Biomedical Signal Processing and Control 72 (2022) 103283

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

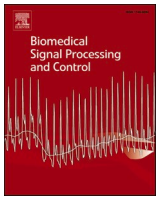

Image translation for medical image generation: Ischemic stroke lesion segmentation

Moritz Platscher [a,*,1], Jonathan Zopes [a], Christian Federau [a]

[a] Institute for Biomedical Engineering, University and ETH Zürich, Gloriastrasse 35, 8092 Zürich, Switzerland

ARTICLE INFO

Keywords:
generative models
image-to-image translation
stroke lesion segmentation
image synthesis

ABSTRACT

Deep learning based disease detection and segmentation algorithms promise to improve many clinical processes. However, such algorithms require vast amounts of annotated training data, which are typically not available in the medical context due to data privacy, legal obstructions, and non-uniform data acquisition protocols. Synthetic databases with annotated pathologies could provide the required amounts of training data. We demonstrate with the example of ischemic stroke that an improvement in lesion segmentation is feasible using deep learning based augmentation. To this end, we train different image-to-image translation models to synthesize magnetic resonance images of brain volumes with and without stroke lesions from semantic segmentation maps. In addition, we train a generative adversarial network to generate synthetic lesion masks. Subsequently, we combine these two components to build a large database of synthetic stroke images. The performance of the various models is evaluated using a U-Net which is trained to segment stroke lesions on a clinical test set. We report a Dice score of **72.8**% [$\mathbf{70.8 \pm 1.0}$%] for the model with the best performance, which outperforms the model trained on the clinical images alone **67.3**% [$\mathbf{63.2 \pm 1.9}$%], and is close to the human inter-reader Dice score of **76.9**%. Moreover, we show that for a small database of only 10 or 50 clinical cases, synthetic data augmentation yields significant improvement compared to a setting where no synthetic data is used. To the best of our knowledge, this presents the first comparative analysis of synthetic data augmentation based on image-to-image translation, and first application to ischemic stroke.

1. Introduction

Ischemic stroke (IS) is the second leading cause of death worldwide [1], and, if not fatal, frequently results in irreversible brain tissue damage and disabilities. IS is caused by occlusion of blood vessels, resulting in restricted or suppressed blood supply to some areas of the brain. Consequently, the success of therapy is closely tied to the time between the onset of symptoms and successful revascularization treatment [2,3]. Therefore, it is of utmost importance to streamline and optimize the diagnostic process. While many clinical centers use computed tomography (CT) as a first imaging modality for stroke due to its wider availability, speed, and relative absence of contraindications, the gold standard for the identification and quantification of IS lesions is diffusion-weighted magnetic resonance imaging (DWI), on which the infarcted region induces a hyper-intensity compared to normal tissue, corresponding to zones of relative diffusion restriction. Manual segmentation of these areas is a relatively simple, but tedious task for a trained expert, while automatic segmentation using standard methods is a difficult task, e.g., due to the complex lesion geometries, non-trivial variability in magnetic resonance signal intensities, varying locations, and numerous image artifacts. Machine learning, and deep learning (DL) in particular, holds the potential to robustly automatize such tasks [4–6]. However, the training of DL models requires vast amounts of labeled data, which are typically not available due to privacy concerns, incompatible data formats, or because the disease is rare. In addition, manual segmentation of medical images is a task that only experienced radiologists can perform, is expensive, error-prone, and time-consuming. Consequently, large databases of annotated medical

* Corresponding author.
E-mail address: platscher@biomed.ee.ethz.ch (M. Platscher).

[1] Submitted 13 Apr 2021. MP is supported by the University of Zurich under the Grant no. FK-20-052. JZ acknowledges support from the SPARK Grant CRSK-3_190697 of the Swiss National Science Foundation. CF is supported by an Ambizione Fellowship grant (PZ00P3_173952) of the Swiss National Science Foundation. Calculations for this project were performed in part using support from Google and on a Titan Xp donated by the NVIDIA Corporation Code available at www.github.com/MoPl90/image_translation.

https://doi.org/10.1016/j.bspc.2021.103283
Received 13 April 2021; Received in revised form 20 September 2021; Accepted 23 October 2021
Available online 2 November 2021
1746-8094/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

images are scarce: For example, the Ischemic Stroke Lesion Segmentation Challenge (ISLES) data set contains 64 IS volumes [7]. In comparison, the ImageNet database [8], a popular data set for object recognition in the non-medical area, contains 14 million images.

Synthetic image generation with deep generative models [9] could bridge the success of DL to medical applications by augmenting available data sets, or replacing them altogether, circumventing the aforementioned concerns related to data security and availability. However, these models, too, require large training data sets in order to obtain good results. Moreover, alongside realistic images, the model must provide lesion labels, which is not straightforwardly implemented in the current deep generative pipelines.

We attempt to synthesize DWIs of stroke lesions with image-to-image translation models (ITMs), i.e. DL models that transform images from one domain of training data to another domain of training data: semantic segmentation maps containing anatomical labels of the brain structure and the label of the stroke lesion are translated into brain DWIs. In addition, new lesion labels are generated by a 3D generative adversarial network. We envision that the anatomical labels will guide the network to produce sharp images with tissue contrast comparable to the real DWIs. At the same time, the lesion labels of the training set will provide sufficient information to learn a pathology-specific contrast modification. The ITM should learn to generate hyper-intensities inside the lesion labels, while ignoring any (statistically underrepresented) misclassification, such as hyper-intensities outside lesion labels or normal tissue misclassified as belonging to a lesion by the human reader. This has a welcome side effect: the network will produce DWIs with lesions that match the input region labeled as such with high confidence, thus providing accurate lesion labels. The synthetic images obtained using such an approach can then be used to augment a clinical database, or, such as here, directly for supervised learning to segment stroke lesions from normal brain tissue.

Related work. The proposed approach is similar in spirit, but more general than Ref. [10], where synthetic IS volumes were generated by fusing real IS lesion contours into healthy brain DWIs by linearly increasing the voxel intensity within the contour to mimic a lesion. This ensures that anatomical boundary conditions are respected and realistic intensity patterns are generated. Part of our proposal is to outsource this task to a generative DL model, which generalizes the allowed transformations given that neural networks are intrinsically non-linear. With their approach the authors achieved an increase in the Dice similarity coefficient (DSC) from 65% using the clinical data alone to 70% when including 2,000 synthetic images, and even up to 72% using 40,000 synthetic DWIs. An advantage of this approach is that it generates coherent 3D data; however, it is limited combinatorically since there is only a limited number of real lesion labels available. Moreover, Refs. [11–13] successfully used generative adversarial networks (GANs) to generate synthetic magnetic resonance images (MRIs) to improve brain tumor segmentation. A different approach was used by [14]: there the authors use an ITM to obtain DWIs from CT images allowing them to improve the segmentation quality of the ischemic stroke core tissue on CTs augmented with synthetic DWIs. Similarly, Ref. [15] used CT-to-MR image-translation to improve lung tumor segmentation. Further approaches using style transfer were recently proposed in Ref. [16], where image-to-image translation was used to translate healthy into diseased MR images for brain tumor segmentation, and in Ref. [17] transforming MR into CT images for general treatment planning. Finally, Ref. [18] introduced an ITM approach that is was used for a variety of medical tasks involving image denoising, motion-correction, and domain adaptation.

Contribution. We provide a new approach to synthesizing both IS lesion labels and DWIs from semantic segmentation maps. While the former task is realized by ITMs, the lesion labels are generated by a 3D GAN. This allows us to build spatially coherent brain volumes including lesion labels. Moreover, this makes our pipeline scalable in the sense that the GAN can be used to produce an unlimited number of synthetic, yet distinct, lesion labels. We thereby demonstrate that DL-based data augmentation is apt to leverage the information contained in limited medical data sets and outperform conventional data augmentation techniques, an idea that was coined by [19] and further developed for segmentation in [20] in a non-medical context. We quantify this statement by comparing the DSC of a U-Net [21] trained on manually labeled data versus U-Nets trained with the synthetic data, each evaluated on a test set of clinical IS data. We highlight several areas of application for our pipeline and emphasize the utility of the approach, given that the trained ITMs can be freely distributed without any data privacy concerns. A number of studies exist that investigate the potential improvement of downstream clinical tasks, e.g., anatomical segmentation and tumor segmentation; however many discuss only one specific model setup. To the best of our knowledge, this is the first comparative study using ITMs for data augmentation in IS lesion segmentation.

