input occluded image,  $M$  is binary occlusion mask, and  $\odot$  means pixel-level dot product. The processing in Eq. (2) are applied to both Coarse and Refine Discriminators with the corresponding reconstructed maps, respectively. The network structure of each discriminator is adopted from the discriminator of the PatchGAN in the Pix2Pix model [50], whose output is a  $32 \times 32$  feature map for the real or fake classification. We also feed the input occluded image  $I$  into the discriminators to simulate the conditional GAN.

**Geoscience Task Network:** In order to accomplish the task-driven inpainting network, we incorporate the sub-network related to the specific geoscience task into the overall pipeline of the proposed method, as shown in Fig. 2. We treat the Geoscience Task Network as a fixed deep neural network that are pre-trained on clean images. This Geoscience Task Network is only used during training stage and discarded during testing stage. With such a design, the reconstructed image by our proposed inpainting method could be directly fed into the existing/deployed deep neural network for geoscience tasks during testing, so the existing deep learning based geoscience model will be compatible to both clean and occluded images without any changes. In other words, the proposed inpainting method can be used as a pre-processing procedure for the occluded geoscience images.

### B. Loss Functions

In this section, we will introduce the detailed loss functions to train the proposed network.

**1) Reconstruction Loss:** Because the geoscience images have complex contexts with limited data, we design the comprehensive reconstruction loss from multiple perspectives. We reconstruct the occluded region based on two considerations: pixel-level intensity mismatching, and human perception difference. Based on these two considerations, we define the comprehensive reconstruction loss based on two particular losses,  $\mathcal{L}_1$  Loss, Perceptual Loss, to compare the reconstructed maps with the ground-truth image.

**$\mathcal{L}_1$  Loss:** We expect to minimize the pixel-level intensity mismatching between the reconstructed image  $R$  and the ground-truth image  $I_{gt}$ , where the pixel-level intensity mismatching is computed as their  $L_1$  distance, as defined as:

$$\mathcal{L}_1 = \|R - I_{gt}\|_1. \quad (3)$$

Minimizing the  $\mathcal{L}_1$  Loss makes  $R$  and  $I_{gt}$  have similar pixel-level intensity.

**Perceptual Loss:** We also want to minimize the human perception difference between  $R$  and  $I_{gt}$ , which is computed by LPIPS (Learned Perceptual Image Patch Similarity) distance [51] with a ImageNet pre-trained VGG16 model, as defined as:

$$\mathcal{L}_p = \text{LPIPS}(R, I_{gt}), \quad (4)$$

where **LPIPS** is a standard operation defined in [51] to compute the  $L_2$  distance of the activation feature maps in

different CNN layers between two input images. The LPIPS metric is recently shown as more similar to the human perception seeing an image. Minimizing the  $\mathcal{L}_p$  Loss will force  $R$  and  $I_{gt}$  to have consistent human perception response.

For the Coarse Network, we only use the  $\mathcal{L}_1$  loss for its reconstruction, while we use the summation of  $\mathcal{L}_1$  Loss and Perceptual Loss for the Refinement Network's reconstruction.

2) *Adversarial Loss*: Let us treat the proposed network in Fig. 2 except the Coarse Discriminator  $D_c$  and Refinement Discriminator  $D_r$  as a coarse-to-fine generator  $G$ . We learn the generator  $G$  and the discriminators  $D_c$  and  $D_r$  by the following adversarial loss:

$$\begin{aligned} \min_G \max_{D_c, D_r} \mathcal{L}_{GAN} = & \mathbb{E}_{I, I_{gt}} [\log D_c(I, I_{gt})] + \\ & \mathbb{E}_{I, G(I)} [\log(1 - D_c(I, G(I)_c))] + \\ & \mathbb{E}_{I, I_{gt}} [\log D_r(I, I_{gt})] + \\ & \mathbb{E}_{I, G(I)} [\log(1 - D_r(I, G(I)_r))], \end{aligned} \quad (5)$$

where  $I$  is the input occluded image,  $I_{gt}$  is the ground-truth reconstructed image for  $I$ ,  $G(I)_c$  is the predicted coarse reconstruction map,  $R_{coarse}$ , and  $G(I)_r$  is the predicted refined reconstruction map,  $R_{refined}$ .  $G$  tries to minimize the loss  $\mathcal{L}_{GAN}$  against two adversarial  $D_c$  and  $D_r$  that would like to maximize it. The coarse and refinement discriminators try to classify the reconstructed region as real or fake in a coarse-to-fine manner. Since the context of geoscience images is complex with limited training data, we use the coarse-to-fine generator  $G$  to reconstruct the occluded regions, and simultaneously we use the coarse-to-fine discriminators to adversarially improve the context reconstruction ability of the generator  $G$ .

3) *Geoscience Task Loss*: The geoscience task loss  $\mathcal{L}_T$  is related to the specific geoscience task. In this paper, we applied the proposed network to three geoscience tasks as example: remote sensing scene recognition, cross-view geolocation and semantic segmentation. For example, in the task of remote sensing scene recognition, the internal purpose is image classification, so we could choose the image classification network like fixed VGG16 [52] pretrained on clean images as the geoscience task network, where the corresponding cross entropy loss for image classification is used as the geoscience task loss. When applying our proposed method for cross-view geolocation, the geoscience task network could be the fixed LPN [4] pretrained on clean images, and the corresponding geoscience task loss is the summation of cross entropy losses over all image parts as defined in [4]. For the semantic segmentation task, we use HRNet [53] pretrained on clean images as the geoscience task network and set the pixel-level cross-entropy loss as the geoscience task loss. Minimizing the geoscience task loss  $\mathcal{L}_T$  will make the reconstructed image well fit the fixed geoscience task network pretrained on clean images.

The overall loss function for training our proposed network is shown below as

$$\mathcal{L}_{overall} = \mathcal{L}_1 + \alpha_1 \mathcal{L}_P + \alpha_2 \mathcal{L}_{GAN} + \lambda \mathcal{L}_T, \quad (6)$$

where  $\alpha_1$ ,  $\alpha_2$  and  $\lambda$  are the weights to balance the image quality of reconstruction and the geoscience task performance,  $\mathcal{L}_T$  depends on the specific geoscience task.

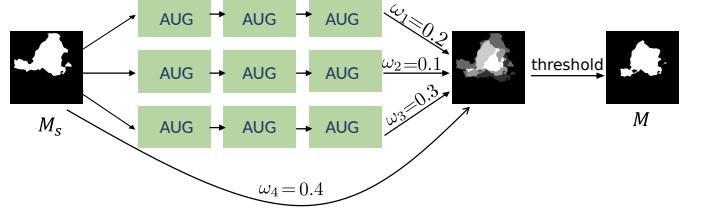


Fig. 4. Illustration of MaskMix based data augmentation. AUG indicates the random augmentation operation of “translate”, “shear” and “rotate”.

