bioRxiv preprint doi: https://doi.org/10.1101/447102; this version posted October 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

REFACING: RECONSTRUCTING ANONYMIZED FACIAL FEATURES USING GANS

David Abramian[a,b] *Anders Eklund*[a,b,c]

[a] Division of Medical Informatics, Department of Biomedical Engineering,
[b] Center for Medical Image Science and Visualization (CMIV),
[c] Division of Statistics and Machine Learning, Department of Computer and Information Science,
Linköping University, Linköping, Sweden

ABSTRACT

Anonymization of medical images is necessary for protecting the identity of the test subjects, and is therefore an essential step in data sharing. However, recent developments in deep learning may raise the bar on the amount of distortion that needs to be applied to guarantee anonymity. To test such possibilities, we have applied the novel CycleGAN unsupervised image-to-image translation framework on sagittal slices of T1 MR images, in order to reconstruct facial features from anonymized data. We applied the CycleGAN framework on both face-blurred and face-removed images. Our results show that face blurring may not provide adequate protection against malicious attempts at identifying the subjects, while face removal provides more robust anonymization, but is still partially reversible.

Index Terms— MRI, anonymization, GANs, image-to-image translation

1. INTRODUCTION

Anonymization is an important topic in medical imaging and data sharing, to guarantee privacy for the test subjects. This is especially important for neuroimaging [1], where head volumes are collected, and for subjects with a specific disease. Furthermore, the General Data Protection Regulation (GDPR) often requires anonymization. Virtually all data sharing initiatives in the neuroimaging field therefore remove facial features from MRI volumes before they are shared with the community. At least two techniques are currently being used: removing all facial features (e.g. using FreeSurfer [2]) or blurring the face [3]. Face removal was used in the 1000 Functional Connectomes Project [4], while face blurring was used in the Human Connectome Project [5].

This study was supported by Swedish research council grant 2017-04889. Funding was also provided by the Center for Industrial Information Technology (CENIIT) at Linköping University, and the Knut and Alice Wallenberg foundation project "Seeing organ function". We thank the Biomedical Image Analysis Group at Imperial College, London, for for sharing the IXI dataset. The Nvidia Corporation, which donated the Nvidia Titan X Pascal graphics card used to train the GANs, is also acknowledged.

Deep learning has been extensively used for medical imaging [6, 7], as these new methods in many cases provide superior performance compared to traditional image processing algorithms. In particular, generative adversarial networks (GANs) [8, 9, 10] have recently become a very popular tool for a multitude of tasks, such as realistic image synthesis, denoising, domain translation, and superresolution [11, 12]. A conditional GAN can for example be used to generate CT images from MRI [13, 14], PET images from MRI [15] or T1-weighted MR images from T2-weighted images [16, 17].

New machine learning techniques, in combination with fast computing, have provided great benefits for the medical imaging field. However, these techniques have also opened the door to certain malicious applications. In this work, we attempt to highlight this problem by demonstrating that a GAN can be used to restore facial features of anonymized T1-weighted images.

2. DATA

The data used was obtained from the IXI dataset [18], a multi-site MRI dataset including T1, T2, PD, MRA and diffusion data from 581 subjects. In this work we employ only the T1 images, which are provided without any facial anonymization. The images have also not been coregistered or normalized to a common space. Table 1 provides more details about the composition of the IXI dataset and in particular the T1 images.

3. METHODS

3.1. Anonymization

Two different anonymization procedures were applied to the data. The first was the mask_face software [3], which applies blurring to the facial surface while conserving the structure beneath the face. The second is the mri_deface function from the FreeSurfer package [2], which zeroes out all the voxels from the subject's face, including deeper facial structures.

bioRxiv preprint doi: https://doi.org/10.1101/447102; this version posted October 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Location	Scanner	Num. subjects	Image dim. (vox.)	Vox. size (mm)
Guy's Hospital	Philips Gyroscan Intera 1.5T	322	$150 \times 256 \times 256$	$1.2 \times 0.938 \times 0.938$
Hammersmith Hospital	Philips Intera 3T	185	$150 \times 256 \times 256$	$1.2 \times 0.938 \times 0.938$
Institute of Psychiatry	GE 1.5T	74	$146 \times 256 \times 256$	$1.2 \times 0.938 \times 0.938$

Table 1. Composition of the IXI dataset.