This paper is structured as follows. In Section 2 we describe the data and models that are used in the image synthesis pipeline and detail how to evaluate the results quantitatively. The results of the analysis are presented and discussed in Section 3, before concluding in Section 4. The Appendix contains supplemental information.

2. Materials and methods

We propose an approach for generating annotated DWIs of brains displaying an IS lesion by means of two consecutive generative models, one for generating realistic stroke lesion labels and another to translate brain segmentation masks into DWIs. Anatomical segmentation masks of healthy brains are readily available and require no medical training, since several automatic segmentation algorithms exist. The synthesis of stroke lesion labels by a GAN, on the other hand, is assumed to introduce general transformations, as well as realistic interpolations of latent space representations of available lesion masks. Finally, the use of conditional generative models to generate synthetic images, significantly improves the convergence behavior of the generators compared to unconditional, latent space models. Thereby we construct a minimal toolbox, that can easily be scaled by employing the interpolation capacities of the GAN to synthesize more lesion masks to generate more and more annotated DWIs. In the following we provide a detailed description of the involved data and methods.

2.1. Data

A database of 804 DWIs of patients that presented with symptoms of IS was obtained at the University Hospital Basel, for which Institutional Review Board approval was granted. This database (henceforth stroke DB) contains 449 DWI positive cases, i.e. diagnosed IS lesions (mean patient age 72 ± 14 years; 200 left-sided ISs, 193 right-sided ISs, 56 bilateral ISs; 194 female and 255 male). The DWI positive cases were separated randomly for training (365) and testing (74). Additionally, 85 DWI negative samples were included in the test set, which amounts to 159 test samples in total. A separate database of 2027 healthy DWI scans (mean age 38 ± 24 years; 1088 female, 939 male) was used for the augmentation pipeline (normal DB). The majority of images in the stroke DB were acquired on a 1.5 T scanner (67% @ 1.5 T; 23% @ 3 T), while most of the normal DB scans used a 3 T scanner (68% @ 1.5 T; 22% @ 3 T). On average, stroke DB scans were acquired with similar echo and recovery times ($T_E = 90 \pm 16\,\text{ms}, T_R = 7400 \pm 1300\,\text{ms}$ [normal DB] vs. $T_E = 100 \pm 2\,\text{ms}, T_R = 7000 \pm 1500\,\text{ms}$ [stroke DB]).

Images were co-registered to the standard Montreal Neurological Institute atlas [22], skull-stripped using FSL's 'brain extraction tool' [23], and re-sampled to a standard resolution of $128 \times 128 \times 40$ voxels using ANTs [24]. For better stability during training, the top and bottom four slices were cropped. The voxel intensities were clipped at the 99.5$^{\text{th}}$ percentile, while the background was clipped at an absolute voxel intensity of 35. Finally, signal intensities were re-scaled to the range $[-1,$

1]. The anatomical segmentation of brains was obtained using 3D U-Net trained on reference images obtained using FreeSurfer [25], as outlined in Ref. [26]. This reduced the processing time for the entire database to a few hours.

2.2. Image translation

While unconditional GANs have been trained to yield state-of-the art results on brain DWIs [27], they do not automatically provide ground truth labels for anatomical structures or pathologies. ITMs, on the other hand, are generative models, which generate samples conditional to an input segmentation map. Thus, a ground truth label is available by construction. Moreover, this approach helps to improve the quality of the generated image by providing boundaries between different instances of a segmentation label. In this section we introduce the ITMs studied in this manuscript, further details can be found in the Appendix.

2.2.1. Pix2Pix

Pix2Pix is a well-studied ITM, which was first proposed by [28] and further developed by [29]. Pix2Pix is widely accepted as the method of choice for paired image translation, i.e. when the images in two domains come in pairs as it is the case for our database. It has previously been applied successfully to medical data, see e.g. Refs. [14,13,30–33]. The original Pix2Pix is based on the U-Net architecture; however, in the high-resolution derivative of the model residual blocks were used [29,34]. We have trained both architectures but found little difference in quality and therefore used the U-Net based version which converges faster. The loss function for this network is a weighted sum of adversarial loss and L_1-norm reconstruction loss,

$$\mathscr{L}_{\text{Pix2Pix}} = \mathbb{E}_y[\log D(y)] + \mathbb{E}_x[1 - \log D(G(x))] + \lambda \|y - G(x)\|_{L_1}. \quad (1)$$

In this equation, $\mathbb{E}_{x/y}$ denotes the expectation value taken w.r.t. a batch x or y from one of the two domains, D represents the discriminator network which operates on the target domain and G is the generator network. We consider different values for the reconstruction loss weight λ, since we have found that a value $\lambda = 100$ (Pix2Pix100) yields qualitatively more appealing results than the recommended value $\lambda = 10$ (Pix2Pix10). For the discriminator architecture we rely on the PatchGAN [28,35,36]. In contrast to ordinary discriminators, PatchGAN does not output a single number to characterize if an image is real or fake, but does so for patches of the image, thus allowing a more refined feedback for the generator on smaller scales. The large scale, or low-frequency, features are sufficiently well captured by the L_1 loss [28].

2.2.2. cycleGAN

A method of unpaired image translation is given by cycleGAN [37]. Two generator-discriminator pairs, one for each image domain, allow the model to be trained via a reconstruction loss, translating from domain A to B using the first generator, then back from B to A using the other generator and computing the pixel-wise (L_1 or L_2 norm) difference between the resulting and the original image. This cycle consistency lends its name to the model and tackles the problem of unpaired data. However, the number of parameters to train is quite vast, reaching 42 million in our setup. In order to handle this large number of free parameters, the model enforces an identity loss, which penalizes deviations of the generators from an identity mapping.[2] Note that cycleGAN has been applied successfully to paired data in medical image generation [38,39].

2.2.3. SPADE

SPADE is the most recent of the ITMs considered in this work. While it has not been widely applied in the medical field, results on ambiental images are promising [40]. Furthermore, the SPADE architecture, which is largely identical to a standard GAN, is very economical, reducing the number of parameters to train compared to other ITMs. In analogy to GANs [9], this model draws a random latent vector which is subsequently upsampled introducing elements of variation in the generative process. The segmentation mask is injected into SPADE normalization layers, which replace the usual batch- or instance-wise normalizations. This guides the model to learn more effectively and converge more rapidly to an optically appealing result [40].

2.2.4. Training

We have implemented all ITMs as 2D CNNs in tensorflow [41] and trained them on a single GPU (Nvidia Titan RTX 24 GB) for 100 epochs on the combined clinical normal and stroke databases. All models used a batch size of 8, a learning rate of $2{\cdot}10^{-4}$ and otherwise are set up as suggested in the references indicated in the previous section. All models are trained on equal footing, using the available segmented volumes to obtain a conditional generative model that differs from its competitors only in the model architecture, but not the training procedure.

Finally, we train a GAN, which serves as the lesion label generator, for 500 epochs on the 3D lesion masks that are part of the training set. The same lesion generator is used for all ITMs in the following lesion injection.

2.3. Lesion injection and DWI synthesis

In order to synthesize IS data to augment existing clinical data, we use the 2027 healthy brain DWIs' anatomical segmentation maps. In principle, one could also introduce synthetic segmentation maps, which entails complications of its own (mostly related to 3D generative modelling, see Section 3.5). Since this is not the main aspect of our analysis, we resort to the available segmentations of the healthy patient cohort. This also guarantees sufficient anatomical variation for the augmentation. These healthy segmentation maps are subsequently modified by injecting a fake lesion, which in turn is generated by a 3D Wasserstein GAN [42] that is trained on the lesion masks of the stroke DB and whose softmax outputs are transformed into hard segmentations by thresholding at a value of 0.5. The injection is achieved by adapting the label in the parenchyma of the DWI's segmentation map; however, demanding a minimum lesion volume of 20 voxels. If this requirement is not met, a new lesion mask is generated until a lesion of at least 20 voxels is generated.[3] This approach ensures that the anatomical features of the generated brain volumes are realistic and we do not generate a large amount of small, scattered lesions. The Wasserstein loss has proven to be most stable in training. Moreover, it is assumed that the lesion generation by a GAN will generalize previous attempts to data augmentation by synthetic data, e.g. Ref. [10]. The underlying assumption is that generative models yield meaningful interpolations in latent space and at the same time generalizations of simple (geometric) transformations of the available clinical lesion masks.

