Reconstruction of Large-Scale Missing Data in Remote Sensing Images Using Extend-GAN

Yongchuan Cui, Peng Liu, Bingze Song, Lingjun Zhao, Yan Ma, and Lajiao Chen

***Abstract*— Numerous studies have been conducted on missing data recovery in remote sensing images, such as cloud removal and dead pixels restoration. Nevertheless, reconstructing continuous, extensive, and complete missing areas still poses a significant challenge. In this letter, we propose a new architecture named Extend-generative adversarial network (GAN), which leverages only a low-resolution image with relaxed requirements on spatial resolution and acquisition time as a condition to reconstruct a high-resolution image with large-scale missing areas. We equip Extend-GAN with learnable adaptive region normalization (LARN) to adjust the intensity distribution of pixels to reduce color distortion. We also introduce a new loss function into the training process of Extend-GAN, namely the structural similarity (SSIM)-based triplet loss, which helps to preserve the between missing parts and known regions. Gaofen-2 and Landsat-9 image pairs are used to validate the proposed method. Extend-GAN performs better when comprehensively evaluated on visual effect, quantitative metrics, processing speed, etc. Code is available at https://github.com/yc-cui/Extend-GAN.**

***Index Terms*— Generative adversarial network (GAN), image reconstruction, remote sensing images, triplet loss.**

I. Introduction

MISSING data recovery in remote sensing images is a classical yet challenging task. Many applications, such as recovering the images of Landsat Enhanced Thematic Mapper Plus (ETM+) (scan line corrector (SLC)-off), repairing the occluded areas of clouds and shadows, or filling the region in mosaic of large-scale images, etc., are often regarded as missing data recovery problems. The nature of these problems is to estimate the missing areas and fill the vacancies with predicted pixels so that the remedied image looks visually and semantically correct and the data usability is also improved.

Early work on missing data recovery of remote sensing images can be roughly divided into three subcategories [1], [2], [3]: spatial-based, spectral-based, and temporal-based methods. However, most of these methods [1], [2], [4], [5], [6], [7] still face noteworthy challenges in practice. For example, spectral-based methods often assume that complete data is available for certain bands to reconstruct missing data in other bands; temporal-based methods often require that the acquisition time of reference images be close to the time of target images. These methods are capable of producing favorable outcomes when the reference data is reasonable. However, in cases where the region of interest presents extensive missing areas, it will severely reduce the applicability of these models since obtaining such supplementary data can be very difficult in real-world situations. Especially in large-scale image mosaics, as in Fig. 1, there are often large areas that need to be filled in. Therefore, it is essential to investigate generating large missing areas using loose reference information while not conflicting with the semantic information of target data (such as hue and feature continuity).

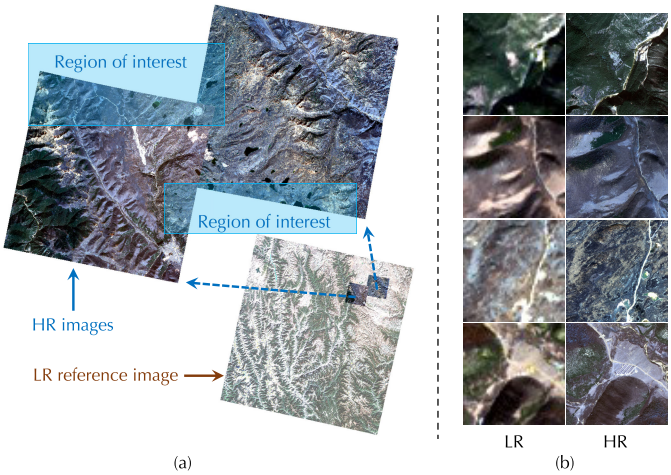

Fig. 1. (a) Regions of interest of HR images are incomplete. LR provides complementary information for the missing parts of HR. (b) Pairs of HR and LR images with the same geographic location.