### C. MaskMix based Data Augmentation

Since the geoscience data is always limited, not as large as the natural image dataset like ImageNet, we propose a novel data augmentation method for training image inpainting models to make better use of limited geoscience training data. Most of existing inpainting methods take the fixed masks, so some occlusion patterns are never processed when training the inpainting model, which limits the generalization capability of models to handle different occlusion scenarios. To address this issue, we propose to augment seed masks by conducting a series of transformations on the masks, which might happen in the real world, and mix them to simulate more complex scenarios. Specifically, given a seed mask  $M_s$  (e.g., the cloud in a satellite image), we transform it via different augmentation operations (e.g., translation, shearing, and rotation) and obtain multiple augmented masks. Intuitively, these processes might simulate the cloud moving and reshaping in the real world. Finally, we mix all augmented masks to a single mask to obtain a complex occlusion pattern. Based on this idea, inspired by AugMix [57], we propose to augment the diversity of seed masks and increase robustness to unforeseen occlusion scenarios by a MaskMix based data augmentation. As shown in Fig. 4, we set three random operations of “translate”, “shear” and “rotate” in parallel three branches to generate several random masks, then the MaskMix based data augmentation is computed by

$$M = \Phi\{w_4 * M_s + \sum_{i=1}^3 w_i * A_i^3(A_i^2(A_i^1(M_s)))\}, \quad (7)$$

where  $M_s$  is the binary seed mask,  $A_i^j$  is the randomly sampled augmentation operations of “translate”, “shear” and “rotate” in the  $i$ -th row and  $j$ -th column as shown in Fig. 4,  $w_i$  is the random sample mixing weight in the  $i$ -th row,  $\Phi\{\cdot\}$  indicates the thresholding operation, and  $M$  is the final augmented binary mask by the MaskMix.  $M$  is fed to train our proposed inpainting framework instead of the seed mask  $M_s$ . Different with the AugMix [57] which augments the original image, the proposed MaskMix aims to augment the binary mask for the image inpainting problem. Please note that the proposed MaskMix based data augmentation is only applied in the model training, not in the model testing. For some special data format (like human 3D skeleton joints [58]), the advanced data



Fig. 5. Datasets of image inpainting for two geoscience tasks: Remote Sensing (RS) scene recognition and cross-view geolocation: (a) RSSCN7 [1] dataset for RS scene recognition (Top: satellite view, Bottom: occluded satellite view), (b) CVUSA [54] dataset for cross-view geolocation (From left to right: ground view, occluded ground view, corresponding satellite view of same location).

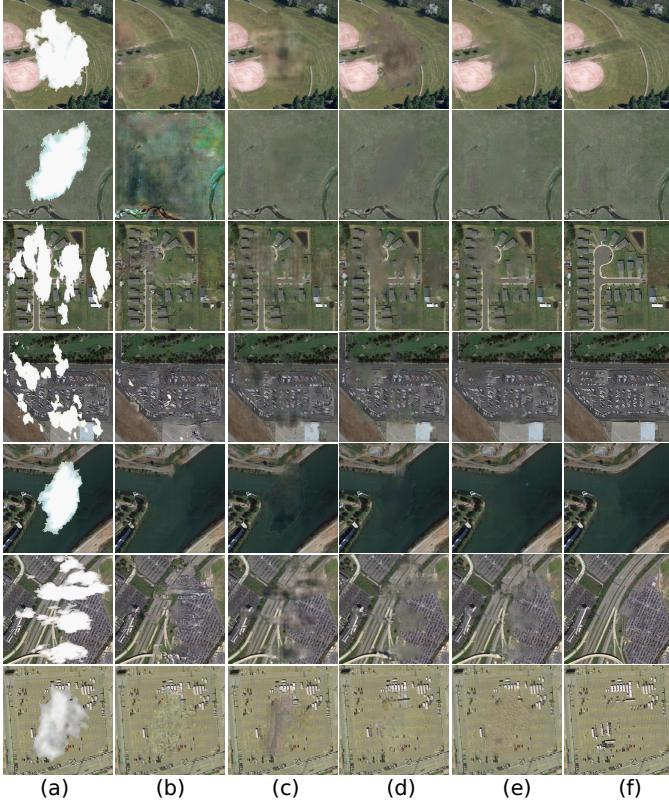


Fig. 6. Qualitative comparisons on the RSSCN7 dataset: (a) Input images, (b-f) image inpainting results by CSA [18], RFR [55], MISF [56], the Proposed method and ground truth.

augmentation method could extract rotate-shear-scale invariant features of 3D skeleton by automatically learning some affine transformation matrix [58]. However, the data format of image inpainting problem does not have some clearly defined affine transformation related features to be augmented. Differently, MaskMix aims to simulate different occlusion patterns in geoscience image inpainting.

#### IV. EXPERIMENTS

In this section, we will evaluate the proposed method on three widely-used geoscience tasks: remote sensing (RS) scene recognition, cross-view geolocation, and semantic segmentation. The RS scene recognition task is to recognize an aerial-view satellite image into predefined classes [1], similar to the image classification problem. The cross-view

geolocation task is to localize the spot by matching an given street-view image to the corresponding aerial-view satellite or UAV image in a gallery [4], similar to the image retrieval problem. The remote sensing semantic segmentation task is to identify the land-cover or land-use category of each part of the remote sensing High Spatial Resolution (HSR) image [59].

##### A. Dataset for RS Scene Recognition

The dataset for the RS scene recognition task is the public aerial-view Google Earth satellite images, i.e., RSSCN7 [1] dataset. The RSSCN7 dataset contains satellite images acquired from Google Earth, which is originally collected for remote sensing scene classification. We conduct image synthesis on RSSCN7 to make it capable of the image inpainting task. It has seven classes: grassland, farmland, industrial and commercial regions, river and lake, forest field, residential region, and parking lot. Each class has 400 images, so there are total 2,800 images in the RSSCN7 dataset. 50% is used for the network training, and another 50% is for network testing in the RSSCN7 dataset. We first extract some thick/nontransparent clouds as 28 anchor masks from some real cloudy satellite images [60], [61]. For each image of RSSCN7 dataset, we randomly pick one mask, and randomly rotate, translate, resize the mask and overlap it to the original image, and we make a constraint that the area ratio of the added occlusion over the whole image is between 15% and 60%. The inpainting examples of occluded RSSCN7 dataset are shown in Fig. 5.

