ADVERSARIAL INPAINTING OF MEDICAL IMAGE MODALITIES

Karim Armanious[1,2], Youssef Mecky[1,3], Sergios Gatidis[2], Bin Yang[1]

[1]University of Stuttgart, Institute of Signal Processing and System Theory, Stuttgart, Germany
[2]University of Tübingen, Department of Radiology, Tübingen, Germany
[3]German University in Cairo, Faculty of Information Engineering and Technology, Cairo, Egypt

ABSTRACT

Numerous factors could lead to partial deteriorations of medical images. For example, metallic implants will lead to localized perturbations in MRI scans. This will affect further post-processing tasks such as attenuation correction in PET/MRI or radiation therapy planning. In this work, we propose the inpainting of medical images via Generative Adversarial Networks (GANs). The proposed framework incorporates two patch-based discriminator networks with additional style and perceptual losses for the inpainting of missing information in realistically detailed and contextually consistent manner. The proposed framework outperformed other natural image inpainting techniques both qualitatively and quantitatively on two different medical modalities.

***Index Terms*—** Magnetic resonance imaging, computed tomography, medical image inpainting, deep learning, GANs

1. INTRODUCTION

Medical imaging is a fundamental tool for diagnostic procedures. Nevertheless, causes for missing or incomplete image information in medical scans are manifold including image artifacts (e.g. metal artifacts in CT and MRI), limited field of view, selective reconstruction of acquired data or superposition of foreign bodies in projection methods. It is clear that missing image information cannot be retrieved in a diagnostic sense, meaning that the actual information is lost. However, for image post processing, completing missing information within medical scans is also of interest.

One example is attenuation correction in PET/MRI, where MR data are used for the estimation of attenuation coefficients. In this case, it is not the detailed local properties of MR data that are required but a more global property. Here, the correction of missing body parts (e.g. due to artifacts or positioning outside the MR field of view [1]) via inpainting can be of high value [2]. In a similar way, inpainting can be advantageous in radiation therapy planning for the correction of MR artifacts before calculation of dose distribution. In general, completion of medical images is of interest whenever automated algorithms for image analysis shall be applied (e.g. for segmentation or classification) that require a complete, artifact-free input. Thus, inpainting can also be part of data curation frameworks in medical imaging.

Current approaches for medical image inpainting rely on texture synthesis [3], interpolation [4, 5], non-local means [6] and diffusion techniques [7]. These classical approaches face difficulty when inpainting more complex regions such as in medical imaging data.

From another perspective, the inpainting of natural images is a hot topic of research in the computer vision community. This is especially true while utilizing GANs [8]. Context encoders (CE) are one of the most widely used natural image inpainting techniques [9]. They are based on training an encoder-decoder network with an adversarial discriminative network. However, the resultant inpainted regions may not always be consistent with their surrounding regions. The Globally and Locally Consistent Image Completion (GLCIC) builds upon CE by expanding the discriminator network into a multi-scale approach [10]. This is achieved by fusing the learned discriminative features from a global discriminator network with those from a more local network before a discriminative decision is taken. However, to ensure consistency with the surrounding regions, further post-processing methods and long training durations are recommended. In [11], instead of post-processing, the consistency of the inpainted regions is improved by separating the two discriminator networks and using a parsing network to enhance the results. Other proposed inpainting techniques include utilizing contextual attention [12], perceptual loss [13] or super-resolution methods [14] among others [15]. A complete overview of such methods is outside the scope of this work.

In this work, we introduce the topic of medical image inpainting using deep learning techniques. As a baseline, we utilize our recently proposed MedGAN framework for medical image translation tasks [16]. MedGAN is an adversarial framework combining a cascaded U-net generator architecture (CasNet [16, 17]) with a new combination of non-adversarial losses. However, we argue that an inpainting task is more challenging than a translation task. This is because not only the inpainted region must be highly realistic, but also it must fit homogeneously into the given context information. Motivated by the recent advances in natural image inpainting

[10, 11], we expand MedGAN with an additional local discriminator network to enhance the inpainting performance.

Our new model, named ip-MedGAN, produces globally consistent and realistic results without the need for further post-processing. We demonstrate the model performance on different medical modalities, MRI and CT. Furthermore, we compare qualitatively and quantitatively with other adversarial inpainting methods.