In recent years, deep generative models, especially generative adversarial networks (GANs), have recently made considerable progress in remote sensing image recovery. Due to its powerful data fitting ability, GANs can integrate multi-source remote sensing data effectively (such as spatial-spectral and spatiotemporal data, even heterogeneous data) [3], [8] to reconstruct missing data more accurately. However, due to the complexity and specialty of remote sensing data, existing GAN-based methods still face challenges in feature extraction and complementation. While most of these methods [2], [6], [7] are effective in addressing small regions with missing data, their performance is limited when it comes to larger-scale continuous missing regions. They suffered from blurred edges and artifacts in the reconstruction of large areas (e.g., half of the image is completely missing), resulting in visual and semantic discontinuity.

Manuscript received 30 May 2023; revised 13 September 2023; accepted 16 September 2023. Date of current version 24 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 41971397, Grant U2243222, and Grant 42071413. *(Corresponding author: Peng Liu.)*

The authors are with the Eight Department, Aerospace Information Research Institute, School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100094, China (e-mail: liupeng@radi.ac.cn).

Digital Object Identifier 10.1109/LGRS.2023.3317898

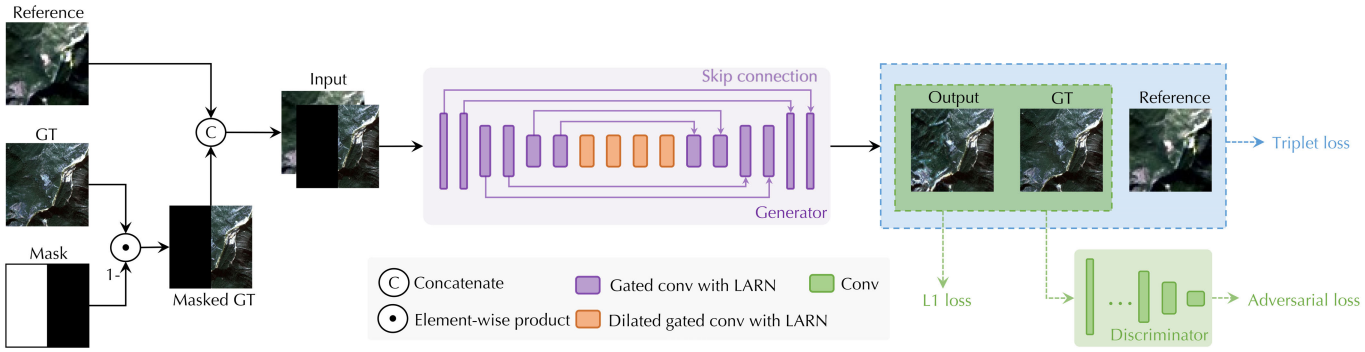

Fig. 2. Overview of Extend-GAN. The generator is an encoder–decoder architecture with skip connections. All convolution layers in the generator are equipped with LARN. Different colored dashed boxes represent different losses calculated with specific inputs.

Achieving excellent reconstruction outcomes lies in a more strategic utilization of the information inherency within both the target and reference image. In this letter, as in Fig. 1, we reconstruct the missing HR image by referring to the LR image, since the corresponding LR data for missing regions is intact and provides global information [refer to Fig. 1(a)]. The LR does not necessarily require being very close to HR in capture time or resolution in practical applications. We propose Extend-GAN, which takes LR images as conditions to reconstruct HR images with large continuous missing areas. Extend-GAN is equipped with a novel normalization and is trained by a new loss function, which significantly enhances the reconstruction quality. Our contributions can be summarized as follows.

1) We proposed the architecture of Extend-GAN, which utilizes LR images as references to effectively restore HR remote sensing images with large continuous missing areas.
2) A new normalization method, namely learnable adaptive region normalization (LARN), is armed in Extend-GAN to align statistics from missing areas to known areas.
3) To train Extend-GAN, we propose a new loss function, namely SSIM-based triplet loss. This loss function provides more reasonable constraints to the structure of reconstructed parts.