##### B. Dataset for Cross-view Geolocation

The public dataset used for the cross-view geolocation task is the CVUSA [54] dataset. It includes the geoscience data collected from the ground view (street view) and satellite view. In CVUSA dataset, all the ground-view images are panoramic images downloaded from Google Street View, while the corresponding satellite-view images are collected from Microsoft Bing Maps. There are 35,532 ground-and-satellite image pairs for training and 8,884 image pairs for testing. The seed masks used to simulate occlusions are downloaded from the public image inpainting masks [27] and added to the ground-view images only. We use the seed masks with 10-20% area ratio in the experiment. In this task, we assume that the satellite-view images are clean without occlusions. The examples of occluded CVUSA dataset for image inpainting are shown in Fig. 5.

TABLE I

QUANTITATIVE EXPERIMENTAL RESULTS ON THE RSSCN7 DATASET FOR RS SCENE RECOGNITION USING FIXED VGG16 AS THE GEOSCIENCE TASK NETWORK

Classification	PSNR	SSIM	MSE	BRISQUE	ACCURACY
Clean Testing [52]	-	-	-	-	93.57
Occluded Testing [52]	12.04	0.75	4837.56	23.33	70.86
CSA [18]	25.45	0.76	312.70	23.64	84.50
RFR [55]	26.01	0.78	275.58	19.72	86.21
MISF [56]	<b>27.44</b>	0.84	<b>228.09</b>	17.36	87.93
GMCNN [19]	25.35	0.79	303.38	16.21	78.14
SPL [62]	26.69	<b>0.84</b>	237.55	<b>13.58</b>	77.71
MAT [63]	24.91	0.77	335.42	22.05	87.79
PDGAN [64]	24.37	0.71	403.83	19.26	82.36
Proposed <sub>b</sub>	26.07	0.79	283.27	22.96	87.00
Proposed-	26.24	0.79	269.85	24.82	88.79
Proposed	25.43	0.77	291.53	24.38	<b>89.36</b>

### C. Dataset for Semantic Segmentation

The dataset used for semantic segmentation task is the public LoveDA [59] dataset. This dataset consists of rural and urban images that are obtained from Google Earth platform, along with their pixel-level labels. The LoveDA dataset contains 5,987 High Spatial Resolution (HSR) images in total, with image resolution of  $1024 \times 1024$  pixels, whose spatial resolution is 0.3 m. In this experiment, we follow the default setting of LoveDA [59] to have 4,191 images for training and the other 1,796 images for testing. As for the occlusion simulation, we employ the same strategy as that in cross-view geolocation task to generate occluded images and masks.

### D. Experimental Setups

**Implementation details:** In the RS scene recognition experiment, the VGG16 based image classification network [52] is used as the geoscience task network. In the cross-view geolocation experiment, the Local Pattern Network (LPN) [4] is used as the geoscience task network. In the semantic segmentation experiment, HRNet [53] is used as the geoscience task network. The task networks are trained independently on the clean images without occlusions, and then they are fixed for task performance evaluation. We denote “Clean Testing” as testing clean images without occlusions on the fixed geoscience task network, and denote “Occluded Testing” as testing the occluded images on the fixed geoscience task network. During the initialization of the proposed network, the first three layers of the encoders are initialized from the ImageNet pre-trained ResNet-34 [48] model, and the other layers are randomly initialized. We set the loss balance weight  $\alpha_1$  as 1,  $\alpha_2$  as 1 and  $\lambda$  as 5 for RS scene recognition and semantic segmentation, 1.2 for cross-view geolocation. During the network training, each image is resized to  $256 \times 256$  for RS scene recognition and cross-view geolocation task,  $512 \times 512$  for semantic segmentation task. We use the PyTorch framework to implement the proposed network and all the experiments are run with a NVIDIA RTX 3090 GPU card. For more details, we will publicize the code after paper acceptance.

**Baselines:** We compare the proposed method with several state-of-art image inpainting or image transfer methods. These

methods are RFR [55], CSA [18], GMCNN [19], SPL [62], MISF [56], PDGAN [64] and MAT [63]. These comparison methods are carefully trained on the related geoscience datasets until convergence. We use *Proposed<sub>b</sub>* to represent our proposed baseline method (without Geoscience Task Network and MaskMix), *Proposed-* as our proposed method (without MaskMix), *Proposed* as our full proposed method.

**Metrics:** The evaluation metrics are two folds. One side is for the image quality evaluation, following most image inpainting researches, we use structural similarity index (SSIM) [65], peak signal-to-noise ratio (PSNR), mean squared error (MSE), and the no-reference image quality score: blind referenceless image spatial quality evaluator (BRISQUE) [66]. The other side is for the geoscience task-related performance evaluation. In the RS scene recognition task, the overall classification accuracy (%) on the whole testing set is used, following [1]. In the cross-view geolocation task, we use Recall@K (R@K) and the average precision (AP) to evaluate the performance following [4], where R@K represents the proportion of correctly matched images in the top-K of the ranking list. AP calculates the area under the Precision-Recall curve, which shows the precision and recall rate of the retrieval performance. For the semantic segmentation task, mean intersection over union (mIoU) metric is deployed to evaluate the performance. Note that for most metrics we mentioned above, higher value indicates better performance; while for MSE and BRISQUE metrics, lower value means better image quality.

### E. RS Scene Recognition Results

Table I shows the quantitative performance on the RSSCN7 dataset. When there are no occlusions, the geoscience task network gets 93.57% overall accuracy, seeing “Clean Testing”. If we feed the occluded images into the fixed geoscience task network pre-trained on clean images, the overall accuracy for “Occluded Testing” is only 70.86%, since it is more difficult for the fixed geoscience task sub-network to make accurate recognition when there are occlusions in image. With each of the image inpainting methods, the recognition accuracy is improved, this improvement indicates that image inpainting could help the geoscience related task. Among the comparison methods, the MISF [56] method got the best image quality in PSNR 27.44 and MSE 228.09, with advanced classification performance of 87.93% Accuracy. SPL [62] reached the best image quality in SSIM 0.84 and BRISQUE 13.58, but had poor Accuracy 77.71%. However, the Proposed method obtained the best classification performance of 89.36% Accuracy and advanced image quality. This phenomenon demonstrates that the proposed method could not only achieve the best task-related performance, but also obtain comparable reconstructed image quality. The qualitative results of image inpainting for RS scene recognition are shown in Fig. 6.

Comparing the Proposed<sub>b</sub>, Proposed-, and Proposed versions of our method, we can see that 1) the proposed coarse-to-fine baseline method is with good-quality reconstruction and reasonable task-related performance; 2) adding the geoscience task network into the proposed

TABLE II

QUANTITATIVE EXPERIMENTAL RESULTS ON THE CVUSA DATASET FOR CROSS-VIEW GEOLOCATION USING FIXED LPN [4] (PRETRAINED ON CLEAN IMAGES) AS THE GEOSCIENCE TASK NETWORK.

