Performance evaluation of a Pix2Pix synthetic Xray image generator for preclinical applications

Vasilis Eleftheriadis[1], Paneta Valentina[1], Eleftherios Fysikopoulos[1] and Panagiotis Papadimitroulas[1]

[1]BIOEMISSION TECHNOLOGY SOLUTIONS, BIOEMTECH, Athens, Greece

Abstract In this work, we evaluate a pix2pix conditional generative adversarial network as a potential solution for generating adequately accurate synthesized morphological X-ray images by translating standard photographic images of mice. Continuing our previous work, we expanded our dataset to 940 paired photographic/X-ray mice images and used data augmentation and transfer learning techniques to improve the performance and generalization of our model. We explore how the expansion of the training dataset affects the models' performance and use five image quality metrics to quantitatively evaluate the models' performance in the present dataset and compare the results to visual inspection by experts to assess the metrics' ability to evaluate image-to-image translation tasks. In conclusion, a combination of metrics is essential for generated image quality evaluation, and we propose an ensemble of the used metrics as an essential methodological step to evaluate the performance of image-to-image deep learning models.

1 Introduction

Scintigraphy is a high throughput alternative to 3D high-end molecular imaging systems and biodistribution ex vivo studies in monitoring the interaction of radioisotope- or fluorescence-labeled pharmaceuticals on animal subjects. It can effectively be used to track a new tracer and study the accumulation from zero point in time post-injection.

Image-to-image translation is a generative deep learning (DL) technique that utilizes generative adversarial networks (GANs), conditional generative adversarial networks (cGANs), and convolutional neural networks (CNNs) to learn complex mapping functions between input and output images to translate a source image into a target image while preserving certain visual properties of the original.

In our previous studies [1, 2] we presented a potential solution for generating adequately accurate synthetic morphological X-ray images by translating standard photographic images of mice using a well-known cGAN for image-to-image translation (pix2pix) [3]. Such an approach would benefit both imaging and targeted radionuclide therapy studies. The results showed that the network predicts an X-ray image with sufficient accuracy and that the calculated metrics are comparable with those presented in literature. In this study, we expanded our dataset, and we used data augmentation and transfer learning techniques to improve the generalization of our method. In addition, we performed a thorough evaluation of the performance comparing different metrics.

2 Materials and Methods

A dataset of 940 input/ground truth image pairs consisting of a photographic image and the corresponding X-ray scan of the same anesthetized mouse has been acquired. The photographic images were captured using commercial imaging devices by BIOEMTECH (eyes-series – www.bioemtech.com).

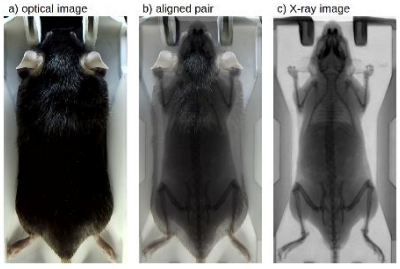

Figure 1: a) Indicative optical image acquired by BIOEMTECH eyes-series imaging devices, b) Corresponding X-ray image, c) Aligned pair.

These scanners, featuring a standard photographic sensor, provide an optical image of the animal (Fig. 1a) [4]. Both systems are adequate for mice imaging providing a useful field of view (FoV) of 50×100 mm^2.

The study involved 86 white and 88 black Swiss albino mice. We acquired 5 input/ground truth images of each animal in different poses upon the hosting bed, leading to a total of 940 input/ground truth images. The mice were divided into two groups to form the training and testing datasets. A group of 76 white and 78 black mice was used to collect 823 paired images for training, while a group of 10 white and 10 black mice was used to collect 117 paired images for testing. Except mouse color, the method was evaluated against different animal hosting beds (plastic bed with white color; plastic bed with black color), which leads to different backgrounds in the input image. The image pairs were properly aligned before training, having a resolution of 512×1024 pixels, corresponding to the 50×100 mm^2 FoV (Fig. 1b). Additionally, to ensure the validity of the training and testing sets, the mice used for training were separate from the individual mice used in the test set.

In this study, we utilized the PyTorch implementation of the pix2pix algorithm [3] and used the cross-entropy loss function. We trained all models for 200 epochs. For our transfer learning experiments, we first trained models on a subset of the dataset and then finetuned those on data from specific scanners using the weights of the pretrained models. For data augmentation we used the open-source Python library Albumentations [5]. By using different geometric and photometric transformations we managed to expand our training dataset up to 5175 paired images. We evaluated the performance of the pix2pix models in the present dataset using five commonly used image quality metrics: (a) mean squared error (MSE), (b) normalized root mean squared error (NRMSE), (c) peak signal-to-noise ratio (PSNR), (d) structural similarity index measure (SSIM) and (e) Fréchet inception distance (FID). Furthermore, we propose and evaluate our models' performance with a novel

weighted average ensemble performance evaluation metric (WAEM) which combines these five metrics.

$$WAEM = \frac{a}{MSE} + \frac{b}{NRMSE} + c \cdot PSNR + d \cdot SSIM + \frac{e}{FID}$$

3 Results

Model performance for increasing dataset size, starting from 50 to 5175 images, is presented in Fig. 2 through the 5 metrics used in this study. By visual inspection of the generated images, none of these metrics was adequate to individually be used to identify the best performing model. The weighted combination of the used metrics (WAEM), relying on a greater extent on FID, was found to be the metric which most consistently agrees with visual inspection and is also included in Fig. 2.