II. Methodology

In this section, we detailedly describe the proposed method. Fig. 2 depicts an overview of the procedure. It showcases the process from raw data through the generator to get output and finally calculate the loss. Firstly, Section II-A presents the network architecture. Then we elaborate on details of LARN and triplet loss in Sections II-B and II-C, respectively.

A. Extend-GAN

1) Generator: The occluded or defective image is denoted as $\mathbf{I} \in \mathbb{R}^{\mathbf{h}\times\mathbf{w}\times\mathbf{c}}$ and is formulated by the original image $\mathbf{O} \in \mathbb{R}^{\mathbf{h}\times\mathbf{w}\times\mathbf{c}}$ and a binary mask $\mathbf{M} \in \mathbb{R}^{\mathbf{h}\times\mathbf{w}\times\mathbf{1}}$ (with value 1 for unknown pixels, 0 otherwise)

$$\mathbf{I} = \mathbf{O} \odot (\mathbf{1} - \mathbf{M}) \tag{1}$$

where $\odot$ is the Hadamard product operator. Given a reference image $\mathbf{R} \in \mathbb{R}^{\mathbf{h}\times\mathbf{w}\times\mathbf{c}}$ (which is an LR image located in the same area in this letter), the objective of the generator is to learn a mapping G where $G(\mathbf{I} \mid \mathbf{R}) \rightarrow \mathbf{O}$. The output is expected to be a plausible image $\hat{\mathbf{O}} \in \mathbb{R}^{\mathbf{h}\times\mathbf{w}\times\mathbf{c}}$ which looks as identical as possible to the ground truth (GT) $\mathbf{O}$. It should be noted that we do not use randomly irregular masks, but rather randomly select half of the image to be the mask (as shown in the mask in Fig. 2). This is because random irregular masks can increase the difficulty of network training, and such masks are not suitable in cases of large continuous complete missing areas.

As demonstrated in Fig. 2, the generator G follows an encoder-decoder style and takes advantage of U-Net [9] skip connections structure to reconstruct pixel-level information. G takes both $\mathbf{I}$ and $\mathbf{R}$ as inputs, which is implemented as concatenation. Gated convolution [10] is introduced to each layer in G to learn different soft masks for different channels dynamically

$$\begin{aligned} \text{Gating}_{i,j} &= \sum\sum \mathbf{W}_G \cdot \mathbf{I} \\ \text{Feature}_{i,j} &= \sum\sum \mathbf{W}_F \cdot \mathbf{I} \\ \mathbf{O}_{i,j} &= \phi\big(\text{Feature}_{i,j}\big) \odot \sigma\big(\text{Gating}_{i,j}\big) \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ is sigmoid function and $\phi(\cdot)$ can be any activation function. $\mathbf{W}_G$ and $\mathbf{W}_F$ are two different convolutional filters [10] for masks and features, respectively. Dilated gated convolution is utilized in the innermost layers to expand the receptive field to fuse more semantic information. To address the color and content inconsistency problem, we present LARN after the gated convolution operator (see Fig. 3), which will be elaborated in Section II-B.

2) Discriminator: The discriminator D is mainly used for adversarial training. The final output of D is a probability in $0-1$ to determine whether the current image is the ground truth or not. Each convolutional layer of D applies spectral normalization. The learning objective for D is given as follows:

$$\mathcal{L}_D = \mathbb{E}_{\mathbf{O}}[\text{ReLU}(1 - D(\mathbf{O}))] + \mathbb{E}_{\hat{\mathbf{O}}}[\text{ReLU}(1 + D(\hat{\mathbf{O}}))] \tag{3}$$

where $\mathbb{E}_{\mathbf{O}}$ represents taking the expectation with respect to $\mathbf{O}$ of its expression, and ReLU is the rectified linear unit function. For G, to fool D via generated images, its corresponding learning objective is expressed as follows:

