RSINET: INPAINTING REMOTELY SENSED IMAGES USING TRIPLE GAN FRAMEWORK

Advait Kumar[1†*]   *Dipesh Tamboli*[‡*]   *Shivam Pande*[2†]   *Biplab Banerjee*[2†]

[1] Dept. of Electrical Engineering   [2] Centre of Studies in Resource Engineering
[†]Indian Institute of Technology Bombay
[‡]School of Electrical and Computer Engineering, Purdue University

ABSTRACT

We tackle the problem of image inpainting in the remote sensing domain. Remote sensing images possess high resolution and geographical variations, that render the conventional inpainting methods less effective. This further entails the requirement of models with high complexity to sufficiently capture the spectral, spatial and textural nuances within an image, emerging from its high spatial variability. To this end, we propose a novel inpainting method that individually focuses on each aspect of an image such as edges, colour and texture using a task specific GAN. Moreover, each individual GAN also incorporates the attention mechanism that explicitly extracts the spectral and spatial features. To ensure consistent gradient flow, the model uses residual learning paradigm, thus simultaneously working with high and low level features. We evaluate our model, alongwith previous state of the art models, on the two well known remote sensing datasets, Open Cities AI and Earth on Canvas, and achieve competitive performance. The code can be referred here: *https://github.com/advaitkumar3107/RSINet*.

***Index Terms*—** Image inpainting, remote sensing, generative adversarial networks

1. INTRODUCTION

Image inpainting is the process of filling in the missing part or conserving the damaged and deteriorated image (which can be physical or digital). Digital image inpainting is an important problem statement in the field of computer vision which has applications in various domains such as restoring damaged images/videos, remote sensing, object removal, text removal, automatic modifications of images/videos, image compression and super resolution [1]. The drawback of the traditional mathematical methods is their low PSNR on complex images such as remote sensing images [2]. Remote sensing images have a variety of boundaries, objects and colours which makes it a semantically tougher problem. Hence, deep learning has been introduced earlier to tackle this challenge such as in [2], which solved three missing information tasks

in remote sensing data using a deep convolutional network combined with spatio-temporal information.

There have been several image inpainting approaches as well, such as [3], which used a UNet based encoder-decoder structure coupled with residual learning to compute the gradients effectively for image restoration. [4] proposed a new type of layer called 'partial convolution' (PartialConv) to improve the current best performing image inpainting models. [5] proposed the EdgeConnect (EC) model, that initially builds up the outline of the image to be restored and then fills it with details. Furthermore, to ensure the relationship between the neighbouring areas and entire image as a whole, [6] proposed a multi GAN approach, that trains a global as well as a local discriminator (GLCIC) on top of their VGG-type model. Moreover, [7] proposed a global and local attention based model to tackle the issue of image irregularities such as holes. In addition, to preserve the semantic style of the original images, additional loss functions have been explored. For instance, [8] introduces style loss, where the textural features are synthesised from the images, which assist in the style transfer. [9] introduces perceptual loss for super-resolution, (instead of pixelwise loss) and get high resolution images.

In the aforementioned methods, even though the methods seemed to perform better individually on the task specific images, they could not prove much effective on the images with high spatial variation such as those from remote sensing domain. This is because each of the individual models was designed to focus on a specific aspect of the image such as colour, edge or texture. Hence, there arises a need to have a common model that simultaneously works with all the image aspects and leads to more accurate image restoration. Inspired from this notion, we propose an image inpainting approach that utilizes multiple GANs to effectively capture the different aspects of the image. In addition, to handle irregularities of the image defects, Convolutional Block Attention Module (CBAM) [10] and the Gated Attention Layers [11] based attention modules are added. The GANs are reinforced with skip connections to ensure a consistent gradient flow. Moreover, to preserve the semantic style of the images and get more robust representation, we incorporate a cocktail of adversarial, style and perceptual loss. Our approach is summarised below:

---

*[*]Equal contribution author
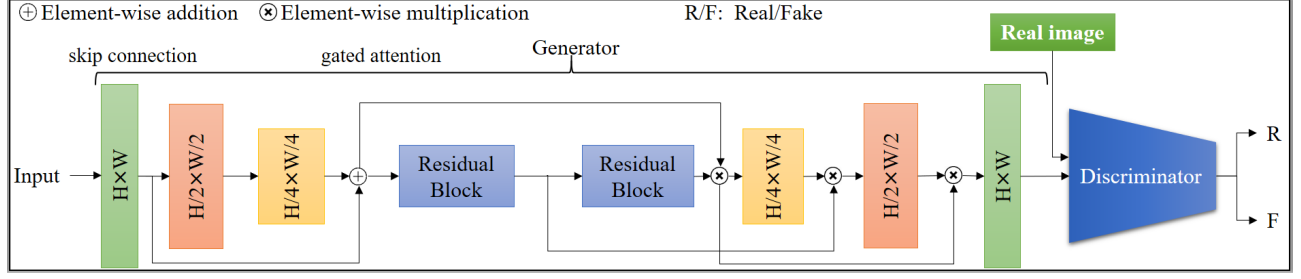
Fig. 1: Architecture of our proposed GAN. The Generator consists of an encoder-residual blocks-decoder structure. The encoder has a skip connection as well as feature maps that serve as a gate for the gated attention layers. The discriminator consists of 6 convolutional layers that successively downsample the input into a 28×28 probabilistic grid.