3.2. GAN model

We employed the CycleGAN unsupervised image-to-image translation framework [10] to reconstruct facial features from anonymized data. CycleGAN is a generative adversarial network which employs two generators and two discriminators, all of which are convolutional neural networks, to simultaneously learn the mappings between the two domains A and B. Because the data is unpaired, the problem of finding a mapping between two domains is underdetermined. To counteract this, CycleGAN employs a cycle consistency constraint that requires that data converted to another domain and back be as close to the original as possible.

We used an implementation of 2D CycleGAN previously developed in our group [17], based on the Keras API [19]. The model is trained to transform images between the anonymized and the original domains. The generators in the GAN have 24 convolutional layers, while the discriminators have 5.

3.3. Training

The model was trained using individual head slices. To generate the training data, 21 sagittal slices were extracted from each subject. This was done for the original dataset as well as for both anonymized versions. Each slice was normalized with the maximum value of its corresponding original volume. All the images were of size 256×256 pixels.

Two different sets of images were used for training the CycleGAN: the first included subjects scanned at Guy's Hospital, and the second included subjects scanned at all three locations. In the first case, 6300 images (300 subjects) were used for training and 462 (22 subjects) for testing, while in the second case 10500 images (500 subjects in total; 284, 151, and 65 from each site respectively) were used for training and 1701 (81 subjects in total; 38, 34, and 9 from each site respectively) were used for testing.

In both cases the training was performed for 200 epochs, with linear decay of the learning rate applied during the second half of the training. The training time for the Guy's data was about 2 days, while for the whole dataset it was close to 4 days on an Nvidia Titan X Pascal graphics card.

3.4. Evaluation

Qualitative evaluation of the results was performed by visual comparison of the original and reconstructed images. Quantitative results are provided in the form of correlation coefficients and structural similarity indices (SSIMs) between the original and the reconstructed test images. The former metric represents the global correlation between two images, while the latter is aimed at predicting the perceived quality of a target image when compared to a reference image. It is important to highlight the difficulty in quantitatively evaluating the perceived realism of an image, since existing metrics may not closely track the nuances of the human visual and pattern recognition systems.

4. RESULTS

4.1. Face blurring

Our approach managed to convincingly reconstruct the facial features for the face-blurred images (see Figure 1). The results for the single dataset experiment were particularly consistent. Figure 3 shows that high degrees of correlation and structural similarity were attained between original and reconstructed images.

Results were also positive for the full experiment, but showing a strong dependence of reconstruction quality on acquisition site. Qualitatively the images from Guy's Hospital attained the best results, followed by those of Hammersmith Hospital, and finally those of the Institute of Psychiatry. From the quantitative results in Figure 4 it is clear that, while the correlation results are approximately constant across all images, the SSIM is capable of capturing the difference in perceived quality in the results from different sites.

4.2. Face removal

Only limited success was achieved for the face-removed images (see Figure 2). While the GAN managed to restore a credible face in many cases, this rarely resembled the original face. The results also suffered from mode collapse, with particular image patches occurring in the output for many different input images. Another common problem is a sharp vertical cutoff on the front of the face, especially on the nose. This is due to the heads of many of the subjects used for training being cut against the boundaries of the volume

The reconstructed faces for the single dataset experiment are generally more shallow compared to the original data. Outcomes were slightly better for the full experiment, with a similar pattern of dependence on the acquisition site as seen in the face blurring case. Figures 3 and 4 show consistently

bioRxiv preprint doi: https://doi.org/10.1101/447102; this version posted October 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