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{\hat{\mathbf{O}}}[\text{ReLU}(D(\hat{\mathbf{O}}))]. \tag{4}$$

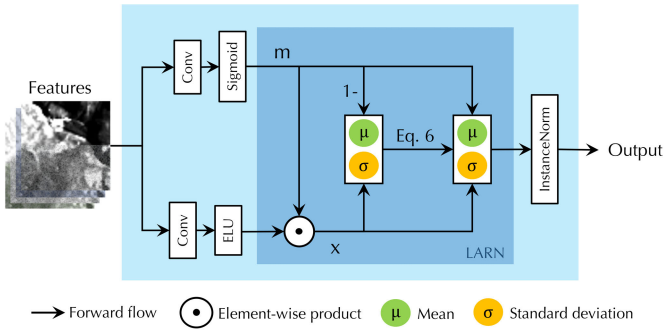

Fig. 3. Illustration of the LARN module.

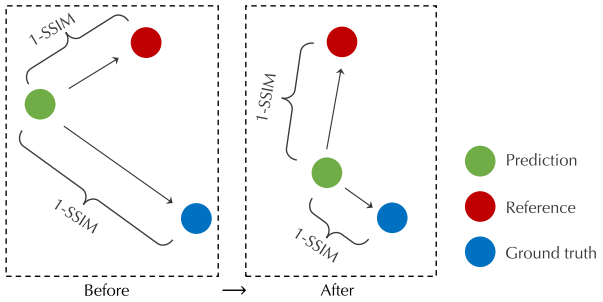

Fig. 4. Illustration of triplet loss based on SSIM. It forces the reconstructed image $\hat{\mathbf{O}}$ close to the original image $\mathbf{O}$ but not too close to the reference image $\mathbf{R}$ in terms of SSIM.

B. Learnable Adaptive Region Normalization

Because the missing areas are large, there is a significant discontinuity of color and textures between the filled part and the known part. Previous research [11], [12] has demonstrated that the mean and variation of features extracted from an image are correlated with its semantics and texture. Inspired by this, we propose an improved adaptive instance normalization (AdaIN) [12] named LARN to automatically transfer the texture style of a known region to the part to be completed. As shown in Fig. 2, each block in G is armed with LARN, which can be described as follows:

$$\mathrm{LARN}(x, m) = (1 - m) \odot x + m \odot \tilde{x} \tag{5}$$

where x and m are features and masks to be received by LARN. $\tilde{x}$ is defined as follows:

$$\tilde{x} = \gamma \cdot \sigma(x, \overline{m}) \left(\frac{x - \mu(x, m)}{\sigma(x, m)} \right) + \beta \cdot \mu(x, \overline{m}) \tag{6}$$

where γ and β are trainable parameters and $\overline{m} = 1 - m$. $\mu(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ calculate the weighted mean and standard deviation of the input elements, respectively, with m indicating the weights of each pixel to be computed. Specifically, the formulation of $\mu(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ are given as follows:

$$\mu(x, m) = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j} \cdot m_{i,j}}{\epsilon + \sum_{i=1}^{H} \sum_{j=1}^{W} m_{i,j}} \tag{7}$$

$$\sigma(x, m) = \sqrt{\frac{\sum_{i=1}^{H} \sum_{j=1}^{W} \left(x_{i,j} - \mu(x, m) \right)^2}{\epsilon + \sum_{i=1}^{H} \sum_{j=1}^{W} m_{i,j}} + \epsilon} \tag{8}$$

where ϵ is a small constant value used to maintain numerical stability. H and W are the height and width of the input feature map, respectively.