1. We introduce a model that combines the characteristics of edge detection GAN, colour filling GAN and global GAN to focus on the different image aspects.

2. We introduce attention layers for guided backpropagation to get more robust spectral-spatial representation and increase the sharpness of the images. We also include style and perceptual losses with the adversarial loss to capture the semantic information of the images.

3. We train our model on the Open Cities AI [12] and the Earth on Canvas [13] datasets where our model outperforms the existing deep inpainting models by at least 0.06 % and 2.24 % respectively.

2. MODEL DESCRIPTION

In this section we discuss our proposed model. It consists of three GANs trained on top of each other, namely, the edge completion GAN $\mathcal{G}_1$, colour filling GAN $\mathcal{G}_2$ and the global GAN $\mathcal{G}_3$. We describe the loss functions and the architecture used for the basic GAN unit in detail in the subsequent sections. The GAN architecture used in RSINet is presented in Fig. 1. All the GANs follow the similar architecture.

Let I_{gt} be the groundtruth image and, C_{gt} and I_{gray} be the groundtruth edge map and grayscale image respectively. In the generator for $\mathcal{G}_1$, we use the masked grayscale image $\tilde{I}_{gray} = I_{gray} \odot (1 - M)$, its masked edge map $\tilde{C}_{gt} = C_{gt} \odot (1 - M)$ ($\odot$ being the Hadamard product), and the image masks M as the pre-condition (proposed in [14] as input to pix2pix GAN, where 1 and 0 denote the masked region and the background respectively). The generator produces the completed edge map for the image (see Eqn. 1).

$$C_{pred} = \mathcal{G}_1(\tilde{I}_{gray}, \tilde{C}_{gt}, M) \tag{1}$$

We use C_{gt} and C_{pred} conditioned on I_{gray} as the input to the discriminator, which predicts whether or not the edge map is real. This predicted edge map is then passed on to the colour filling model (Eqn. 2). The colour filling model uses the masked RGB image, $\tilde{I}_{gt} = I_{gt} \odot (1 - M)$, conditioned on the completed edge map $C_{comp} = C_{gt} \odot (1 - M) + C_{pred} \odot M$, as taken from the previous model. The model outputs the completed RGB image with the right colours filled at the right places.

$$I_{pred} = \mathcal{G}_2 \left(\tilde{I}_{gt}, C_{comp} \right) \tag{2}$$

$\mathcal{G}_1$ and $\mathcal{G}_2$ individually gave blurry outputs on the tougher satellite images. Hence, we train a global GAN $\mathcal{G}_3$ (Eqn. 3) on top of the output from the colour filling model which refines the output and makes the images sharper. This was based upon the suggestion that using the same encoder-decoder architectures for image inpainting gave better results since both the models learnt the same features semantically [15].

$$I_{refined} = \mathcal{G}_3(I_{comp}, \tilde{I}_{gt}) \tag{3}$$

2.1. Loss Functions

The edge completion model is trained the adversarial loss [14] and the feature matching loss [16]. Adversarial training is thought of as a min-max game between two players. Here, the generator tries to 'fool' the discriminator, i.e. make the discriminator predict with a high probability that the generator's output belongs to the input data, while the discriminator is trained to differentiate between the original and the generated samples. The adversarial loss is given in Eqn. 4.

$$L_{adv} = \mathbb{E}_{C_{gt}, I_{gray}} \left[\log \mathcal{D}_1(C_{gt}, I_{gray}) \right] + \\ \mathbb{E}_{I_{gray}} \left[\log(1 - \mathcal{D}_1(C_{pred}, I_{gray})) \right] \tag{4}$$

In Eqn. 4, $\mathcal{D}_1$ is the discriminator for edge completion GAN.