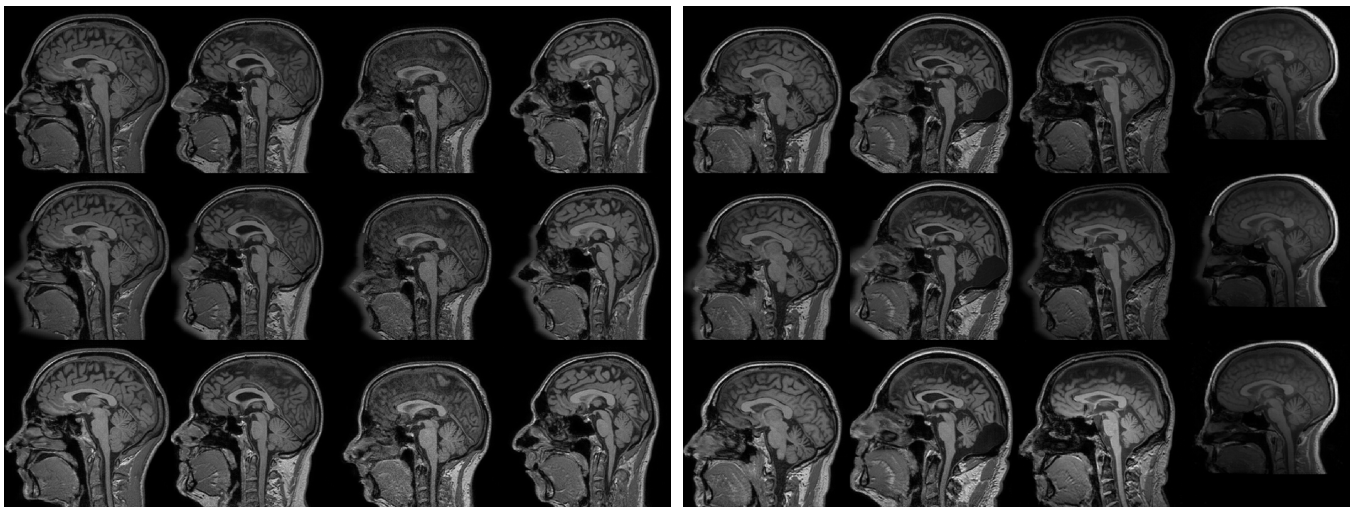


Fig. 1. Left: results for training using only subjects from Guy's hospital, Right: results for training using data from all 3 sites. Top row: original image, middle row: face-blurred image, bottom row: reconstructed image. CycleGAN learns to perform a deconvolution, to reconstruct the anonymized face.

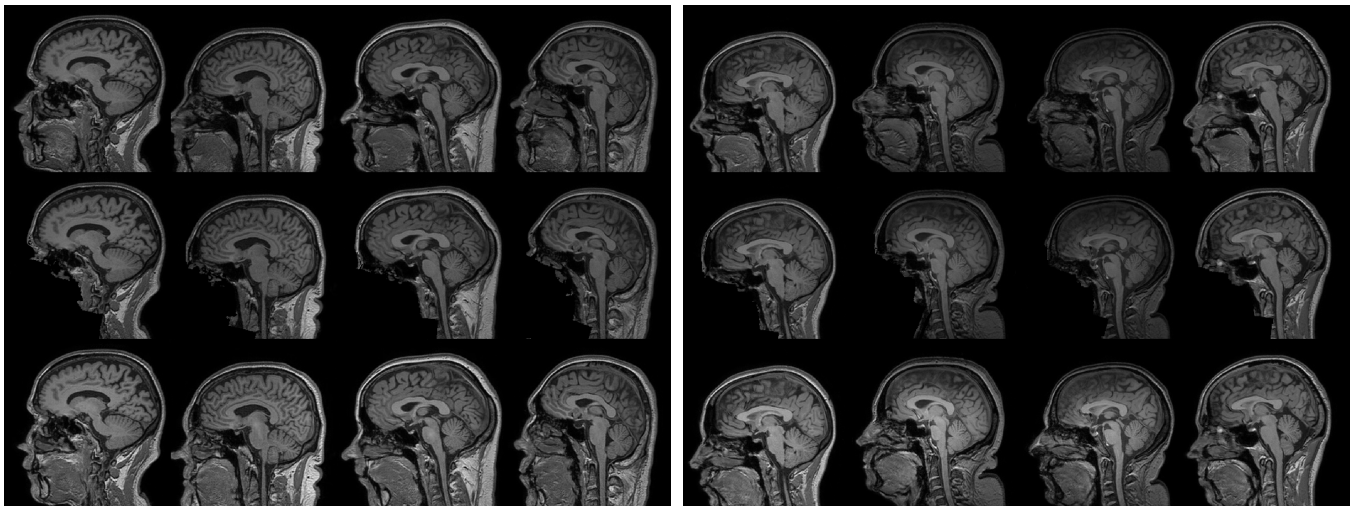
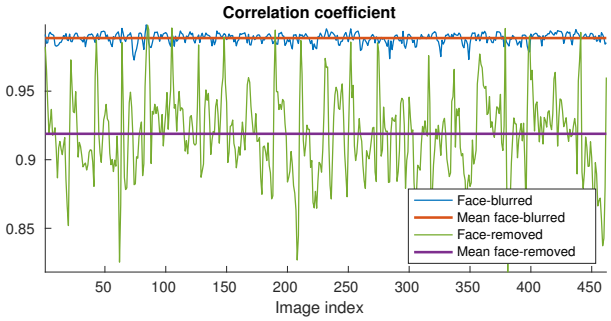