As shown in Fig. 3, LARN receives a feature map x and a soft mask $m \in [0, 1]$ with the same size as the feature map from the gated convolution. It was plugged into each block of the generator in Fig. 2. The features of missing and known regions can be calculated through x and mask m. Then, align the mean and standard deviation of known regions of each channel to unknown regions [(6)]. In order to promote the adaptability of fusion to the spectrum difference between missing and known regions, two learnable parameters, γ and β, are added to LARN to allow the network to automatically learn the ratio of mean and standard deviation shifts.

C. Triplet Loss Based on SSIM

Different from the task of super-resolution of the LR reference image, in the reconstruction of missing areas, the newly generated features need to connect with the known region by structure consistent, especially edges. With a common loss, G does not learn the structural information of the original part, resulting in unreasonable features (such as textures not consistent with the original part after the completion). This is supported by our experimental results below, which show that the structural similarity (SSIM) between the filled area and the reference image is always greater than it with the ground truth: $\mathrm{SSIM}(\hat{\mathbf{O}}, \mathbf{R}) > \mathrm{SSIM}(\hat{\mathbf{O}}, \mathbf{O})$. For this reason, we propose a triplet loss based on SSIM to guide the optimization of the network to the right direction.

The triplet loss is based on metric learning, which was first proposed in [13] as a loss for training face recognition models. As shown in Fig. 4, it utilizes a distance-based loss function to adjust the distance of embedding feature space of the positive and negative sample pair. It aims to learn embeddings that are closer for positive pairs and farther for negative ones. In this letter, we define $\langle \hat{\mathbf{O}}, \mathbf{O} \rangle$ as positive pair and $\langle \hat{\mathbf{O}}, \mathbf{R} \rangle$ as negative one. The triplet loss is expressed as follows:

$$\mathcal{L}_{\mathrm{triplet}} = \max \left(f(\mathbf{O}, \hat{\mathbf{O}}) - f(\mathbf{R}, \hat{\mathbf{O}}) + \alpha, 0 \right) \tag{9}$$

where $f(\cdot, \cdot)$ is defined as follows:

$$f(\mathbf{X}, \mathbf{Y}) = 1 - \mathrm{SSIM}(\mathbf{X}, \mathbf{Y}). \tag{10}$$

In (9), the distance in embedding space between positive and negative samples can be controlled by adjusting the hyper-parameter α. For a certain sample, if $f(\mathbf{O}, \hat{\mathbf{O}}) - f(\mathbf{R}, \hat{\mathbf{O}}) + \alpha < 0$, that means the SSIM of $\hat{\mathbf{O}}$ and $\mathbf{O}$, even if α is subtracted, is still larger than that of $\hat{\mathbf{O}}$ and $\mathbf{R}$. In this situation, the loss after taking maximum is 0, which means that this sample has learned the structural information from the ground truth, and there is no need to continue optimizing. If $f(\mathbf{O}, \hat{\mathbf{O}}) - f(\mathbf{R}, \hat{\mathbf{O}}) + \alpha > 0$, then there is still room for optimization, and taking maximum of (9) will result in a positive value of the loss, and we can continue to optimize this portion to make it smaller or equal to 0.

Besides $\mathcal{L}_{\mathrm{adv}}$ and $\mathcal{L}_{\mathrm{triplet}}$, We adopt the ℓ_1 distance between $\mathbf{O}$ and $\hat{\mathbf{O}}$ as the reconstruction loss, formulated as follows:

$$\mathcal{L}_{\mathrm{rec}} = \| \mathbf{O} - \hat{\mathbf{O}} \|_1. \tag{11}$$

TABLE I
QUANTITATIVE EVALUATION RESULTS OF DIFFERENT METHODS. BOLD TEXTS INDICATE THE BEST. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