The feature matching loss, L_{FM} compares the activation maps in the intermediate layers of the discriminator. This is similar to perceptual loss [9] (used in the colour filling model), where activations are compared with those from a pre-trained VGG network. However, a VGG network is not trained to produce the edge maps, so we use the L_{FM} instead of the L_{perc} here. The L_{FM} is in Eqn. 5 defined as :

$$L_{FM} = \mathbb{E} \left[\sum_{i=1}^{L} \frac{1}{N_i} \left\| \mathcal{D}_1^{(i)} \left(C_{gt} \right) - \mathcal{D}_1^{(i)} \left(C_{pred} \right) \right\|_1 \right] \tag{5}$$

144

Here L is the number of layers in the discriminator. N_i is the number of elements in the ith convolutional layer and $\mathcal{D}_1^{(i)}$ is the activation in the ith layer of the $\mathcal{D}_1$.

The colour filling model is trained on 4 losses. The adversarial loss, L1 loss [17], style loss [8] and perceptual loss [9]. To ensure proper scaling the L1 loss is normalized by the mask size. The adversarial loss is similar to the previous model. The perceptual loss gives the L1 distance between the activations from a few specific layers of a pre-trained network as well as our model (Eqn. 6).

$$L_{perc} = \mathbb{E}\left[\sum_{i=1} \frac{1}{N_i} \|\phi_i\left(I_{gt}\right) - \phi_i\left(I_{pred}\right)\|_1\right] \quad (6)$$

Here ϕ_i represents the activation map of the ith layer of the pre-trained VGG19 network.

Style loss returns the differences in covariances of these activation maps by constructing the gram matrix from the activation maps. Given the feature maps of sizes $C_j \times H_j \times W_j$, style loss is computed in (Eqn. 7).

$$L_{style} = \mathbb{E}_j\left[\left\|G_j^\phi\left(\tilde{I}_{pred}\right) - G_j^\phi\left(\tilde{I}_{gt}\right)\right\|_1\right] \quad (7)$$

Here G_j^ϕ is the gram matrix constructed from activation maps ϕ_j. The global GAN model is also trained on the weighted sum of the 4 losses as discussed above, namely, L_{adv}, L_{perc}, $L1$ and L_{style}. The final loss for RSINet is given in Eqn. 8. The value of λ_1, λ_2 and λ_3 is empirically fixed to 1 during implementation.

$$L_{final} = L_{adv} + \lambda_1 L1 + \lambda_2 L_{perc} + \lambda_3 L_{style} \quad (8)$$

3. EXPERIMENTS

The model has a VGG type architecture [18] (an encoder followed by a decoder). The encoder part consists of 3 convolutional layers each successively halving the image dimensions and doubling the number of channels, followed by 8 residual blocks and finally 3 convolution layers again. This converts the feature representation of the residual blocks back to the image size (the output). A skip connection (consisting of 2 layers) has been added to the second convolutional layer of the network. We have also added 3 gated attention layers in between, for further refinement of the feature maps.

3.1. Datasets

We worked primarily with two famous satellite imagery datasets. The first one is the Open Cities AI challenge dataset [12], which consists 500 randomly sampled images size of 1024×1024, divided into 16 images of size 256×256. The second one is the Earth on Canvas dataset [13] which consists of 1400, 256×256 images from 14 classes. It is highly uncorrelated and a good test for our model's performance. Both the datasets have been divided into 60%, 20% and 20% ratio for train, validation and test sets respectively.

3.2. Protocol

We kept the parameters for the global GAN similar to those of colour filling GAN. Adam optimizer [19] is used for training all the models, with learning rates of 10^{-3} and 10^{-4} for generator and discriminator respectively. For evaluation, peak signal-to-noise ratio (PSNR) is used. For the GLCIC model, its 160×160 output was interpolated to 256×256 while computing the PSNR.

Table 1: Accuracy analysis for irregular masks on Open Cities AI and Earth on Canvas datasets for different methods

| Model Name	PSNR (DB)	
	Open Cities AI	Earth on Canvas
EC original [5]	37.113	33.145
EC (our model)	37.433	33.784
Only colour filling GAN (our model)	30.347	31.198
PartialConv [4]	35.912	31.246
EC+Global GAN (our model)	**37.456**	**34.542**

3.3. Discussions and Critical Analysis

The performance for all the models is presented in Table 1. It is visible that our model outperforms all the models for Open Cities AI dataset (PSNR 37.456) and Earth on Canvas dataset (PSNR 34.542).

We also deal with 3 types of ablations; *rectangular masks* wherein a random rectangle of the input image was replaced with white pixels, covering between 5%-30% of the image area during training, *salt and pepper masks* where, random image pixels as sampled from a gaussian distribution were whitened, covering anywhere between 5%-95% of the image area and *irregular masks*, as shown in [4]. Furthermore, we also present the visualizations of inpainting for the different masks in Fig. 2 for the three kinds of masks. The method has shown a good performance in reconstructing the images for all the three kinds of masks.