2. METHODS

Our model, ip-MedGAN, is based on a conditional GAN (cGAN) architecture with the inclusion of a patch-based local discriminator network and additional non-adversarial losses. In Fig.1 an outline of the proposed model is presented.

2.1. Conditional Generative Adversarial Networks

A cGAN framework consists of two convolutional networks, the generator G and the global discriminator D [18]. In the proposed framework, G receives as input the context image y. It is a 2D medical image of size 256×256 with a randomly cropped square region of size 64×64. Thus, the missing portion of the image is $\frac{1}{16}$ of the original image size. The generator utilizes the given context information to inpaint the missing region and to form a synthetically completed image $\hat{x}$. On the other hand side, the discriminator receives as input the target image x with no missing information or the generated image $\hat{x}$. It utilizes a binary cross entropy loss function to classify which of input images is a synthetic output from the generator, $D(\hat{x}, y) = 0$, and which belongs to the real target distribution, $D(x, y) = 1$. The networks are trained via a game-theoretical approach where the generator attempts to fool the discriminator into misclassifying $\hat{x}$ as a real image, while the discriminator constantly improves its classification performance to avoid being fooled. The following min-max optimization task represents this training procedure:

$$\min_G \max_D \mathcal{L}_{\text{adv}} = \mathbb{E}_{x,y}\left[\log D(x,y)\right] + \mathbb{E}_{\hat{x},y}\left[\log\left(1 - D\left(\hat{x},y\right)\right)\right] \tag{1}$$

where $\mathcal{L}_{\text{adv}}$ is the adversarial loss function.

For the generator network, a CasNet architecture is utilized which cascades multiple U-net networks, with batch normalization and skip-connections, in an end-to-end manner. This is utilized to distribute the generative task over the more extensive network and thus produce more detailed outputs. CasNet has been shown as an effective method of increasing the overall network capacity and stabilizing the training while avoiding depth-related problems such as vanishing gradients and exponential increase in the number of parameters [17]. Further architectural details are presented in [16].

The discriminator network is identical to the architecture proposed in [19]. It is a patch discriminator which divides the input images into overlapping 70×70 patches, before classifying each patch as real or fake. For the final classification decision, the score from all patches is averaged out. We consider this discriminator network as a global discriminator because it takes as input the complete output and target images and not just the inpainted and target patches. By focusing on smaller patches which collectively span the entire image, the global discriminator ensures that the inpainted regions fit homogeneously into the given context information.

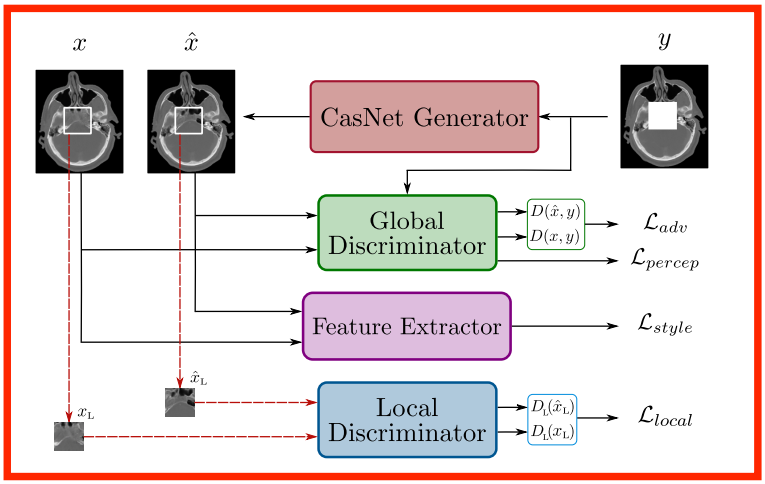

Fig. 1. An overview of the proposed ip-MedGAN framework utilized for the inpainting of medical image modalities.