Models	Metrics								Time(ms)↓	Params(M)↓	MACs(G)↓
	PSNR↑	SSIM↑	SSIM-D↑	CC↑	UQI↑	SAM↓	ERGAS↓	MAE↓			
Shao *et al.* [6]	24.6481	0.7424	0.0448	0.6744	0.5169	0.0545	62.3787	0.0291	27.558	7.9726	50.161
Boundless [14]	25.4584	0.7768	0.0250	0.7264	0.5126	0.0509	56.2431	0.0269	26.947	10.719	35.805
MISF [15]	26.5151	0.7999	0.0419	0.7726	0.5168	0.0453	50.1720	0.0241	28.334	25.846	148.05
HAN [16]	26.5576	0.7915	0.0551	0.7824	0.5490	0.0430	50.3666	0.0232	28.147	**7.6400**	32.521
Ours	**27.1135**	**0.8011**	**0.0766**	**0.8062**	**0.5696**	**0.0396**	**47.3426**	**0.0216**	**26.551**	10.461	**30.029**

In summary, the joint loss for the generator is written as follows:

$$\mathcal{L} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{triplet}}\mathcal{L}_{\text{triplet}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} \tag{12}$$

where λ_{adv}, λ_{triplet}, and λ_{rec} are the tradeoff parameters, and we empirically set $\lambda_{\text{adv}} = 0.01$, $\lambda_{\text{triplet}} = 0.1$ and $\lambda_{\text{rec}} = 1$.

III. EXPERIMENTS

A. Experimental Setup

1) Dataset: We use Gaofen-2 (GF2) HR images and Landsat-9 (LC9) LR images to test the model. Examples of image pairs of the training data are shown in Fig. 1(b). After registration, training image pairs were cropped to patches with a size of 512 × 512 with the same geographic extent. A total of 609 GF2-LC9 image pairs were used for training and 68 images for testing.

2) Implementation Details: All the experiments are performed with one Nvidia GeForce RTX 4090 GPU. The network architecture is implemented with PyTorch v1.13.1. We use the Adam optimizer (with $\beta_1 = 0.5$ and $\beta_2 = 0.9$) to update the weights of the network iteratively. The initial learning rate of the generator and discriminator are set to $1e-5$ and $1e-4$, respectively. α in $\mathcal{L}_{\text{triplet}}$ is set to 0.1. Models are trained with a minibatch size of 8 for 2400 epochs. When training, image pairs with size 512 × 512 are randomly cropped to 256 × 256, and then flip horizontally or vertically with a probability of 0.5. We do not use extra optimization tricks. All the models are trained under the same settings for a fair comparison. The average value of the various metrics is computed after 5 iterations of this procedure.

3) Metrics: For quantitative comparisons, the peak signal-to-noise ratio (PSNR), the SSIM, the correlation coefficients (CCs), the relative dimensionless global error in synthesis index (ERGAS), the universal quality index (UQI), mean absolute error (MAE), the spectral angle mapper (SAM) index are employed as full-reference metrics. To assess whether the model has successfully extracted the structural information of GT, we use the term SSIM-D which is formulated as $\text{SSIM-D}(\mathbf{O}, \mathbf{R}, \hat{\mathbf{O}}) = \text{SSIM}(\mathbf{O}, \hat{\mathbf{O}}) - \text{SSIM}(\mathbf{R}, \hat{\mathbf{O}})$ to represent the extent to which the predicted image is more similar to GT compared to the reference image (the larger the indicator, the better).

B. Experimental Results

The proposed method is compared with four advanced models based on neural networks, including Shao et al. [6], Boundless [14], MISF [15] and HAN [16]. To make a fair comparison, we use the same data loader when conducting experiments for these models.

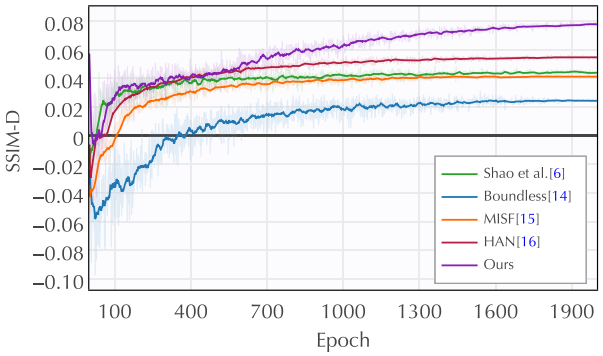

Fig. 5. Convergence curve of SSIM-D. The upper is better.