3.3.1. Effect of Loss Function

As seen from Fig. 3, a mixture of losses works best for inpainting (presented on Open Cities AI). This hybrid loss function consisting of perceptual, style, L1 and adversarial losses for the generator, and feature matching and adversarial losses for the discriminator performs well for this semantically tough image inpainting task. Perceptual and style losses help in improving the quality of the reconstructed image by capturing image semantics and minimising the difference of activations in all the layers of the target and the output images. L1 loss improves the model performance due to its high noise robustness. Feature matching loss used in the discriminator helps to keep the GAN stable and prevent mode collapse.

145
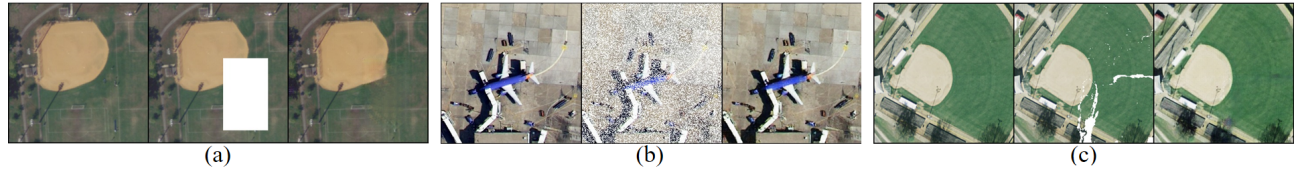
Fig. 2: Inpainting results on Earth on Canvas dataset with (a) rectangular mask, (b) salt-pepper noise, (c) irregular mask. For each example, the image on the left is the original image, the image in the centre has been corrupted by a mask while the one in the right is the reconstructed image after inpainting.
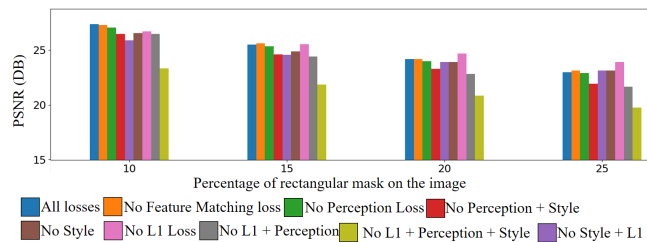
Fig. 3: Performance of various loss functions presented on Open Cities AI dataset. It can be observed that relatively better performance is observed when the model is trained on all the losses.

4. CONCLUSION

We present a novel approach to tackle the problem of image inpainting in remote sensing domain. Our approach builds on the existing EdgeConnect model and strengthens it by incorporating attention mechanism and global GAN, that gives realistic results by combining global and local features. Furthermore, we incorporate multiple losses such as perceptual loss and style loss, which further boost our model's performance. We evaluate our model on Open Cities AI and Earth on Canvas datasets, where our approach gives competitive results with respect to previous benchmarks. In future, we would consider more semantically challenging datasets and explore the problem in a multimodal scenario.

Acknowledgement

5. REFERENCES

[1] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari, "Image inpainting: A review," *Neural Processing Letters*, pp. 1–22, 2019.

[2] Junyu Dong, Ruiying Yin, Xin Sun, Qiong Li, Yuting Yang, and Xukun 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.

[3] Yang Liu, Jinshan Pan, and Zhixun Su, "Deep blind image inpainting," 2017.

[4] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro, "Image inpainting for irregular holes using partial convolutions," 2018.

[5] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," 2019.

[6] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, July 2017.

[7] S. M. Nadim Uddin and Yong Jung, "Global and local attention-based free-form image inpainting," *Sensors*, vol. 20, pp. 3204, 06 2020.

[8] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge, "Image style transfer using convolutional neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016.

[10] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," 2018.

[11] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018.

[12] Milind Naphade, Zheng Tang, Ming-Ching Chang, David C Anastasiu, Anuj Sharma, Rama Chellappa, Shuo Wang, Pranamesh Chakraborty, Tingting Huang, Jenq-Neng Hwang, et al., "The 2019 ai city challenge," in *CVPR Workshops*, 2019, pp. 452–460.

[13] Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, and Mihai Datcu, "A zero-shot sketch-based intermodal object retrieval scheme for remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2021.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," 2016.

[15] Suriya Singh, Anil Batra, Guan Pang, Lorenzo Torresani, Saikat Basu, Manohar Paluri, and CV Jawahar, "Self-supervised feature learning for semantic segmentation of overhead imagery.," in *BMVC*, 2018, vol. 1, p. 4.

[16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," 2016.

[17] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz, "Loss functions for neural networks for image processing," 2015.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.