The generated segmentation map is subsequently decomposed into 2D slices and fed into the ITM in order to generate a fake DWI. In this manner we obtain a database of 2027 fake DWIs per ITM. Fig. 1 shows an example of this lesion implantation procedure, which is also sketched in the graphical abstract. The left column shows the healthy input DWI volume, whose corresponding semantic segmentation map is shown next to it. The third column shows the segmentation map after implanting a lesion label according to the output of the GAN. Subsequently, we have generated realistic DWIs from axial slices of these 3D segmentation maps using three ITMs, namely Pix2Pix, SPADE and cycleGAN, as shown in the right-most column for one axial slice.

[2] Clearly, this identity-loss cannot be used when the input is a multi-channel segmentation map, while the output is a single-channel image. Thus, we do not include an identity loss for training.

[3] We have also investigated smaller thresholds, but found no improvement, as this tends to increase the false-positive rate of the segmentation networks.

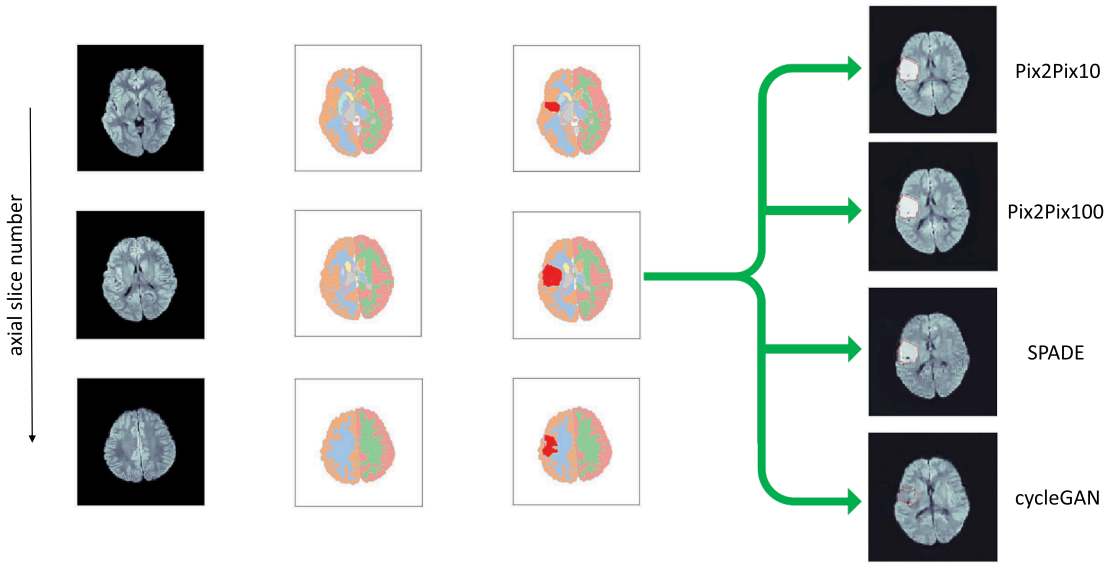

Fig. 1. Implantation of a fake lesion and generation of DWIs. Left: 3 slices of an original healthy DWI. *Center left*: The segmentation of the original volume. *Center right*: Segmentation map with the implanted lesion label generated by the GAN. *Right*: IS volumes generated from the segmentation maps using different ITMs as indicated.

2.4. Evaluation of lesion segmentation

In order to assess the performance of the various ITMs quantitatively, we train a 3D U-Net [21] segmentation network on the clinical, clinical and synthetic, and synthetic-alone databases (for each ITM discussed above) and subsequently evaluate the resulting segmentation masks on a test set comprising 74 stroke patient and 85 healthy patient DWIs, which has been separated from the training data. See Fig. 2 for a visual representation of this procedure.

The U-Net is a well-established standard in biomedical image segmentation and has been used for various tasks, including stroke [10,43], and still yields results that match the state-of-the-art [44]. Training is stopped if the validation loss has not increased for 100 epochs. To this end, 15% of the training data are used for validation. To compare the models' performance quantitatively, we consider a range of 100 epochs, chosen at the end of the training history, to evaluate each model on the test set.

The U-Net architecture is outlined in detail in the Appendix. The network's output is a feature map of equal dimension as the input; however, comprising as many image channels as segmentation labels (in our case two: background and lesion), from which a hard segmentation can be obtained by applying an argmax over the channel axis.

The models are trained on a combination of cross-entropy loss and DSC, which are computed using the soft-max output of the U-Net P and the ground truth G as

$$\mathscr{L}_{\mathrm{U-Net}}(P,G) = -DSC(P,G) + \sum_i G_i \log P_i \tag{2}$$

$$\mathrm{DSC}(P,G) \equiv \frac{2|P \cap G|}{|P| + |G|} = \frac{2\sum_i P_i G_i}{\sum_i P_i + \sum_j G_j}, \tag{3}$$

where $|\cdot|$ denotes the cardinality of a set and the second equality in (3) holds for a pixel-/voxel-based binary segmentation map, and the sums run over all entries. For better comparability, we also report the DSC between the segmentation by two human readers (each with 2 years of experience).

Additionally, we evaluate the quality of the segmentations by a number of additional metrics. First, we evaluate the relative volume difference, given that the lesion volume is an important quantity to triage patients for treatment. Further, we evaluate the Hausdorff distance (HD) and the average symmetric surface distance (ASSD), two quantities that are often reported in order to quantify the quality of the predicted lesion shapes compared to the ground truth. The surface dis-

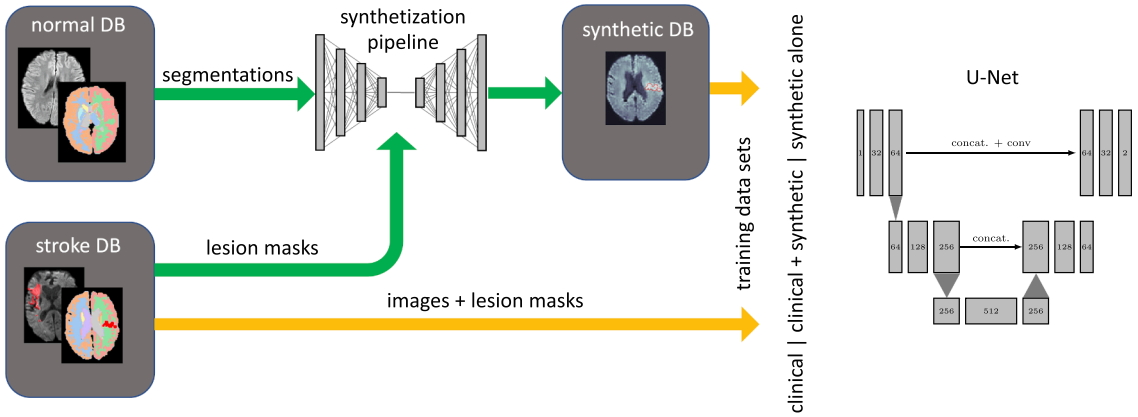

Fig. 2. Evaluation method. Once all ITMs are trained, each of them is used to generate a set of synthetic DWIs which are used as part of the training sets for the U-Net segmentation network. Ultimately, the U-Net is evaluated on a test set of clinical cases. The synthetization pipeline is shown in the graphical abstract.

tance $d_A(b)$ is defined as the minimal Euclidean distance of point $b \in B$ from surface A, i.e. the minimal distance between b and any point $a \in A$. From this, the ASSD is constructed as the symmetric sum,

$$\mathrm{ASSD}(A,B) = \frac{1}{|A|+|B|}\left(\sum_{a\in A} d_B(a) + \sum_{b\in B} d_A(b)\right), \tag{4}$$

while the (symmetrized) HD is the largest symmetric surface distance,

$$\mathrm{HD}(A,B) = \frac{1}{2}\left(\max_{a\in A} d_B(a) + \max_{b\in B} d_A(b)\right). \tag{5}$$

Finally, we evaluate recall and precision of the models. All calculations are relying on the python package `MedPy` [45].

The U-Net is trained on clinical, synthetic, and combined data sets for 500 epochs (clinical data alone), or 300 epochs (all others) using a batch size of 5 and a learning rate of 10^{-4}. To ensure the best benchmark possible, we use standard affine data augmentation including flips, rotations, sheer and translation for the clinical data [46]. In some cases, we further investigate if the performance can be improved by fine-tuning the model trained on synthetic cases using only a handful of clinical cases.