1) Comparison of Model Performance: Table I shows the evaluation results, where the proposed method outperforms the other approaches on all indicators. The model proposed by Shao et al. [6] yields relatively lower performance. This suggests that this model may not be suitable for large missing areas. The PSNR can be improved by 0.6 dB compared with the second-best model. The convergence curve of SSIM-D in masked areas is demonstrated in Fig. 5. With triplet loss and LARN, images that are structurally more similar to GT are significantly improved. This indicates that our model extracts the structural information of GT, rather than simply replicating the contents of the reference image to the missing areas.

Speed is also crucial for recovering large-scale missing data. In Table I, our model exhibits the shortest time compared with other models. In terms of the number of parameters and computational cost, the proposed model is with a much smaller size and floating point operations. Our full model has 10.461 M parameters and costs around 26.551 ms to process a 256 × 256 image.

The visual comparison of the testing data is presented in Fig. 6. It can be observed that our model outperforms the other models in synthesizing textures and edges, as evidenced by more visually convincing results with fewer artifacts. Conversely, the other models appear to neglect such structural information in their synthesis process. According to the error heatmap, it can be seen that our model has a smaller reconstruction error, which further illustrates the superiority of the proposed method.

2) Ablation Study: We verify the necessity of LARN and triplet loss through ablation experiments. As is demonstrated

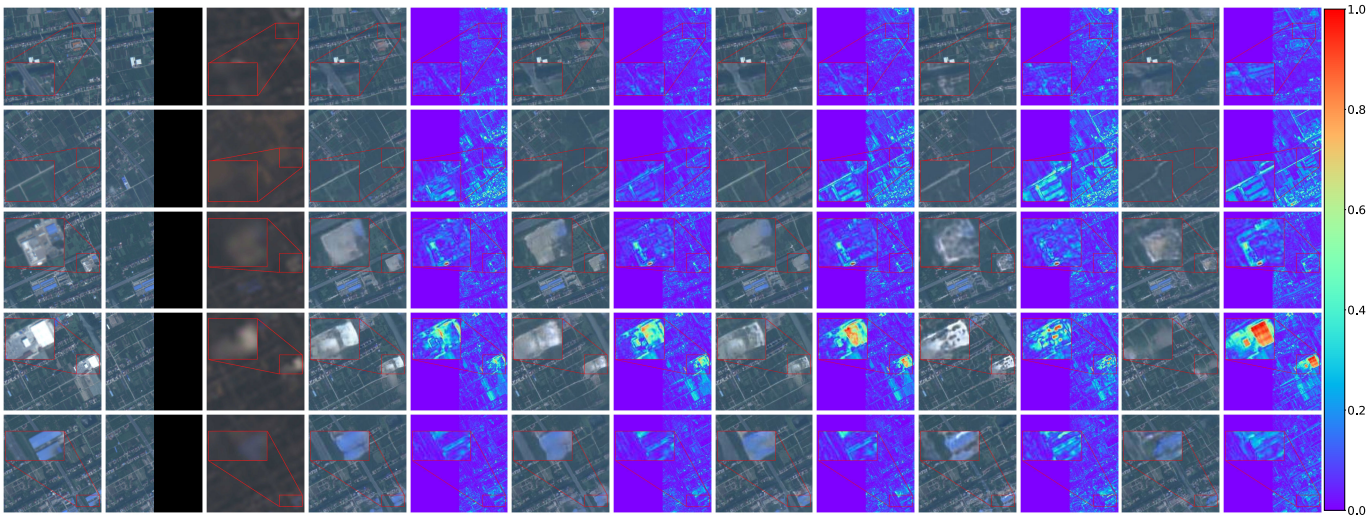

Fig. 6. Qualitative results of different methods. From left to right: (a) GT, (b) Masked GT, (c) Reference. Results: (d) Ours, (e) Error map of ours, (f) HAN [16], (g) Error map of HAN [16], (h) MISF [15], (i) Error map of MISF [15], (j) Boundless [14], (k) Error map of Boundless [14], (l) Shao et al. [6], and (m) Error map of Shao et al. [6]. **Please zoomed-in view for detailed comparison**.