Fig. 2. Left: results for training using only subjects from Guy's hospital, Right: results for training using data from all 3 sites. Top row: original image, middle row: face-removed image, bottom row: reconstructed image. CycleGAN learns to add a face, but in many cases it is not the correct face.

bioRxiv preprint doi: https://doi.org/10.1101/447102; this version posted October 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



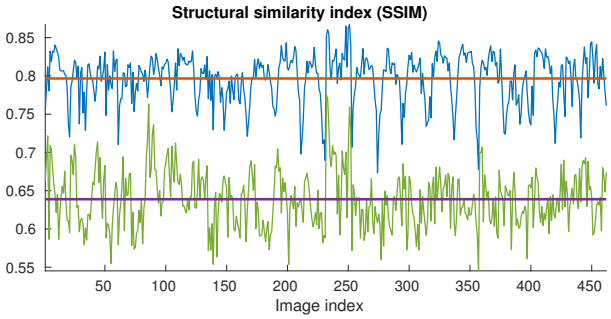
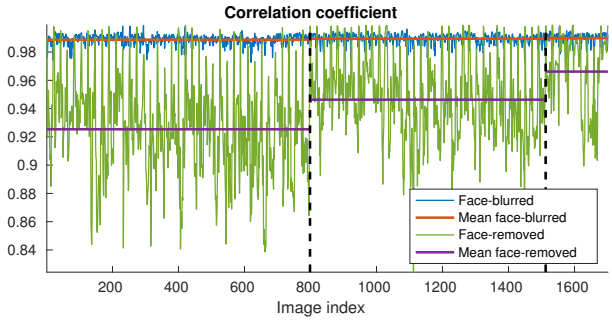


Fig. 3. Correlation and structural similarity between original and reconstructed test images after training using only subjects from Guy's hospital. As expected, it is easier to reconstruct face blurred images, compared to face removed images.



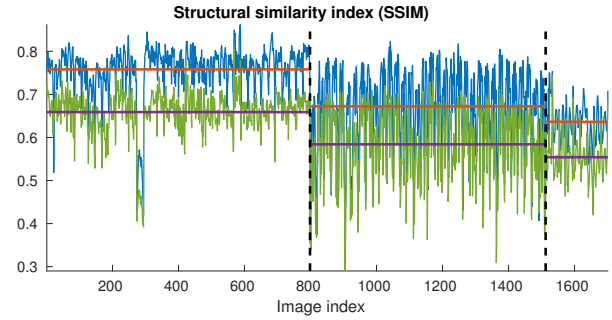


Fig. 4. Correlation and structural similarity between original and reconstructed test images after training using data from all 3 sites. Vertical lines separate images by aquisition site (left to right: Guy's Hospital, Hammersmith Hospital, Institute of Psychiatry). Means were calculated separately for images from each site.

worse quantitative results for the face-removed images compared to the face-blurred ones. The changes in correlation coefficient across acquisition sites happen in opposite direction to the perceived changes in quality, showing again the superiority of SSIM as a measure of perceived quality.

5. DISCUSSION

The successful reconstructions raise some concerns about the potential vulnerability of certain common anonymization methods used for MRI data. While the current implementation has only been shown to work on test data coming from the same dataset used to train the model, pre-trained networks are commonly used outside of their original domain. As an example, the Human Connectome Project [5] provides face-blurred data of 1113 subjects, which could have been used to test the generalization properties of our procedure. However, to respect the privacy of these subjects, we have not attempted to apply our trained GANs on them. In addition, the generalization properties of the network could be improved by training it on datasets acquired from multiple sites and with different scanning parameters.