In order to assess the impact of the training set sizes, we have conducted a number of follow-up experiments, where we evaluate the metrics discussed above and calculate mean and standard deviation over the test set for models trained on fractions of the clinical data and the synthetic data. We consider two cases: One, where we used the normal DB's segmentation masks to synthesize fake data, and one where we used the 428 DWI-negative stroke DB cases instead. Finally, we use these 428 cases' segmentations four times to inject lesions with different random seeds to generate a synthetic database of size comparable to that of the normal DB. Moreover, we analyze the impact of different conventional data augmentation techniques, and using no augmentation at all.

3. Results

3.1. Synthetic data generation and qualitative evaluation

Fig. 3 shows some examples of generated images. The synthetic DWIs generated by the ITMs display anatomical structures that are coherent in location, shape, and size across the volume, including the lesions. The first two images from the left have been generated using Pix2Pix. Panel (a) was generated using a reconstruction loss weight $\lambda = 10$, while Panel (b) was created using $\lambda = 100$. The results are of good quality, subjectively realistic and, as anticipated, come accompanied by a high-fidelity lesion label. Only closer inspection of a number of volumes reveals that the larger reconstruction loss weight produces slightly sharper images with more details both inside the lesion and in the brain tissue. We have therefore tested both models quantitatively in the following section.

The lesion intensity produced by SPADE [Fig. 3(c)] is noticeably lower than in the images generated by Pix2Pix, but also the contrast outside the lesion is much lower: in the posterior part of the synthetic brain, barely any gray/white matter contrast is present. While this might seem a disadvantage at first sight, it could actually be beneficial for training a U-Net to reliably detect lesions with lower signal increase.

Next, notice that cycleGAN has failed to recognize and synthesize any lesion hyper-intensity, while producing very high inter-tissue contrast [Fig. 3(d)]. We have experimented with fine-tuning the model on IS data, training on IS data alone, and modifying the network architecture; however, this result seems to be unavoidable. We speculate that this is closely related to the cycle consistency requirement of cycleGAN, i.e. that the two generators map an image back to itself as close as possible, and large class imbalance between lesion and non-lesion voxels. Similarly, Ref. [37] identified 'failed cases' which indicate that cycleGAN generalizes poorly. Due to this failure, we have excluded cycleGAN from the quantitative analysis, and leave a dedicated study using cycleGAN for data augmentation for future work. We stress that this does not preclude any application of cycleGAN to this task; however, it will require an approach that differs from the one chosen here, and thus the comparability of ITMs is not given.

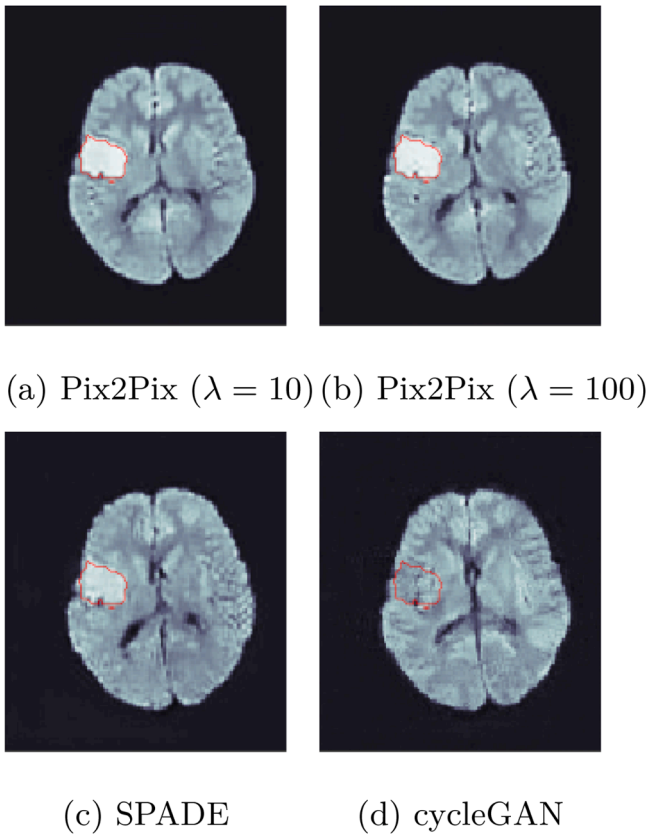

Fig. 3. Generated stroke DWIs from the segmentation mask in Fig. 1.

To conclude this section, we present in Fig. 4 the intensity histograms of the synthetic and real data sets after pre-processing and normalized to the range $[0, 1]$. Notice how the normal and clinical (i.e. stroke) DBs significantly differ in the region of intensities below 0.5. Moreover, the synthetic image intensities closely follow the normal volume intensities. This does not come as a surprise as we have used the normal segmentation masks to synthesize stroke images. One is led to the conclusion that some systematic difference, either of anatomical nature or rooted in

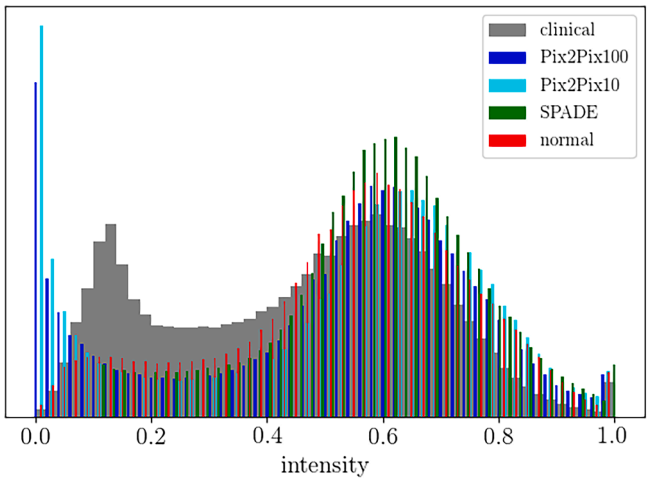

Fig. 4. Intensity distribution of the synthetic and clinical stroke images.

the image acquisition, is responsible for this discrepancy. We will now turn to the assessment of the quality of image generation and return to this question in Section 3.5.

3.2. Model selection and test set evaluation

The results are shown in Fig. 5 and summarized in Table 1. In our set-up, the U-Net trained on the clinical data reaches a mean test set DSC of $(63.2 \pm 1.9)\%$ and a maximal DSC of 67.3%. The U-Nets trained on the combined synthetic data from Pix2Pix and clinical data outperform the clinical baseline U-Net and we find mean DSCs of $(70.3 \pm 1.1)\%$ (Pix2Pix; $\lambda = 100$) and $(70.8 \pm 1.0)\%$ (Pix2Pix; $\lambda = 10$), while SPADE performs slightly worse (62.9 ± 1.2). The maximum DSC achieved is 72.8% for Pix2Pix with $\lambda = 10$. The other models achieve maximal DSC of 66.5% (SPADE) and 72.5% (Pix2Pix; $\lambda = 100$). This is a key result of our study highlighting that synthetic data does not simply reproduce available information in the training data, but instead interpolates and effectively generalizes the training data.

In contrast to this, the U-Net performs much worse when trained on synthetic data alone, and with much greater variance. We find the test set DSCs $(57.1 \pm 2.7)\%$ (Pix2Pix; $\lambda = 100$), $(55.7 \pm 3.3)\%$ (Pix2Pix; $\lambda = 10$), and $(47.3 \pm 2.2)\%$ (SPADE). Interestingly, and different from the previous result, the larger reconstruction loss yields better results for Pix2Pix in agreement with our qualitative observation. To put these findings into perspective we report the human inter-reader DSCs of $76.6 \pm 13.9\%$ for the training and $76.9 \pm 13.5\%$ for the test set.

3.3. Model fine-tuning

These results indicate that synthetic data alone is not apt to replace the clinical data set. Whether this can be attributed to the quality of the generated images or differences in anatomical features between the databases used for image synthesis and evaluation, will be investigated in the upcoming section. First, we address the question whether the models trained on synthetic data alone can be fine-tuned to yield better results.