2.2. Patch-Based Local Discriminator

Inspired by recent natural image inpainting techniques, the proposed model extend MedGAN by including an additional discriminator network titled the local discriminator D_{L} [10, 11]. In contrast to the global discriminator, D_{L} receives as input only the inpainted and target regions, $\hat{x}_{\text{L}}$ and x_{L} respectively. This allows the local discriminator to focus on the details of the inpainted region rather than on the global context information in the complete image. D_{L} is also a patch-based network which divides the input regions in 34×34 overlapping patches for classification. It is trained in an adversarial setting along with the generator network analogous to Eq. 1:

$$\min_G \max_{D_{\text{L}}} \mathcal{L}_{\text{local}} = \mathbb{E}_{x_{\text{L}}}\left[\log D_{\text{L}}(x_{\text{L}})\right] + \mathbb{E}_{\hat{x}_{\text{L}}}\left[\log\left(1 - D_{\text{L}}\left(\hat{x}_{\text{L}}\right)\right)\right] \tag{2}$$

2.3. Non-Adversarial Losses

To improve the inpainted results, additional non-adversarial losses are utilized to train the generator network. The first is the style reconstruction loss which guides the generator to match the style and textures of the target images x onto the generated output $\hat{x}$ [20, 21]. This loss is calculated using intermediate features maps extracted from a pre-trained feature extractor network. A VGG-19 network pre-trained on the ImageNet classification task is utilized [22]. The extracted feature maps are used for the calculation of the Gram matrices, $G_n(x)$ and $G_n(\hat{x})$, which represent the correlation between the features in the spacial extend for x and $\hat{x}$, respectively [16]. The style reconstruction loss is calculated as

二值损失是否可以替换为逻辑回归?

梯度消失

把输入拆成重叠的小方块，进行判断，然后取平均值作为结果

the weighted average of squared Frobenius norm of the Gram matrices:

$$\mathcal{L}_{\text{style}} = \sum_{n=1}^{N} \lambda_{sn} \frac{1}{4d_n^2} \|G_n(\hat{x}) - G_n(x)\|_F^2 \quad (3)$$

where d_n and $\lambda_{sn} > 0$ are the spatial depth and the weight, respectively, of the extracted features from the n^{th} layer of the feature extractor network and N is the total number of layers.

The second non-adversarial loss utilized within the framework is the perceptual loss. It focuses on minimizing pixel-wise variations as well as perceptual discrepancies between the output and target images, which results in more globally consistent generated images [23]. To evaluate the perceptual loss, the mean absolute error (MAE) between the image inputs, x and $\hat{x}$, and their intermediate feature maps, extracted from the global discriminator network, is calculated. The perceptual loss is then a weighted average of the MAE:

$$\mathcal{L}_{\text{percep}} = \sum_{n=0}^{B} \lambda_{pn} \|D_n(\hat{x}, y) - D_n(x, y)\|_1 \quad (4)$$

where $\lambda_{pn} > 0$ and D_n are the weight and the extracted feature maps of the n^{th} layer of the global discriminator, respectively. B is the total number of layers for the global discriminator and D_0 represents the raw input images.

2.4. The ip-MedGAN framework

To summarize the proposed framework, ip-MedGAN incorporates a CasNet generator together with a global discriminator, which ensures the homogeneity of the inpainted region with the surrounding context information. The framework additionally utilizes a local discriminator to enhance the details of the inpainted output. The generator also minimizes the perceptual and style reconstruction losses for textural and perceptual refinement. The final loss function is given as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{adv}} + \lambda_2 \mathcal{L}_{\text{local}} + \lambda_3 \mathcal{L}_{\text{style}} + \lambda_4 \mathcal{L}_{\text{percep}} \quad (5)$$

where λ_1, λ_2, λ_3 and λ_4 represents the contributions of the different loss functions. 四个loss

3. DATASETS AND EXPERIMENTS

The proposed inpainting framework was evaluated on two different medical modalities, CT and MRI. For CT, a dataset of the brain region from 50 volunteers was collected on a clinical CT scanner (Siemens Biograph mCT). The acquired data was resampled from an original resolution of $0.85 \times 0.85 \times 5\text{mm}^3$ to $1 \times 1 \times 1\text{mm}^3$. For the training and validation datasets, two-dimensional slices were extracted and scaled to a matrix-size of 256×256 pixels, from 40 and 10 volunteers respectively. For MRI, 44 anonymized T2-weighted (FLAIR) data sets of the head region acquired on a 3T scanner were used. The MR data was also resampled to $1 \times 1 \times 1\text{mm}^3$ and rescaled to 2-D slices of matrix size 256×256 pixels. Scans from 33 patients were used for training and 11 patients for validation. Randomly placed square patches of size 64×64 were removed from the datasets to form the model's input context images y.