Methods	PSNR	SSIM	MSE	BRISQUE	Recall@1	Recall@5	Recall@10	Recall@top1%	AP
Clean Testing [4]	-	-	-	-	-	93.14	95.28	98.83	84.77
Occluded Testing [4]	11.87	0.70	8700.67	17.51	24.55	38.89	45.29	68.18	28.12
CSA [18]	25.42	0.81	218.19	16.43	52.60	70.81	76.81	91.46	56.85
RFR [55]	23.56	0.77	313.11	20.71	48.22	66.32	72.58	88.28	52.47
GMCNN [19]	25.08	0.81	231.19	16.06	46.15	63.76	69.91	86.37	50.27
MISF [56]	25.49	0.83	219.60	16.48	69.61	85.91	89.95	97.07	73.31
SPL [62]	<b>26.56</b>	<b>0.85</b>	<b>172.35</b>	16.09	70.26	86.20	90.26	96.88	73.86
MAT [63]	24.74	0.79	255.73	21.82	55.34	73.65	79.62	92.75	59.57
PDGAN [64]	22.92	0.73	378.70	16.53	45.81	64.64	71.65	88.54	50.30
Proposed <sub>b</sub>	26.28	0.82	181.75	15.98	68.38	85.11	89.43	96.81	72.16
Proposed-	26.11	0.82	188.81	<b>15.42</b>	74.54	89.00	92.17	97.77	77.81
Proposed	26.01	0.82	197.58	16.18	<b>75.19</b>	<b>89.42</b>	<b>92.56</b>	<b>98.00</b>	<b>78.44</b>

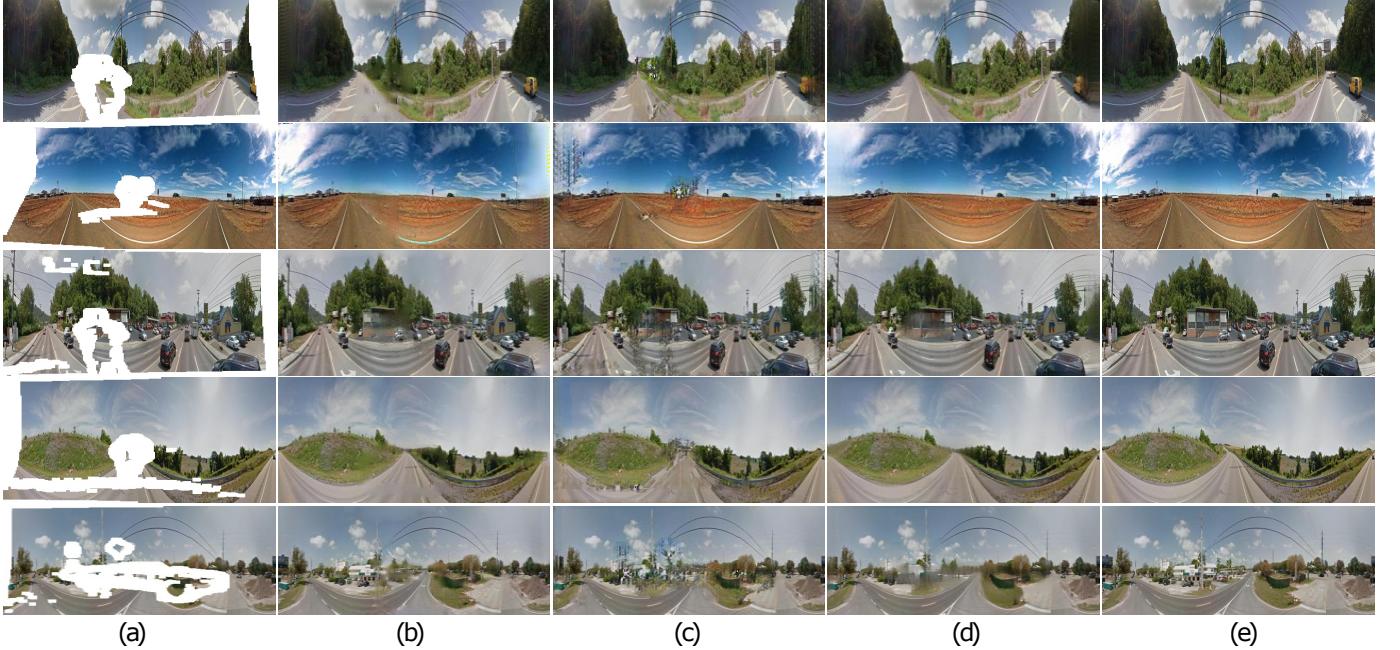


Fig. 7. Qualitative comparisons on the CVUSA dataset: (a) Input ground/street-view images, (b-d) image inpainting results by CSA [18], RFR [55], and the Proposed method, (e) ground truth.

framework is helpful in the RS scene recognition task, with only slightly decrease in reconstructed image quality; 3) further including MaskMix in the proposed framework could continue to improve the task-related performance and reconstructed image quality.

#### F. Cross-view Geolocation Results

Table II shows the quantitative results on the CVUSA dataset for the cross-view geolocation task. With the fixed LPN [4] pretrained on clean images as the geoscience task network, clean images could obtain 84.77% AP while occlusions will reduce it to 28.12%. It demonstrates that occlusion leads to the significant challenge to the cross-view geolocation problem. By each image inpainting method, the task-related performance could be increased. Among them, the Proposed method gets the best AP as 78.44%, much larger

than other image inpainting methods. This is because that the Proposed method is designed to improve the geoscience task-related performance without changing the fixed geoscience task network. Reconsidering 1) the large increase from 28.12% AP to our 78.44% improvement and 2) no need to change the fixed geoscience task network pretrained on clean images, our proposed method is quite promising to process and understand the occluded geoscience images. Specifically, the Proposed<sub>b</sub> already obtains better performance than other comparison methods on both image quality evaluation metrics and geoscience task-related evaluation metrics. When combined with geoscience task network (Proposed-) and MaskMix (Proposed), the task-related performance, R@K and AP, could be further improved, which is consistent with the experimental results for RS scene recognition.

It is worth mentioning that the image quality (PSNR, SSIM) of final Proposed method is relatively high but not

TABLE III

QUANTITATIVE EXPERIMENTAL RESULTS ON THE LOVEDA DATASET FOR RS SEMANTIC SEGMENTATION USING FIXED HRNET [53] (PRETRAINED ON CLEAN IMAGES) AS THE GEOSCIENCE TASK NETWORK.