To this end, we fine-tune a model previously trained on synthetic data alone on either 10 or 50 randomly chosen clinical training samples for another 150 epochs. In Fig. 6 we duplicate the previous result, where the model was trained on synthetic data alone (orange). Subsequently, we show the results after training on 10 (red) and 50 clinical cases (dark red), respectively. Notice the difference between the two distinct architectures, Pix2Pix and SPADE. Increasing the tuning set size from 10 to 50 yields only a minor improvement for SPADE: from $(51.0 \pm 1.2)\%$ for

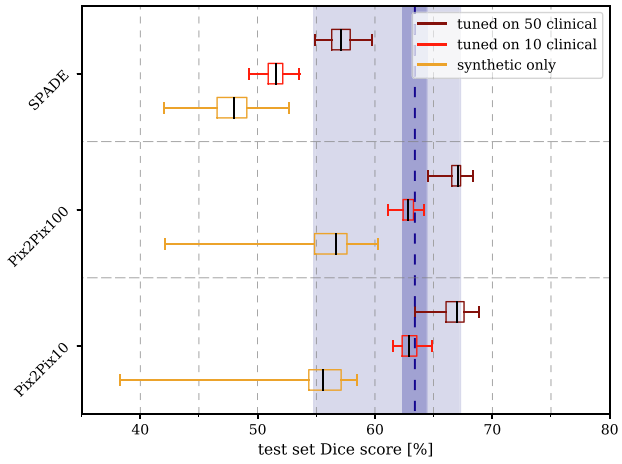

Fig. 5. Box plot of the various ITMs evaluated on the clinical test set. Boxes indicate standard deviation and whiskers min- and max-values.

Table 1
Summary of the resulting DSCs (in percent) of the different models trained with various amounts of synthetic data. The last two rows correspond to the fine-tuning procedure described in Section 3.3.

data set	baseline	Pix2Pix10	Pix2Pix100	SPADE
clinical	**63.2 ± 1.9**	–	–	–
+ all synthetic	–	**70.8 ± 1.0**	70.3 ± 1.1	62.9 ± 1.2
synthetic	–	55.7 ± 3.3	**57.1 ± 2.7**	47.3 ± 2.2
+ 10 clinical	–	**62.9 ± 0.8**	62.3 ± 0.9	51.0 ± 1.2
+ 50 clinical	–	66.4 ± 1.1	**66.8 ± 0.7**	56.4 ± 1.5

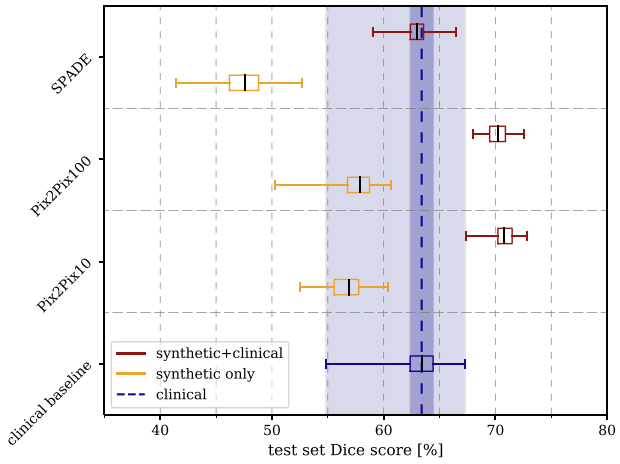

Fig. 6. Box plots for fine-tuned models. The performance of models trained on synthetic data alone can be improved to match the clinical case, or even outperform it in some cases.

10 samples to $(56.4 \pm 1.5)\%$ for 50 samples. After tuning on 50 samples, SPADE's performance matches that of the lower spectrum of the clinical baseline model. The situation is quite distinct for Pix2Pix, which for 10 tuning samples matches the performance of the clinical baseline. Tuning on 50 samples yields enough information for Pix2Pix to outperform the clinical model with a mean DSC of $(66.4 \pm 1.1)\%$ (max. 68.9%; $\lambda = 10$) and $(66.8 \pm 0.7)\%$ (max. 68.3%; $\lambda = 100$). In all cases the variance of DSCs is much lower compared to the clinical baseline.Had we trained the U-Net on only 10 [50] cases, the outcome would be much worse: we find a mean DSC of $(2.9 \pm 1.6)\%$ [$(12.5 \pm 7.0)\%$].

Therefore, we may conclude that models trained on synthetic data alone do not match the performance of those trained on clinical data sets. However, these models can be fine-tuned with only few clinical training cases. This ultimately yields results that are comparable to the clinical baseline model trained on all 365 clinical IS cases; however, with much less variance. Moreover, in a situation where only a limited number of clinical cases is available, this approach yields a significant improvement as we highlight in the upcoming section.

In Fig. 7 we show some examples for segmentations on two clinical test cases, one with a big lesion (top row) and one with small, disconnected lesions (bottom row). The U-Nets have been trained on clinical data (left-most panels), or on the combined data sets as indicated (other panels). Alongside the labeled areas, the ground truth (red) and the corresponding DSC (for the whole volume) are indicated.

3.4. Further experiments

Let us now focus on the model that obtained the highest DSC in Section 3.2, i.e. Pix2Pix ($\lambda = 10$) and consider it in more detail. Table 2 and Fig. 8 show the results of the four different types of experiments we have conducted. The first for rows contain the results of training the U-

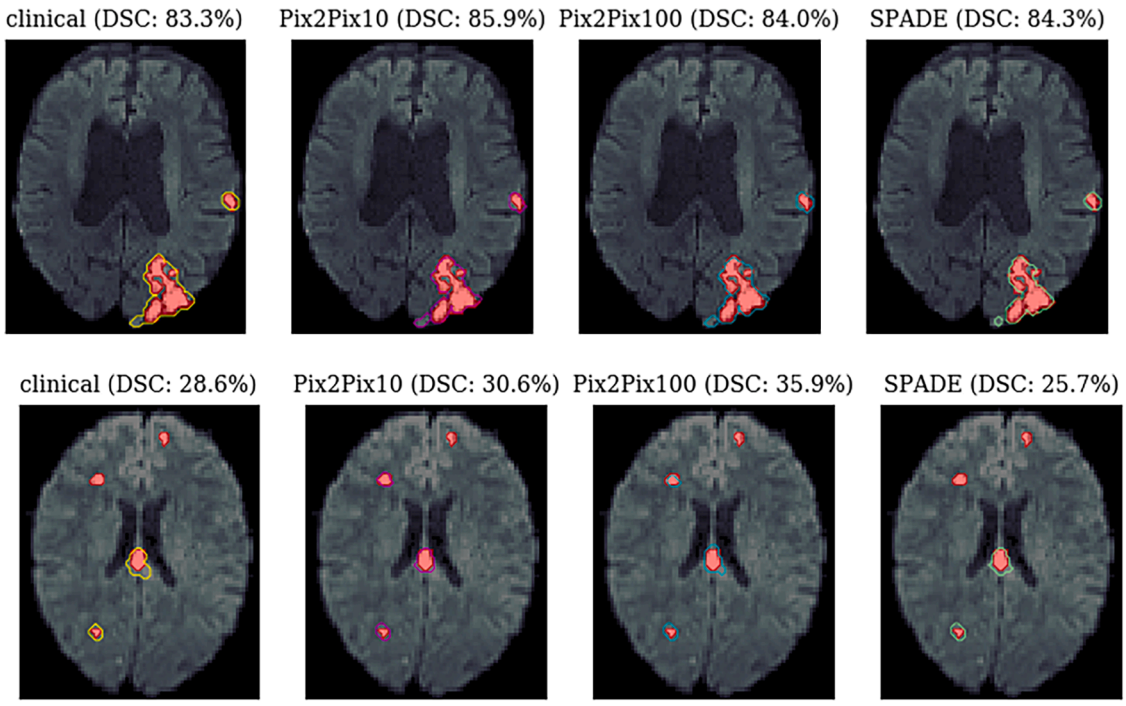

Fig. 7. Two segmentation examples: A large lesion (*top*), which is segmented well by all models. And smaller, disconnected lesions (*bottom*). The latter case shows greater DSC variance among the models.