Table I. Quantitative comparison of inpainting techniques

Model	(a) CT inpainting			
	SSIM	PSNR(dB)	MSE	UQI
CE	0.6235	19.07	1260.2	0.9307
GLCIC	0.7137	22.18	1169.1	0.9290
MedGAN	0.8044	29.74	368.9	0.9681
ip-MedGAN	**0.8346**	**31.45**	**284.4**	**0.9737**
Model	(b) MRI inpainting			
	SSIM	PSNR(dB)	MSE	UQI
CE	0.1383	14.29	2624.7	0.8492
GLCIC	0.2287	15.01	2286.6	0.8229
MedGAN	0.3034	15.91	1809.5	0.7830
ip-MedGAN	**0.3818**	**18.32**	**1121.2**	**0.9262**

To evaluate the performance of ip-MedGAN, qualitative and quantitative comparisons with other inpainting techniques were carried out. Specifically, we compare against CE and GLCIC [9, 10]. To ensure a faithful representation of the comparison methods, verified open-source implementation were utilized along with the hyperparameters from the original publications [24, 25]. We also compare against the MedGAN image translation approach [16]. For the weighting of the different utilized loss functions, $\lambda_1 = 0.8$, $\lambda_2 = 0.2$ and $\lambda_3 = \lambda_4 = 0.0001$ was used for ip-MedGAN, with the original MedGAN framework utilizing instead $\lambda_1 = 1.0$ and $\lambda_2 = 0$. All models were trained for 200 epochs on a single NVIDIA Titan X GPU. Training time was on average 24 hours while inference time is approximately 100 milliseconds. For the quantitative comparisons, several evaluation metrics were used: the Universal Quality Index (UQI) [26], Structural Similarity Index (SSIM) [27], Peak Signal to Noise Ratio (PSNR) and the Mean Squared Error (MSE).

4. RESULTS AND DISCUSSION

The quantitative and qualitative comparisons of the inpainting performance between the proposed ip-MedGAN framework and other adversarial techniques are presented in Table I and Fig. 2 respectively. From a qualitative perspective, CEs produced inpainted regions which did not fit homogeneously into the given context information within the input images. Consequently, this method resulted in the worst quantitative scores in Table I. The GLCIC framework enhanced the inpainting performance by producing more globally consistent results. However, the inpainted regions were blurry and lacked sharpness. This may be attributed to the relatively short training time, while the original GLCIC paper recommended training for two months on a multi-GPU sys-