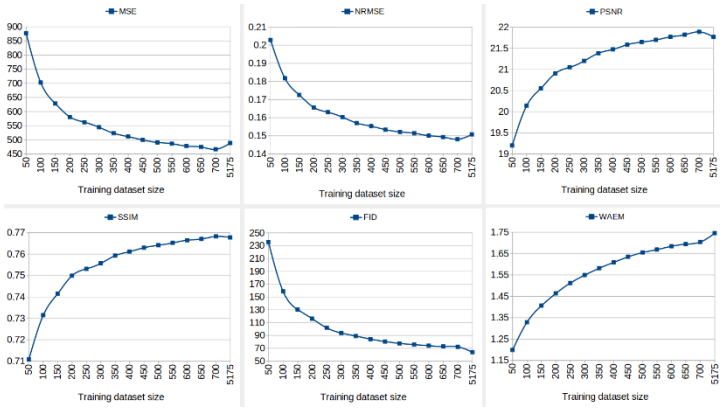

Figure 2: Graphs of the 6 metrics used in this study for the performance evaluation of the pix2pix models. The X-axis represents the size of the training dataset.

Fig. 3 shows indicative optical to X-ray translations of the same optical image in the test dataset for different training set sizes for visual comparison. The increase of the dataset using data augmentation (Fig. 2 & 3) was found to boost performance, while transfer learning techniques did not improve the generalization of our model, in our case study.

4 Discussion

Increasing the training dataset size with real input/ground truth pairs of images (up to 700 in the graph) the performance of the pix2pix models is improved according to all evaluation metrics. Nevertheless, when we further increase the training dataset size using data augmentation to 5175 images, four out of five metrics (MSE, NRMSE, PSNR and SSIM) show a drop in the model's performance. In contrast, FID metric gets even lower values implying that the final model with the augmented training dataset performs better. Fig. 3 depicts the visual resemblance of the generated images to the ground truth that drastically improves by increasing the training dataset size either with real images and/or using data augmentation. We identified FID as the most indicative metric for image-to-image translation. However, we observed that when 2 models have similar FID scores the models that translate the photographic images into synthetic x-rays that resemble the ground truth better are the ones that score better on the rest of the performance evaluation metrics (MSE, NRMSE, PSNR and SSIM), independently of which model had the best FID score. Our proposed weighted metric (WAEM) was found to clarify these cases and gave a straightforward measure of performance.

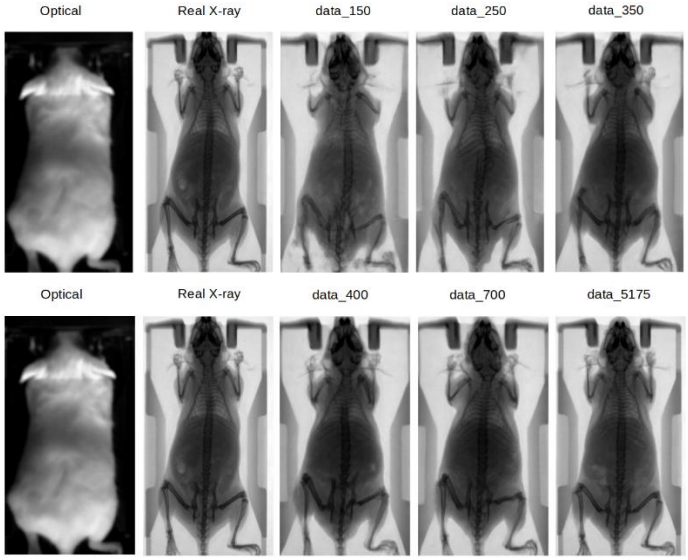

Figure 3: Indicative optical to X-ray translations of the same optical image in the test dataset for different training set sizes.

5 Conclusion

In the present work we evaluate the performance of pix2pix for image-to-image translation in our case study, in terms of dataset size, data augmentation, transfer learning techniques and the performance metrics themselves, that proved to be the most challenging issue. Our proposed weighted average ensemble performance evaluation metric (WAEM), which combines the five metrics used in this study (and are the most frequently used metrics in this field) could help researchers evaluate the performance of image-to-image DL models according to their case/dataset too.

References

[1] E. Fysikopoulos et al. "Optical to Planar X-ray Mouse Image Mapping in Preclinical Nuclear Medicine Using Conditional Adversarial Networks". J Imaging 7(12):262 (2021). DOI: 10.3390/jimaging7120262.

[2] E. Fysikopoulos et al. "Photograph to X-ray Image Translation for Anatomical Mouse Mapping in Preclinical Nuclear Molecular Imaging". In: Su, R., Zhang, YD., Liu, H. (eds) Proceedings of MICAD 2021. Lecture Notes in Electrical Engineering, vol 784. Springer, Singapore. DOI: 10.1007/978-981-16-3880-0_31.

[3] P. Isola et al. "Image-to-image translation with conditional adversarial networks". Proceedings of the IEEE CVPR Proceedings; Honolulu, HI, USA. 21–26 July 2017. DOI: 10.1109/CVPR.2017.632.

[4] M. Rouchota et al. "A prototype PET/SPET/X-rays scanner dedicated for whole body small animal studies". Hell. J. Nucl. Med. 2017;20:146–153. DOI: 10.3390/jimaging7120262.

[5] A. Buslaev et al. "Albumentations: Fast and Flexible Image Augmentations". Information 2020, 11, 125. DOI: 10.3390/info11020125.