Table 2
Experiments with varying training set sizes and compositions. Results indicate mean and standard deviation over the test set The first section compares models with fractional clinical data as indicated. The second section compares different data augmentations. In the third and forth column, the results using synthetic data from Pix2Pix are shown ($\lambda = 10$) synthesized from the normal DB (third column) and the stroke DB (fourth column). In each section, results are ordered by increasing DSC and we include the additional metrics discussed in Section 2.

model	DSC [%]	rel. volume loss [%]	HD [mm]	ASSD [mm]	precision [%]	recall [%]
10% clinical	26.5 ± 29.7	58.2 ± 0.3	42.7 ± 32.7	14.0 ± 20.9	52.7 ± 42.2	21.2 ± 26.5
25% clinical	43.1 ± 29.7	42.9 ± 0.3	40.5 ± 30.2	8.6 ± 16.0	68.1 ± 32.7	38.3 ± 31.8
50% clinical	61.1 ± 24.5	44.6 ± 2.4	47.6 ± 29.3	7.8 ± 16.6	71.3 ± 22.7	57.8 ± 27.0
100% clinical	62.7 ± 23.9	78.7 ± 3.5	33.2 ± 25.5	4.7 ± 10.7	75.6 ± 22.5	59.6 ± 27.2
geometrical augment + noise	62.3 ± 25.5	86.5 ± 3.5	35.3 ± 28.4	6.5 ± 14.8	74.7 ± 24.8	59.4 ± 28.4
no augmentation	58.9 ± 23.8	51.1 ± 1.6	38.3 ± 26.0	6.3 ± 13.9	73.7 ± 22.4	54.7 ± 27.8
noise	55.9 ± 24.8	166.4 ± 11.4	43.4 ± 28.5	6.4 ± 11.7	72.3 ± 24.9	51.0 ± 27.1
geometrical + intensity augmentation	53.3 ± 28.0	33.1 ± 0.3	34.9 ± 28.4	6.5 ± 12.7	80.1 ± 25.4	45.1 ± 28.5
intensity augmentation	54.9 ± 26.9	387.2 ± 26.7	42.2 ± 25.7	8.0 ± 16.7	71.4 ± 26.9	51.0 ± 29.1
10% clinical + Pix2Pix (normal DB)	55.3 ± 26.0	29.9 ± 0.3	38.6 ± 29.1	7.0 ± 13.1	79.2 ± 24.7	46.6 ± 26.4
25% clinical + Pix2Pix (normal DB)	56.3 ± 27.1	27.9 ± 0.3	39.4 ± 28.0	7.5 ± 16.4	79.8 ± 25.0	48.5 ± 28.2
50% clinical + Pix2Pix (normal DB)	58.0 ± 27.1	28.3 ± 0.3	36.6 ± 29.5	6.3 ± 14.2	79.1 ± 24.4	51.3 ± 28.9
100% clinical + Pix2Pix (normal DB)	67.3 ± 19.7	15.5 ± 0.2	32.2 ± 26.4	4.1 ± 10.6	81.8 ± 15.6	61.8 ± 23.9
10% clinical + Pix2Pix (stroke DB)	51.9 ± 26.5	27.4 ± 0.3	47.6 ± 28.8	8.1 ± 11.7	72.8 ± 24.9	44.6 ± 27.3
25% clinical + Pix2Pix (stroke DB)	55.9 ± 26.7	27.5 ± 0.3	42.5 ± 27.9	5.2 ± 6.2	76.9 ± 20.7	50.4 ± 29.2
50% clinical + Pix2Pix (stroke DB)	56.0 ± 26.2	35.3 ± 0.6	38.0 ± 26.0	5.6 ± 8.4	77.1 ± 20.8	50.9 ± 29.9
100% clinical + Pix2Pix (stroke DB)	58.6 ± 26.6	41.8 ± 1.0	37.3 ± 29.3	7.0 ± 14.4	77.0 ± 25.1	54.1 ± 29.7
100% clinical + 4 × Pix2Pix (stroke DB)	53.7 ± 22.8	137.7 ± 8.9	54.6 ± 26.5	8.7 ± 10.3	68.3 ± 26.3	48.6 ± 23.4

Net on a fraction of 10/25/50% of the clinical data and evaluated on the full test set. Notice that in all cases, the error is rather large and decreases only slightly when more clinical data is included, indicating that the data set is rather diverse. The DSC increases quickly as the training set size grows; however, only a small improvement is found when going from 50% to the full data set. This behavior is contrasted by the relative volume loss, which is in fact largest for the full training set. Closer inspection reveals that cases with large volume loss have small lesion volumes (see also Fig. 7), and thus the volume loss does not seem to be a suitable metric when lesions of varying sizes are studied. All other metrics confirm that using the full training set yields the best performance.

In the second class of experiments, we utilized the full training set and used different data augmentation techniques, including adding random noise and randomly remapping the image intensities as suggested in Ref. [47].[4] However, none of the metrics indicates that any

[4] Recall that, unless indicated otherwise, we always use geometrical data augmentation.

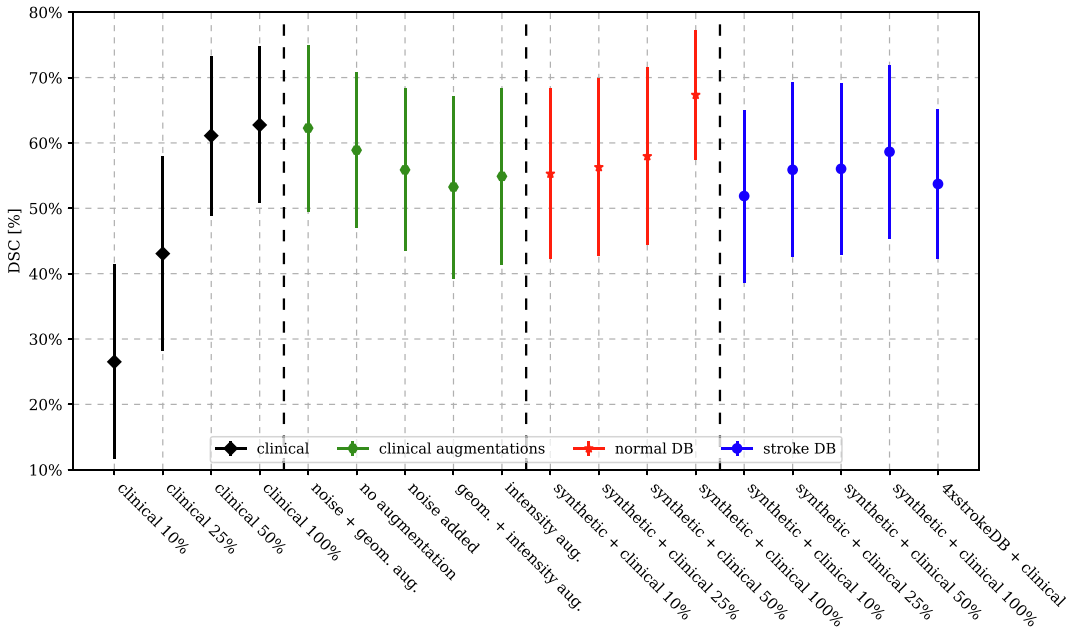

Fig. 8. Results of the additional experiments, which correspond to the first column of Table 2.

other data augmentation technique performs better than geometrical transformation alone.

In the former case, significant improvements can be obtained for the small fractional training data sets, e.g. with less than 40 DWI-positive clinical cases (10%), a DSC above 50% can be obtained. Other metrics indicate the same behavior. Interestingly, the relative volume loss is monotonically decreasing with increasing number of clinical cases, indicating that the model's predictions are more stable for smaller lesions.

One might have suspected that using segmentation masks from the stroke DB (which on average have larger ventricles due to the higher age of the individuals), would further improve the performance; however this is not the case. Even when recycling the segmentation masks four times (which yields a data set of size equal to the normal DB) and augmenting with different lesion labels, we do not find any improvement. Thus, we may conclude that the increased variability in the normal DB is beneficial to the performance of the U-Net. Moreover, we conclude that the deficiencies of the models trained on synthetic data alone are not rooted in the out-of-distribution nature of the segmentation masks we used, but rather a limitation of the ITMs themselves, or the training set size, respectively.

Finally, we have evaluated the trained models on the ISELS 2015 DWI images [7]. We found that the models under consideration unanimously yield DSCs around 50% with little variation among models. However, the variation across this test set is rather large with a standard deviation of around 30%. Closer inspection of some of the cases reveals that the ISLES data is much more coarsely segmented, such that the DSCs are actually limited by the segmentation quality of this test set rather than the models' capacities. Qualitatively, the results are of high quality and only a few cases, with infraction in the brain stem are missed by our models. See Fig. 9 for an example, where the model provides a very detailed outline of the segmentation. The ground truth segmentation, on the other hand is very coarse, including even CSF and therefore structures that are not part of the parenchyma. This strongly supports our hypothesis that the quality of the data inhibits the scores of the models rather than the model performance itself.