TABLE II

ABLATION STUDY OF LARN AND TRIPLET LOSS. ↓: LOWER IS BETTER. ↑: HIGHER IS BETTER

Modules		Metrics							
LARN	$\mathcal{L}_{triplet}$	PSNR↑	SSIM↑	SSIM-D↑	CC↑	UQI↑	SAM↓	ERGAS↓	MAE↓
✗	✗	26.4625	0.7348	0.0536	0.7847	0.5633	0.0417	50.2091	0.0241
✗	✓	26.5942	0.7950	**0.0768**	0.7965	0.5453	0.0442	50.7563	0.0230
✓	✗	26.9266	0.7884	0.0695	0.7794	0.5621	0.0404	48.1925	0.0231
✓	✓	**27.1135**	**0.8011**	0.0766	**0.8062**	**0.5696**	**0.0396**	**47.3426**	**0.0216**

in Table II, if only triplet loss is employed, SSIM and SSIM-D will raise dramatically (SSIM + 6%, SSIM-D + 2%). When only LARN is employed, most indicators also improve. When both LARN and triplet loss are applied, all indicators are further improved (e.g.,, PSNR + 0.7 dB).

IV. CONCLUSION

In this letter, a new GAN-based network, namely Extend-GAN, is designed for large-scale missing information generation of remote sensing images. We propose LARN for adaptive style transfer and a triplet loss based on SSIM to preserve the structure of the original image. The image pairs from Gaofen-2 and Landsat-9 satellites are used to construct training and testing datasets to validate the proposed method. Both the spatial resolution and spectral characteristics are different between the target and reference image, which reduces some restrictions for proposed methods. The comprehensive experiments, including an ablation study, are carried out to compare our method with the other two methods. Simulated and real-world scene reconstruction experiments all highlight the superior speed and accuracy of our method.

REFERENCES

[1] H. Shen et al., "Missing information reconstruction of remote sensing data: A technical review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 61–85, Sep. 2015.

[2] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, "Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4274–4288, Aug. 2018.

[3] P. Liu, J. Li, L. Wang, and G. He, "Remote sensing data fusion with generative adversarial networks: State-of-the-art methods and future research directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 295–328, Jun. 2022.

[4] P. Duan, S. Hu, X. Kang, and S. Li, "Shadow removal of hyperspectral remote sensing images with multiexposure fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5537211.

[5] Q. Wang, L. Wang, Z. Li, X. Tong, and P. M. Atkinson, "Spatial–spectral radial basis function-based interpolation for Landsat ETM+ SLC-off image gap filling," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7901–7917, Sep. 2021.

[6] M. Shao, C. Wang, T. Wu, D. Meng, and J. Luo, "Context-based multiscale unified network for missing data reconstruction in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[7] Y. Zi, F. Xie, X. Song, Z. Jiang, and H. Zhang, "Thin cloud removal for remote sensing images using a physical-model-based CycleGAN with unpaired data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[8] B. Song et al., "MLFF-GAN: A multilevel feature fusion with GAN for spatiotemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410816.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. Conf.*, Nov. 2015, pp. 234–241.

[10] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4470–4479.

[11] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[12] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[14] D. Krishnan et al., "Boundless: Generative adversarial networks for image extension," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10520–10529.

[15] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1869–1878.

[16] Y. Deng, S. Hui, R. Meng, S. Zhou, and J. Wang, "Hourglass attention network for image inpainting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 483–501.