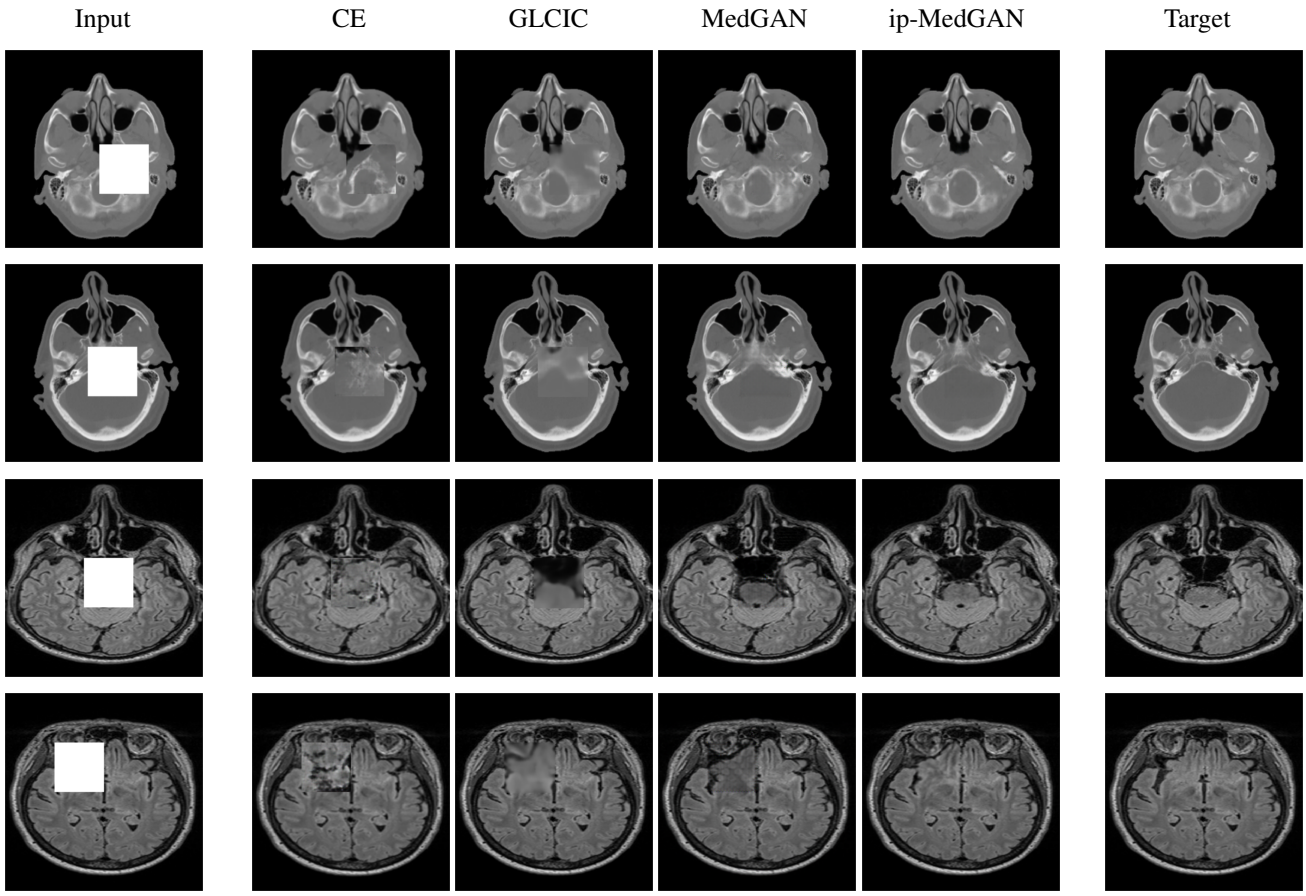

Fig. 2. Qualitative comparison of the inpainting results between the proposed ip-MedGAN framework and other adversarial inpainting techniques. The first and last two rows represent inpainting of CT and MRI modalities respectively.

tem with additional post-training image post-processing [10]. MedGAN produced noticeably enhanced results from the aspect of sharpness and global consistency with the surrounding information. However, the inpainted regions by this method lacked details and contained unrealistic tilting artifacts. By introducing an additional patch-based local discriminator, the proposed ip-MedGAN framework surpasses the limitation of MedGAN by enhancing the textural quality and details of the inpainted regions thus removing any tilting artifacts. This was also reflected quantitatively with the ip-MedGAN framework resulting in the best scores across the chosen metrics.

From another perspective, the proposed ip-MedGAN framework is not without limitation. The training procedure requires the location of the missing regions for the local discriminator. However, this is not necessary for the generator network during inference. This localization may not be readily available in the medical context without the incorporation of an additional segmentation network as a pre-processing step. Moreover, only randomly placed square regions of a fixed size were considered to the input images. However, in the medical context, distortions due to metallic implants in MRI or CT and other similar cases are of arbitrary shapes.

5. CONCLUSION

In this work, we introduce the inpainting of medical images to complete missing or distorted information. This is beneficial for further image post-processing tasks, such as PET/MRI attenuation correction and radiation therapy planning, rather than for diagnostic purposes. To achieve this goal, an adversarial framework is proposed which incorporates two patch-based discriminator networks and additional non-adversarial losses. It ensures that the inpainted results are both detailed and globally consistent in the given context information. The performance of the proposed framework was validated both qualitatively and quantitatively in comparison to other natural image inpainting techniques.