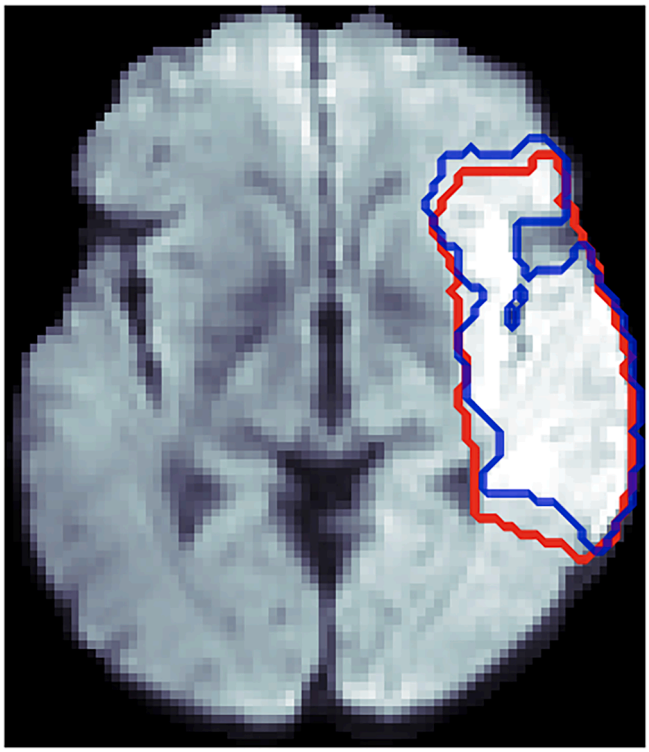

Fig. 9. Segmentation result on a sample from the ISLES 2015 DWI training data. Red: ground truth, blue: model prediction. Notice the coarse ground truth where cerebrospinal fluid is labeled as belonging to a lesion, whereas the model prediction correctly follows the brain tissue boundaries.

3.5. Discussion

The results illustrate that DL-based data augmentation can be used to optimize model performance. Depending on the number of available clinical samples, the improvement can be substantial, as summarized in Tables 1 and 2.

Training on the clinical set augmented by synthetic data, Pix2Pix with a reconstruction loss weight of $\lambda = 10$ is the choice that yields the best results. Training instead on synthetic data alone, we find that increasing the reconstruction loss weight to $\lambda = 100$ results in the highest DSC. With 10 training samples to fine-tune this model, it yields results comparable to the clinical data set, while it produces better segmentations with only 50 fine-tuning samples. This result matches the qualitative finding that a larger reconstruction loss yields sharper images than the recommended $\lambda = 10$; however, the differences are rather marginal. A more economical model with less adjustable parameters is the SPADE framework, which is attractive for two reasons: First, less trainable weights signify less time for the model to converge and require less training data. And secondly, since SPADE samples from a latent space, it introduces elements of stochasticity as opposed to Pix2Pix, which produces deterministic outputs.

As stated in Section 2.1, the mean age within each database differs significantly [72 ± 14 years (stroke) compared to 38 ± 24 years (normal)]. Consequently, the stroke DB displays anatomical peculiarities that can be attributed to the advanced age of the patients, most notably enlarged ventricle volumes. For example, the mean lateral ventricle volumes are (9.1 ± 6.6) ml in the normal DB versus (20 ± 12) ml in the stroke DB. We have verified that all other anatomical features are consistent between the databases. Moreover, being mostly recorded on a 1.5T scanner, the stroke DB images have lower signal-to-noise ratios than those in the normal DB. To investigate whether this poses a systematic issue, we have generated synthetic DWIs from lesion-free segmentation masks in the stroke DB, but did not find any improvement, see Fig. 8.[5] Thus we conclude that synthetic data is beneficial if it introduces additional anatomical variation.

Finally, we have seen that all models perform well on large lesions, as can be seen in the upper panels of Fig. 7. Conversely, the case shown in the bottom panels, which entails several smaller lesions, is not segmented equally well by all models; especially the U-Net trained only on clinical data produces false positive predictions. This appears to be a common feature and *a posteriori justifies the lower lesion threshold of 20 voxels, which inhibits too many false predictions. We have also trained models without any restriction on the lesion volume; however, found only worse performance and more false-positive labels.*

Our study has several limitations that need to be addressed in future work. First, this is a single-center study using data from only one institution for training. While we did attempt to use the ISLES 2015 data for evaluating how well our models generalize, more quantitative analyses are necessary to guarantee robustness. Second, while DWIs are indeed the gold standard for lesion detection and quantification, CT is the clinically more common imaging modality, even though the lesion contrast is much lower. In the future, an important question will be how generative models can be used to generate more than one contrast. However, the issue of domain adaptation in medical imaging is subject of ongoing research, see e.g. Refs. [47–49].

Comparing our results to Ref. [10], two key aspects should be highlighted: There, the authors used masked lesions extracted from real DWIs of stroke patients to implant them into real DWIs of healthy patients, thus creating a similar database of synthetic images to train a segmentation network. While this creates fully realistic 3D images (at least outside the lesions), the approach is combinatorically limited, while our approach could in principle generate a continuum of new, unseen images either from a (continuous) latent space representation in the case of SPADE and the lesion generating GAN, or continuous deformations of segmentation masks. While there, results similar to ours were obtained using clinical data augmented with some 2000 synthetic cases, the authors did not discuss the use of synthetic images alone. A key result of our study is that a U-Net trained only on synthetic images generated by ITMs, cannot keep up with the results of a model trained on clinical data. However, such models can be fine-tuned on just a handful of clinical cases yielding result comparable to those obtained in a purely clinical setting with many high-quality segmentations.

To address the question of 3D coherence, we have also attempted to train 3D ITMs. However, training 3D generative models is difficult, because these models tend to quickly exhaust available memory capacities such that the image resolution or the training batch size are limited. Moreover, GANs often do not converge when trained on 3D data. Thus, it turns out that the proposed pipeline is the most economical setup, reliably producing 3D brain volumes and IS lesions of accepTable 3D coherence. This is owing to the fact that we used real anatomical segmentations to feed into the ITMs.

To put the results of this study into perspective, we report the results of a some recent approaches to the issue of stroke lesion segmentation. The aforementioned Ref. [10] reported a maximal mean DSC of 72.1% obtained on *the same* test set as ours. This is to be compared with our maximal DSC of 70.8%, in agreement within one standard deviation. Another approach comparable to ours was presented in Ref. [50] using a generative model to obtain DWIs from CT input images. This approach yielded a DSC of 62.4% compared to a baseline DSC of 57.2% without the generator model. This highlights possible applications of our method in cross-modality analyses, but also confirms our finding that the use of generative models generically enhances the quality of segmentations. Notice that the best DSC there compares to the DSC we obtain on clinical DWIs, but was obtained on CT, where IS is much less pronounced as is the case for DWIs. Next, Ref. [51] chose to approach the segmentation of stroke lesions by using Markov random fields on FLAIR images obtaining a DSC of 58.2%. Given the different contrast considered in this reference, this number cannot be directly compared to our results. However, the proposed method systematically outperforms segmentation approaches based on convolutional neural networks, thus providing a potential direction to investigate in the future. Similarly, Ref. [52] proposed a so-called zoom-in zoom-out data augmentation method that significantly improved the quality of neural network based segmentations of stroke lesions on T1-weighted MR images. This indicates that shape deformations could be a promising way to improve the quality of segmentations (also on synthetic images).

4. Conclusions & Outlook

We have investigated the applicability of DL-based image-to-image translation models to generalize data augmentation of medical image data sets. Our results highlight that DL-based algorithms yield significantly better performance than traditional data augmentation techniques. This is in line with multiple recent studies, which have investigated the performance of different image-to-image translation models. However, only few publications exist to date, which subsequently study the potential improvement of DL-based segmentation and diagnosis systems, due to the synthetically enhanced training data sets [15,16]. Our work shows that this increase in data set size improves the quality of the segmentation network significantly. In addition, it has a number of interesting consequences, such as using synthetic data to train disease detection algorithms, e.g. for IS lesion segmentation as in our case, avoids any data privacy obstructions. Thus, data in the form of trained ITMs can be made available to a much broader audience,which promises to accelerate the advances in machine learning applications in diagnostic medicine. Moreover, the technique can be used to augment available data sets with synthetic data.

Following up on our results, several directions are conceivable. As we have seen, training on synthetic data alone does not provide competitive results. A more careful data preparation could provide contrasts that are more aligned with those of the clinical data; however, at the price of fine-tuning the model to a single data set. Alternatively, using the idea of life-long learning [53,54], a modified network architecture could be constructed, such as the one by [55], which allows users

[5] This is also true if the ITM is trained only on stroke DB data and not all data combined.

to easily adapt the models to their data, even if acquired under entirely different circumstances. Incorporating more than one contrast for segmentation is another promising avenue. Finally, the proposed pipeline is straightforwardly generalized to other pathologies that can be identified on DWI, but also on other imaging modalities of the entire human body. A particularly interesting question is whether the diagnosis of rare diseases can be improved by augmenting available data sets by artificial data in order to compensate the intrinsic imbalance of the clinical training data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Sebastian Kozerke and Thomas Joyce for useful discussions.