Method	PSNR	SSIM	MSE	BRISQUE	MIoU
Clean Testing [53]	-	-	-	-	0.4403
Occluded Testing [53]	12.13	0.81	4011.74	30.06	0.3554
CSA [18]	29.75	0.93	89.24	25.94	0.4231
RFR [55]	31.10	0.94	75.53	25.46	0.4132
MISF [56]	31.66	0.94	<b>61.05</b>	26.51	0.4271
SPL [62]	<b>31.81</b>	<b>0.95</b>	66.45	25.31	0.4176
MAT [63]	29.30	0.91	95.61	<b>15.70</b>	0.2886
PDGAN [64]	25.30	0.76	215.61	17.41	0.2191
Proposed <sub>b</sub>	29.31	0.93	192.65	26.16	0.4283
Proposed-	29.35	0.93	185.65	26.16	0.4320
Proposed	30.51	0.94	79.88	26.05	<b>0.4326</b>

the best, while SPL [62] has the best scores on three image quality metrics but much lower AP than proposed method. *This also verifies that better image quality not always lead to higher geoscience task performance.* Our goal is applying inpainting to reach much better geoscience task performance along with relatively good image quality, without changing the fixed geoscience task network pretrained on clean images. Some previous computer vision research also shows that only enhancing the image reconstruction quality does not necessarily improve the next computer vision task performance. As shown in [67], deraining even decreases object detection accuracy on the rainy images. As studied in [68], removing haze cannot largely improve its classification performance. The qualitative results of image inpainting for cross-view geolocation are shown in Fig. 7.

#### G. RS Semantic Segmentation Results

Table III shows the quantitative performance on the LOVEDA dataset. When there are no occlusions, the segmentation task network gets 0.4403 mIoU. If the occluded images are fed into the semantic segmentation task network pre-trained on clean images of the LOVEDA dataset, the mIoU for “Occluded Testing” is only 0.3554. The mIoU performance could be improved by different inpainting methods. These improvements demonstrate that inpainting methods help to reduce the impact of occluded regions on geoscience task. Among the comparison methods, the SPL method obtains the advanced image quality of PSNR 31.81, SSIM 0.95, with mIoU 0.4176. However, the proposed method achieves the best mIoU 0.4326, with comparable PSNR and SSIM score of 30.51 and 0.94. This phenomenon indicates that although the proposed method did not reach the highest image reconstruction quality, it could obtain the best task-related performance. The qualitative results of image inpainting for semantic segmentation are shown in Fig. 8. The proposed method could generate comparable image reconstruction quality as other methods while obtain the best task-related results at the same time.

#### H. Discussion: Effectiveness of Coarse & Refined Inpaintings

In this section, we study the effectiveness of coarse and refined inpaintings respectively. Taking the Proposed

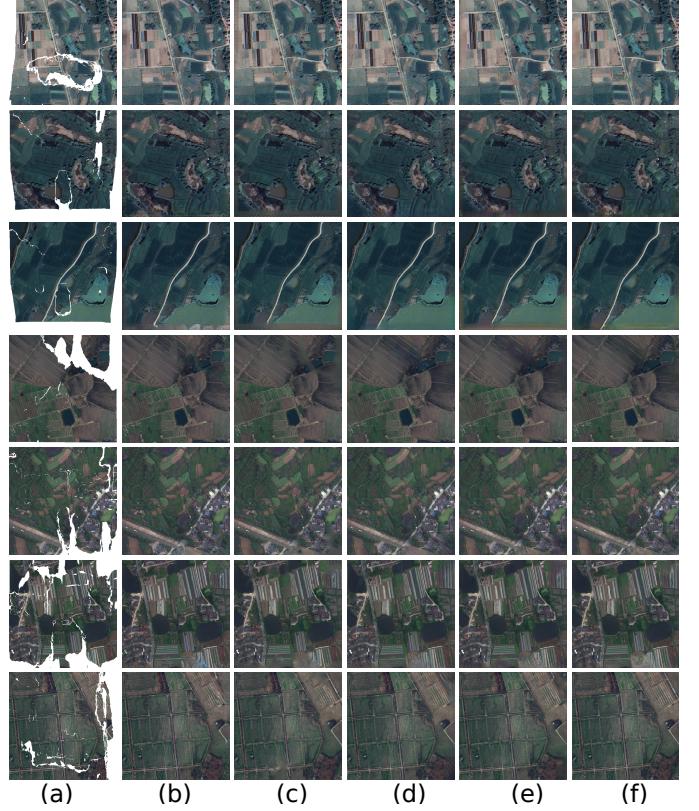


Fig. 8. Qualitative comparisons on the LOVEDA dataset: (a) Input images, (b-e) image inpainting results by CSA [18], RFR [55], MISF [56] and the Proposed method, (f) ground truth.



Fig. 9. Illustration of the coarse and refined reconstruction maps by the proposed image inpainting method using Proposed<sub>b</sub> as example. From left to right: input satellite image with occlusions, coarse reconstruction map, refined reconstruction map, and ground truth.

method on RSSCN7 dataset as an example, the coarse-only reconstruction gets classification accuracy of 86.64%, with SSIM 0.75; while the coarse-to-fine reconstruction obtains better classification accuracy of 89.36%, with SSIM 0.79. Therefore, the coarse-to-fine learning performs better than the coarse-only learning. As shown in Fig. 9, compared to the coarse reconstruction map, the refined reconstruction map looks more smooth with less defects in our experiment.

#### I. Discussion: Weight of Task Loss

In this section, using the RS Scene Recognition task as example, we study the impact of task loss weight  $\lambda$  in Eq. (6).

TABLE IV

EXPERIMENTAL RESULT OF THE PROPOSED METHOD FOR RS SCENE RECOGNITION TASK ON RSSCN7 DATASET WITH DIFFERENT TASK LOSS WEIGHTS  $\lambda$ .

Setting	PSNR	SSIM	MSE	BRISQUE	ACCURACY
$\lambda = 0$	26.07	0.79	283.27	22.96	87.00
$\lambda = 3$	26.24	0.79	269.85	24.82	88.79
$\lambda = 5$	25.43	0.77	291.53	24.38	89.36
$\lambda = 10$	23.64	0.75	380.27	25.94	84.21

Table IV shows the experimental results when setting different  $\lambda$ . When  $\lambda$  is very small (e.g.,  $\lambda = 0$ ), the network tends to reconstruct a better image quality, by somewhat ignoring the task performance. With increased  $\lambda$  (e.g.,  $\lambda = 5$ ), the network aims to obtain a higher task performance. When  $\lambda$  is too large (e.g.,  $\lambda = 10$ ), the tradeoff between quality and task might be imbalanced, leading to a lower performance. In the RS Scene Recognition experiment, the best task performance is achieved by setting  $\lambda = 5$ , where the task loss is about 28% of the overall loss in numbers. We hope this discovery could be helpful to other related research works.