In the future, we plan to expand the proposed framework to include a segmentation network to bypass the need for manual localization of the missing regions during training. Furthermore, we plan to investigate the generalization performance of the proposed model to inpaint arbitrary shapes. Finally, verification of the performance of the inpainted results in further clinical post-processing tasks will be thoroughly investigated in comparison to other traditional approaches [2].

6. REFERENCES

[1] K.M. Koch, B.A. Hargreaves, K. Butts Pauly, W. Chen, G.E. Gold, and K.F. King, "Magnetic resonance imaging near metal implants," *Journal of Magnetic Resonance Imaging*, vol. 32, no. 4, pp. 773–787, 2010.

[2] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 882–889, 2003.

[3] Mirko Arnold, Anarta Ghosh, Stefan Ameling, and Gerard Lacey, "Automatic segmentation and inpainting of specular highlights for endoscopic imaging," *Journal on Image and Video Processing*, 2010.

[4] Mashail Alsalamah and Saad Amin, "Medical image inpainting with RBF interpolation technique," *International Journal of Advanced Computer Science and Applications*, vol. 7, 2016.

[5] Z. Feng, S. Chi, J. Yin, D. Zhao, and X. Liu, "A variational approach to medical image inpainting based on mumford-shah model," in *International Conference on Service Systems and Service Management*, 2007.

[6] Nicolas Guizard, Kunio Nakamura, Pierrick Coupé, Vladimir S. Fonov, Douglas L. Arnold, and D. Louis Collins, "Non-Local Means Inpainting of MS Lesions in Longitudinal Image Processing," *Frontiers in Aging Neuroscience*, vol. 9, 2015.

[7] Pavel Vlanek, "Fuzzy image inpainting aimed to medical imagesl," in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 2018.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio, "Generative adversarial nets," in *Conference on Neural Information Processing Systems*, 2014, pp. 2672–2680.

[9] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros, "Context encoders: Feature learning by inpainting," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544, 2016.

[10] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa, "Globally and Locally Consistent Image Completion," *ACM Transactions on Graphics (Proc. of SIGGRAPH 2017)*, vol. 36, 2017.

[11] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang, "Generative face completion," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5892–5900, 2017.

[12] Jiahui Yu, Zhe L. Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang, "Generative image inpainting with contextual attention," *Conference on Computer Vision and Pattern Recognition*, 2018.

[13] Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C.-C. Jay Kuo, "Spg-net: Segmentation prediction and guidance network for image inpainting," in *The British Machine Vision Conference*, 2018.

[14] Chao Yang, Xin Lu, Zhe L. Lin, Eli Shechtman, Oliver Wang, and Hao Li, "High-resolution image inpainting using multi-scale neural patch synthesis," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Liang Liao, Ruimin Hu, Jing Xiao, and Zhongyuan Wang, "Edge-aware context encoder for image inpainting," 2018.

[16] Karim Armanious, Chenming Yang, Marc Fischer, Thomas Küstner, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang, "MedGAN: Medical image translation using GANs," http://arxiv.org/abs/1806.06397v1, 2018, arXiv preprint.

[17] Sohil Shah, Pallabi Ghosh, Larry S. Davis, and Tom Goldstein, "Stacked U-Nets: a no-frills approach to natural image segmentation," https://arxiv.org/abs/1804.10343, 2018, arXiv preprint.

[18] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," http://arxiv.org/abs/1411.1784, 2014, arXiv preprint.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," in *Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5967–5976.

[20] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[21] Justin Johnson, Alexandre Alahi, and Fei-Fei Li, "Perceptual losses for real-time style transfer and super-resolution," 2016, pp. 694–711.

[22] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," http://arxiv.org/abs/1409.1556, 2014, arXiv preprint.

[23] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Transactions on Image Processing*, vol. 27, 2018.

[24] Narihiro Tada, "CE implementation," https://github.com/jazzsaxmafia/Inpainting.

[25] Taeksoo Kim, "GLCIC implementation," https://github.com/tadax/glcic.

[26] Zhou Wang and A. C. Bovik, "A universal image quality index," in *IEEE Signal Processing Letters*, March 2002, vol. 9, pp. 81–84.

[27] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," in *IEEE Transactions on Image Processing*, 2004, vol. 13, pp. 600–612.