Restoring facial features from blurred data requires that CycleGAN learns a deconvolution. Given the amount of information remaining in the image, such as some of the facial bones, and even the trajectory traced by the blurred face, this problem did not prove to be exceedingly difficult. Restoring the complete face from zeroed out data poses a much more challenging inpainting problem. Even then, some success was achieved using an established GAN approach. In other domains, great success has been achieved by using dedicated inpainting architectures [20], which can be another way to perform refacing.

In regards to the model employed, the choice to use a 2D GAN was made on the basis of constraints in available memory and processing time. When applied to every slice of a volume, such an approach would show discontinuities between contiguous slices. For this reason, no attempt was made to identify the subjects using volume renderings. Future work might examine the possibility of using a 3D GAN, which we expect would yield better results and solve the discontinuity problem. Another possible avenue for future investigation would be the use of a supervised learning model such as Pix2Pix [9], since the available data is paired.

A potential legitimate application for refacing can be the improvement of morphometric estimates from anonymized data. It has been shown that even minimal anonymization procedures such as facial blurring can have an impact on morphometric estimates, such as subcortical volume and cortical thickness [21]. Recovering the face of each subject could improve the correctness of these estimates.

bioRxiv preprint doi: https://doi.org/10.1101/447102; this version posted October 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

6. REFERENCES

[1] R. Poldrack and K. Gorgolewski, "Making big data open: data sharing in neuroimaging," *Nature Neuroscience*, vol. 17, pp. 1510–1517, 2014.

[2] Amanda Bischoff-Grethe, I Burak Ozyurt, Evelina Busa, Brian T Quinn, Christine Fennema-Notestine, Camellia P Clark, Shaunna Morris, Mark W Bondi, Terry L Jernigan, et al., "A technique for the deidentification of structural brain MR images," *Human brain mapping*, vol. 28, no. 9, pp. 892–903, 2007.

[3] Mikhail Milchenko and Daniel Marcus, "Obscuring surface anatomy in volumetric imaging data," *Neuroinformatics*, vol. 11, no. 1, pp. 65–75, 2013.

[4] Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, et al., "Toward discovery science of human brain function," *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4734–4739, 2010.

[5] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al., "The WU-Minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[6] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram Van Ginneken, and Clara I Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[7] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers, "Deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2017.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *CoRR*, vol. abs/1703.10593, 2017.

[11] Xin Yi, Ekta Walia, and Paul Babyn, "Generative adversarial network in medical imaging: A review," *arXiv preprint arXiv:1809.07294*, 2018.

[12] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay, "GANs for medical image analysis," *arXiv preprint arXiv:1809.06222*, 2018.

[13] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen, "Medical image synthesis with context-aware generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 417–425.

[14] Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum, "Deep MR to CT synthesis using unpaired data," in *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2017, pp. 14–23.

[15] Wen Wei, Emilie Poirion, Benedetta Bodini, Stanley Durrleman, Nicholas Ayache, Bruno Stankoff, and Olivier Colliot, "Learning myelin content in multiple sclerosis from multimodal MRI through adversarial training," *arXiv preprint arXiv:1804.08039*, 2018.

[16] Salman Ul Hassan Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Çukur, "Image synthesis in multi-contrast MRI with conditional generative adversarial networks," *arXiv preprint arXiv:1802.01221*, 2018.

[17] Per Welander, Simon Karlsson, and Anders Eklund, "Generative adversarial networks for image-to-image translation on multi-contrast MR images - A comparison of CycleGAN and UNIT," *arXiv preprint arXiv:1806.07777*, 2018.

[18] "IXI Dataset," `http://brain-development.org/ixi-dataset/`.

[19] François Chollet et al., "Keras," `https://keras.io`, 2015.

[20] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 107, 2017.

[21] Avram J Holmes, Marisa O Hollinshead, Timothy M O'Keefe, Victor I Petrov, Gabriele R Fariello, Lawrence L Wald, Bruce Fischl, Bruce R Rosen, Ross W Mair, et al., "Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures," *Scientific data*, vol. 2, pp. 150031, 2015.