#### J. Discussion: Weights of $\mathcal{L}_P$ and $\mathcal{L}_{GAN}$

In this section, we investigate the influence of loss weights  $\alpha_1$  for  $\mathcal{L}_P$  and  $\alpha_2$  for  $\mathcal{L}_{GAN}$  in Eq. (6) on the proposed method. Using RS Scene Recognition task as example, Table V shows performance of the proposed method under different settings of  $\alpha_1$  and  $\alpha_2$ . We found that the increasing of  $\alpha_1$  and  $\alpha_2$  only leads to slight performance difference, so the proposed method is generally robust against the changing of these two parameters.

#### K. Discussion: Different Discriminator Configurations

**PatchGAN Discriminator Size Change:** The coarse discriminator and refined discriminator in our network adopt the same PatchGAN [50] architecture with three downsampling layers. As the input image size is  $256 \times 256 \times 3$ , the final output of PatchGAN is a  $32 \times 32$  feature map. By adjusting downsampling layers in either coarse discriminator or refined discriminator, we investigate the influence of different PatchGAN size configurations on the proposed network. Note that the weights of coarse discriminator and refined discriminator are not shared, so they are independent even with the same PatchGAN size configuration. Using RS Scene Recognition task as example, Table VI shows that performance of the proposed method is relatively robust with different PatchGAN discriminator size configurations.

**Discriminator Structure Change:** We conduct an experiment to test two different discriminator structures. The proposed method uses two PatchGAN discriminators [50], consisting of 3 convolution layers, for coarse discriminator  $D_c$  and refined discriminator  $D_r$ , denoted as  $\{\text{PatchGAN}, \text{PatchGAN}\}$  for  $\{D_c, D_r\}$ . Then, we replace the coarse discriminator as the discriminator of DCGAN [69], consisting of 5 convolution layers and one fully-connected layer. The output of the PatchGAN discriminator is a  $32 \times 32$  feature map

TABLE V

EXPERIMENTAL RESULT OF THE PROPOSED METHOD FOR RS SCENE RECOGNITION TASK ON DIFFERENT WEIGHTS OF  $\alpha_1$  AND  $\alpha_2$ .

$\mathcal{L}_P$	$\mathcal{L}_{GAN}$	PSNR	SSIM	MSE	BRISQUE	ACCURACY
$\alpha_1 = 1$	$\alpha_2 = 1$	25.43	0.77	291.53	24.38	89.36
	$\alpha_2 = 5$	25.90	0.79	272.75	22.00	89.14
	$\alpha_2 = 10$	26.09	0.78	260.53	20.83	89.71
$\alpha_1 = 5$	$\alpha_2 = 1$	25.87	0.78	274.68	20.86	89.14
	$\alpha_2 = 5$	25.64	0.77	279.81	24.65	89.57
	$\alpha_2 = 10$	25.93	0.78	268.47	24.68	88.93
$\alpha_1 = 10$	$\alpha_2 = 1$	25.31	0.78	305.04	23.81	88.36
	$\alpha_2 = 5$	25.51	0.77	277.05	24.35	88.64
	$\alpha_2 = 10$	26.01	0.78	263.05	22.16	87.42

to describe false or true for the corresponding image patch regions, while the output of the DCGAN discriminator is a single scalar indicating whether its input is false or true. We keep the refined discriminator unchanged as PatchGAN, then this new setting is denoted as  $\{\text{DCGAN}, \text{PatchGAN}\}$  for  $\{D_c, D_r\}$ . Using RS Scene Recognition task as example, Table VII shows that performance of the proposed method is relatively robust with these different discriminator structures.

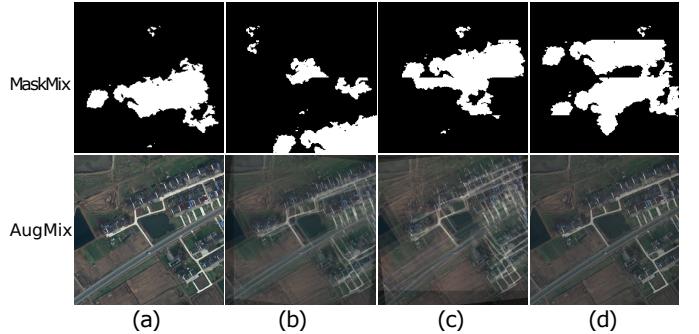


Fig. 10. Visualization of the difference between AugMix [57] and the proposed MaskMix. (a) original mask and image, (b-d) augmented masks and images by MaskMix and AugMix respectively.



Fig. 11. Illustration of the failure case by the proposed image inpainting method. From left to right: input satellite image with occlusion, result by MISF [56], result by proposed method, and ground truth.

#### L. Discussion: AugMix vs MaskMix

The Fig. 10 shows the difference between AugMix [57] and the proposed MaskMix. With different motivation for inpainting, AugMix augments/changes the original image but does not modify the mask for inpainting, while the proposed

TABLE VI

EXPERIMENTAL RESULT OF THE PROPOSED METHOD FOR RS SCENE RECOGNITION TASK ON RSSCN7 DATASET WITH DIFFERENT PATCHGAN DISCRIMINATOR SIZE CONFIGURATIONS.  $d_c$  AND  $d_r$  ARE NUMBERS OF DOWNSAMPLING LAYERS FOR COARSE AND REFINED DISCRIMINATORS RESPECTIVELY. FINAL OUTPUT FEATURE MAP SIZES IN PATCHGAN ARE SHOWN IN PARENTHESES.

Setting	PSNR	SSIM	MSE	BRISQUE	ACCURACY
$d_c = 3$ ( $32 \times 32$ ), $d_r = 3$ ( $32 \times 32$ )	25.43	0.77	291.53	24.38	89.36
$d_c = 4$ ( $16 \times 16$ ), $d_r = 3$ ( $32 \times 32$ )	25.99	0.78	267.38	23.62	89.21
$d_c = 3$ ( $32 \times 32$ ), $d_r = 4$ ( $16 \times 16$ )	25.28	0.78	298.51	23.52	88.57

TABLE VII

EXPERIMENTAL RESULT OF THE PROPOSED METHOD FOR RS SCENE RECOGNITION TASK ON RSSCN7 DATASET WITH DIFFERENT DISCRIMINATOR STRUCTURES.  $D_c$  AND  $D_r$  INDICATE COARSE DISCRIMINATOR AND REFINED DISCRIMINATOR RESPECTIVELY.

Discriminators: $\{D_c, D_r\}$	PSNR	SSIM	MSE	BRISQUE	ACCURACY
{PatchGAN, PatchGAN}	25.43	0.77	291.53	24.38	89.36
{DCGAN, PatchGAN}	26.09	0.78	267.94	23.37	89.00

MaskMix augments/changes the mask but does not modify the original image. As shown in Fig. 10, AugMix might change the object/content location in the image due to the shearing, translation, rotation, etc, so AugMix cannot be directly applied for the tasks of Cross-view Geolocation and RS Semantic Segmentation, since these two tasks rely on precise object/content location information. Therefore, we compare the experimental results on RS Scene Recognition task by replacing MaskMix with AugMix for the proposed method. As shown in Table VIII, our MaskMix could not only obtain better classification Accuracy than AugMix, but also generate images with better image quality (higher PSNR and SSIM, lower MSE and BRISQUE).

#### M. Running Time & Failure Case

During the testing stage, our running time per  $512 \times 512$  RGB color image inpainting is 0.1s, and our running time per  $256 \times 256$  RGB color image inpainting is 0.033s. This demonstrates the proposed network is efficient for geoscience image inpainting task and could be used as an image pre-processing step.

As shown in Fig. 11, if an object (like an island) of geoscience image is fully covered by occlusion or the context of the to-be-reconstructed geoscience region is quite complex, the proposed method could not well reconstruct them because the proposed method does not have enough prior knowledge of the geoscience context. This disadvantage of the proposed method could be overcome if feeding with more context images, e.g., other geoscience images of the same location in different time. This is different with the common image inpainting problem with rich prior knowledge, e.g., the face image inpainting.

## V. CONCLUSIONS

In conclusion, this paper proposed a task-driven approach for image inpainting of geoscience images. Our goal is to improve the performance of geoscience tasks while maintaining a relatively high image quality, without changing

TABLE VIII

EXPERIMENTAL RESULT OF THE PROPOSED METHOD FOR RS SCENE RECOGNITION TASK ON RSSCN7 DATASET WITH DIFFERENT AUGMENTATION METHODS.

Setting	PSNR	SSIM	MSE	BRISQUE	ACCURACY
MaskMix	25.43	0.77	291.53	24.38	89.36
AugMix	24.27	0.73	357.08	25.87	89.21

the already deployed geoscience task network pre-trained on clean images. To achieve this goal, we proposed a coarse-to-fine task-driven learning based deep CNN model with MaskMix-based data augmentation. The effectiveness and accuracy of the proposed method were demonstrated through experiments on the RSSCN7 dataset for remote sensing scene recognition, the CVUSA dataset for cross-view geolocation, and the LoveDA dataset for remote sensing semantic segmentation. The results show that our proposed method outperforms the existing methods in the geoscience task performance, making the deployed geoscience task model unchanged but more robust in occlusions. Furthermore, our approach might be applicable to other task-driven image inpainting problems and we hope that this research could inspire more researchers to contribute to this field.

## REFERENCES

- [1] Q. Zou, L. Ni, T. Zhang, and Q. Wang, “Deep learning based feature selection for remote sensing scene classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, 2015.
- [2] S. Song, H. Yu, Z. Miao, Q. Zhang, Y. Lin, and S. Wang, “Domain adaptation for convolutional neural networks-based remote sensing scene classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1324–1328, 2019.
- [3] Y. Tian, C. Chen, and M. Shah, “Cross-view image matching for geolocation in urban environments,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3608–3616.
- [4] T. Wang, Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zhenga, and Y. Yang, “Each part matters: Local patterns facilitate cross-view geo-localization,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.
- [5] L. Bruzzone and D. F. Prieto, “An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images,” *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 452–466, 2002.
- [6] B. Du, L. Ru, C. Wu, and L. Zhang, “Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9976–9992, 2019.
- [7] H. Liu, Y. Jia, J. Hou, and Q. Zhang, “Global-local balanced low-rank approximation of hyperspectral images for classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.
- [8] J. Xie, N. He, L. Fang, and P. Ghamisi, “Multiscale densely-connected fusion networks for hyperspectral images classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 246–259, 2020.
- [9] L. Sun, C. Ma, Y. Chen, Y. Zheng, H. J. Shim, Z. Wu, and B. Jeon, “Low rank component induced spatial-spectral kernel method for hyperspectral image classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3829–3842, 2019.
- [10] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, “Dota: A large-scale dataset for object detection in aerial images,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [11] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, “Learning roi transformer for oriented object detection in aerial images,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2849–2858.
- [12] X. Li, X. Zhang, W. Huang, and Q. Wang, “Truncation cross entropy loss for remote sensing image captioning,” *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

- [13] W. Huang, Q. Wang, and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, 2020.
- [14] H. Bi, L. Xu, X. Cao, Y. Xue, and Z. Xu, "Polarimetric sar image semantic segmentation with 3d discrete wavelet transform and markov random field," *IEEE Transactions on Image Processing*, vol. 29, pp. 6601–6614, 2020.
- [15] R. Liu, L. Mi, and Z. Chen, "Afnet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.
- [16] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.
- [17] F. Xie, M. Shi, Z. Shi, J. Yin, and D. Zhao, "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 8, pp. 3631–3640, 2017.
- [18] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *IEEE International Conference on Computer Vision*, 2019.
- [19] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 331–340.
- [20] C. Wang, X. Chen, S. Min, J. Wang, and Z.-J. Zha, "Structure-guided deep video inpainting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 2953–2965, 2020.
- [21] H. Wu and J. Zhou, "Iid-net: Image inpainting detection network via neural architecture search and attention," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1172–1185, 2021.
- [22] S. Xu, D. Liu, and Z. Xiong, "E2i: Generative inpainting from edge to image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1308–1322, 2020.
- [23] D. Jin and X. Bai, "Patch-sparsity-based image inpainting through a facet deduced directional derivative," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1310–1324, 2018.
- [24] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [25] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1279–1289, 2016.
- [26] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.
- [27] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *European Conference on Computer Vision*, 2018, pp. 85–100.
- [28] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [29] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [30] Y. Wang, Y.-C. Chen, X. Tao, and J. Jia, "Vnet: A robust approach to blind image inpainting," *arXiv:2003.06816*, 2020.
- [31] Q. Guo, X. Li, F. Juefei-Xu, H. Yu, Y. Liu, and S. Wang, "Jpgnet: Joint predictive filtering and generative network for image inpainting," in *ACM Multimedia Conference*, 2021.
- [32] S. Black, S. Keshavarz, and R. Souvenir, "Evaluation of image inpainting for classification and retrieval," in *IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1060–1069.
- [33] C. Wang, M. Shao, D. Meng, and W. Zuo, "Dual-pyramidal image inpainting with dynamic normalization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.
- [34] J. Li, Z. Xiong, and D. Liu, "Reference-guided landmark image inpainting with deep feature matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8422–8435, 2022.
- [35] J. Liu, M. Gong, Z. Tang, A. Qin, H. Li, and F. Jiang, "Deep image inpainting with enhanced normalization and contextual attention," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.
- [36] Q. Cheng, H. Shen, L. Zhang, and P. Li, "Inpainting for remotely sensed images with a multichannel nonlocal total variation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 175–187, 2013.
- [37] H. Shen and L. Zhang, "A map-based algorithm for destriping and inpainting of remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 5, pp. 1492–1502, 2008.
- [38] D. Lin, G. Xu, Y. Wang, X. Sun, and K. Fu, "Dense-add net: An novel convolutional neural network for remote sensing image inpainting," in *IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 4985–4988.
- [39] J. Dong, R. Yin, X. Sun, Q. Li, Y. Yang, and X. Qin, "Inpainting of remote sensing sst images with deep convolutional generative adversarial network," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 173–177, 2018.
- [40] A. Kuznetsov and M. Gashnikov, "Remote sensing image inpainting with generative adversarial networks," in *International Symposium on Digital Forensics and Security*. IEEE, 2020, pp. 1–6.
- [41] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu, "A remote sensing image dataset for cloud removal," *arXiv:1901.00600*, 2019.
- [42] H. Pan, "Cloud removal for remote sensing imagery via spatial attention generative adversarial network," *arXiv:2009.13015*, 2020.
- [43] M. Pang, B. Wang, Y.-m. Cheung, Y. Chen, and B. Wen, "Disp+ v: A unified framework for disentangling prototype and variation from single sample per person," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [44] D. Zhang, J. Zhang, Q. Zhang, J. Han, S. Zhang, and J. Han, "Automatic pancreas segmentation based on lightweight dcnn modules and spatial prior propagation," *Pattern Recognition*, vol. 114, p. 107762, 2021.
- [45] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [46] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.
- [47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [48] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *International Joint Conference on Artificial Intelligence*, 2018, pp. 684–690.
- [49] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.
- [50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
- [51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.
- [52] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
- [53] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 867–875.
- [54] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7760–7768.
- [55] X. Li, Q. Guo, D. Lin, P. Li, W. Feng, and S. Wang, "Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1869–1878.
- [56] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv:1912.02781*, 2019.
- [57] B. Xu, X. Shu, and Y. Song, "X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 3852–3867, 2022.
- [58] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," *arXiv preprint arXiv:2110.08733*, 2021.
- [59] K. Tan, Y. Zhang, and X. Tong, "Cloud extraction from chinese high resolution satellite imagery by probabilistic latent semantic analysis and object-based machine learning," *Remote Sensing*, vol. 8, no. 11, p. 963, 2016.

- [61] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.
- [62] W. Zhang, J. Zhu, Y. Tai, Y. Wang, W. Chu, B. Ni, C. Wang, and X. Yang, "Context-aware image inpainting with learned semantic priors," *arXiv preprint arXiv:2106.07220*, 2021.
- [63] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10758–10768.
- [64] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, "Pd-gan: Probabilistic diverse gan for image inpainting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9371–9381.
- [65] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [66] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [67] M. Hnewa and H. Radha, "Object detection under rainy conditions for autonomous vehicles: A review of state-of-the-art and emerging techniques," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 53–67, 2020.
- [68] Y. Pei, Y. Huang, Q. Zou, X. Zhang, and S. Wang, "Effects of image degradation and degradation removal to cnn-based image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1239–1253, 2019.
- [69] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.



**Qin Zou** received his B.E. degree in Information Engineering and Ph.D. degree in Computer Vision from Wuhan University in 2004 and 2012, respectively. From 2010 to 2011, Qin was a visiting Ph.D. student at the Computer Vision Lab, University of South Carolina. Currently, he is an Associate Professor at the School of Computer Science, Wuhan University. He is serving as Associate Editor for IEEE Transactions on Intelligent Vehicles and IET Biometrics.



**Shaoyue Song** received the B.E. degree in biomedical engineering and the Ph.D. degree in signal and information processing from Beijing Jiaotong University, Beijing, China, in 2014 and 2021, respectively. She is currently an Assistant Professor with the Faculty of Information Technology, Beijing University of Technology, Beijing, China. Her current research interests include salient object detection and image classification.



**Huiming Sun** received his B.S. degree in Automation from Chang'an University in 2019. He is currently working towards his Ph.D. degree in the Department of Electrical Engineering and Computer Science at Cleveland State University, Cleveland, OH, USA. His research interests include computer vision, deep learning, and remote sensing.



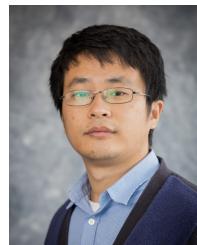
**Jin Ma** received his B.S. degree in information and computational science and M.E. degree in software engineering from Xi'an Jiaotong University, China, in 2016 and 2019 respectively. Now he is pursuing his Ph.D. degree in Computer Science at Cleveland State University. His research interests include computer vision and deep learning related topics.



**Yuewei Lin** received the B.S. degree in electrical engineering from Sichuan University, Chengdu, China; the M.E. degree from Chongqing University, Chongqing, China; and the Ph.D. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA. He is currently a Computational Scientist with Brookhaven National Laboratory, Upton, NY, USA. He is also with Stony Brook University as a research associate professor. His research interests include computer vision, machine learning, image/video analysis, and their applications for scientific data analysis.



**Qing Guo** received a Ph.D. degree in computer application technology from the School of Computer Science and Technology, Tianjin University, China. He was a research fellow with the Nanyang Technology University, Singapore, from Dec. 2019 to Aug. 2020 and Dec. 2021 to Sep. 2022. He was assigned as the Wallenberg-NTU Presidential Postdoctoral Fellow with the Nanyang Technological University, Singapore, from Sep. 2020 to Dec. 2021. He is currently a research scientist at the Center for Frontier AI Research, Agency for Science, Technology, and Research (A\*STAR), Singapore, and the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A\*STAR), Singapore. His research interests include computer vision, AI security, and image processing. He is a member of IEEE.



**Hongkai Yu** received the Ph.D. degree in computer science and engineering from University of South Carolina, Columbia, SC, USA. He is currently an Assistant Professor in the Department of Electrical Engineering and Computer Science at Cleveland State University, Cleveland, OH, USA. His research interests include computer vision, deep learning, and smart city. He is the Area Chair for ACM Multimedia 2022 and IEEE MIPR 2022 conferences. He is the Editorial Board member of the journal Green Energy and Intelligent Transportation.