Appendix A. Model architectures

In this appendix, we use the following abbreviations: Ck is a convolution, CTk a transposed convolution layer with k filters, D is a dropout layer, (L)ReLU is a (leaky) rectified linear unit activation layer, M a max-pooling layer, U a nearest neighbor upsampling layer, BN a batch normalization layer, and IN instance normalization.

A.1. U-Net

The 3D U-Net [21] consists of a downsampling and an upsampling branch, connected by a bottleneck layer. In the downsampling branch, we include 4 blocks D16-D32-D64-D-128, the upsampling path is U128-U64-U32-U16-C*n*, where n is the number of segmentation labels, in this case $n = 2$ (IS and non-IS), and the final convolution uses a softmax activation. The downsampling blocks D*k* are consecutive layers C*k*-ReLU-D-C*k*-ReLU-D-M with stride-1 convolutions, while the upsampling blocks U*k* are CT*k*-BN-ReLU-C*k*-ReLU-D-C*k*-ReLU-D with stride-1 convolutions and stride-2 transposed convolutions. To achieve the U-Net's characteristic skip connection, the output of each downsampling block is concatenated with the output of the transposed convolution in the upsampling branch at the matching image resolution. Between the two branches, the so-called bottleneck is a block C256-ReLU-D-C256-ReLU-D. The dropout rate is 2% and all convolutional kernels are $3 \times 3 \times 3$. The U-Net operates at an image resolution of $128 \times 128 \times 32$ and has a total of $5.6 \cdot 10^6$ trainable parameters. We train the model using a batch size of 5, reflecting memory restrictions.

A.2. Pix2Pix

The 2D Pix2Pix generator architecture is based on the U-Net, comprising a down- and an upsampling path. The details of the generator are largely identical to the proposed architecture by [28]. However, we replaced the batch-wise normalization with an instance-wise normalization [56]. The downsampling path is D64-D128-D256-D512-D512-D512-D512-D512, where the downsampling blocks D*k* are C*k*-IN-LReLU. Each convolutional layer contains a 4×4 kernel, stride 2, and is followed by the InstanceNorm. In the upsampling path we use upsampling blocks U256-U128-U64, each U*k* consists of a sequence of layers *U*-C*k*-LReLU-D-IN-LReLU and stride-1 convolutions with kernel size 4×4. The downsampling block's output is concatenated before the final ReLU layer with a leakiness of 0.2. The generator model has a total of $41.8 \cdot 10^6$ trainable parameters, and we use a batch size of 8 for training.

A.3. SPADE

The SPADE generator samples a latent vector from a 128-dimensional random normal distribution centered at the origin with isotropic variance of unit magnitude. The latent vector is then passed to a dense layer with output size $4 \cdot 4 \cdot 128$ and reshaped into a feature map with dimension $4 \times 4 \times 128$. The following five residual blocks follow the original design by [40], including the SPADE normalization layers, which process the segmenattion maps. The design of each block is SPADE-LReLU-C*k*-SPADE-LReLU-C*k*, where the convolutions have a kernel size of 3×3 and unit stride as in the original setup; however, notice that we work with a reduced number of filters k (128–64-32–16-8), the leaky ReLU layers have a slope of 0.2, and each SPADE block is followed by an upsampling layer until the desired resolution of $128 \times 128 \times 32$ is reached. A final 3×3 convolution is added at the end with a tanh activation and a single filter to match the data dimensions. The model is trained with a batch size of 8.

A.3.1. Discriminator

All ITMs use the PatchGAN discriminator [28,35,36], a convolutional network whose output is not a single number as is the case for ordinary discriminator networks. Instead PatchGAN outputs an array of numbers, which, due to the convolutional character of the network, are connected only to a patch of the input image. The size and number of these patches depends on the number of convolutional layers. In the original Pix2Pix setup, where 256×256 images were considered, the receptive field was 70×70. Here we have chosen three convolutions, C32-IN-LReLU-C64-IN-LReLU-C128-IN-LReLU, which reduces the receptive field from 70×70 to 34×34 – a choice which reflects the fact that the DWI slices under consideration measure 128×128. The final layer is followed by a stride-one, size-one filter, 4×4 convolution to map to the desired output shape. All other convolutions have kernel sizes 4×4, stride 2, and filters as indicated; and the leaky ReLU has a slope of 0.2. The discriminator model has a total of 888,898 parameters.

A.4. Fake lesion generation

We have experimented with several architectures and loss functions and found that a Wasserstein GAN [42] with a partial gradient loss [57] yields the best results. This model can be trained on 3D lesion masks in a stable manner, and therefore provides the ideal building block for our synthesis pipeline. The output activation of this model is a softmax function, which we transform into a lesion prediction by thresholding. The GAN is trained on the 449 lesion masks that are available in the IS database with a batch size of 8 for 500 epochs. We use Instance normalization and Leaky ReLU

(leakiness 0.2) activations after each convolution-upsampling step.

References

[1] WHO, "The top ten causes of death," URL:https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2018, accessed: 08/10/2020.

[2] J. Tsai, M. Mlynash, S. Christensen, S. Kemp, S. Kim, N. Mishra, C. Federau, R. Nogueira, T. Jovin, T. Devlin, N. Akhtar, D. Yavagal, R. Bammer, M. Straka, G. Zaharchuk, M. Marks, G. Albers, M. Lansberg, Time from imaging to endovascular reperfusion predicts outcome in acute stroke, Stroke 49 (4) (Apr. 2018) 952–957.

[3] J. Saver, M. Goyal, A. Lugt, B. Menon, C. Majoie, D. Dippel, B. Campbell, R. Nogueira, A. Demchuk, A. Tomasello, P. Cardona, T. Devlin, D. Frei, R. Rochemont, O. Berkhemer, T. Jovin, A. Siddiqui, W. Zwam, S. Davis, and M. Hill, "Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: A meta-analysis," JAMA, vol. 316, pp. 1279–1288, 09 2016.

[4] N. Hainc, C. Federau, B. Stieltjes, M. Blatow, A. Bink, C. Stippich, The bright, artificial intelligence-augmented future of neuroimaging reading, Frontiers in Neurology 8 (2017) 489.

[5] J. Bernal, K. Kushibar, D.S. Asfaw, S. Valverde, A. Oliver, R. Martí, and X. Lladó, "Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review," Artificial Intelligence in Medicine, vol. 95, pp. 64–81, 2019. [Online]. Available: URL:http://www.sciencedirect.com/science/article/pii/S0933365716305206.

[6] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: A review, Medical Image Analysis 58 (2019), 101552. URL:https://www.arxiv.org/abs/1809.07294.

[7] O. Maier et al., "ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI Medical Image Analysis," URL:https://doi.org/10.1016/j.media.2016.07.009, 2016, URL:www.isles-challenge.org.

[8] Stanford Vision Lab, Stanford University, Princeton University, "ImageNet," URL: http://www.image-net.org, 2016, accessed: 08/10/2020.

[9] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, in: in Advances in neural information processing systems, 2014, pp. 2672–2680. URL:https://www.arxiv.org/abs/1406.2661.

[10] C. Federau, S. Christensen, N. Scherrer, J.M. Ospel, V. Schulze-Zachau, N. Schmidt, H.-C. Breit, J. Maclaren, M. Lansberg, and S. Kozerke, "Improved segmentation and detection sensitivity of diffusion-weighted stroke lesions with synthetically enhanced deep learning," Radiology: Artificial Intelligence, vol. 2, no. 5, p. e190217, 2020. [Online]. Available: URL:https://doi.org/10.1148/ryai.2020190217.

[11] H.-C. Shin, N.A. Tenenholtz, J.K. Rogers, C.G. Schwarz, M.L. Senjem, J.L. Gunter, K.P. Andriole, M. Michalski, Medical image synthesis for data augmentation and anonymization using generative adversarial networks, in: A. Gooya, O. Goksel, I. Oguz, N. Burgos (Eds.), Simulation and Synthesis in Medical Imaging, Springer International Publishing, Cham, 2018, pp. 1–11.

[12] B. Yu, L. Zhou, L. Wang, J. Fripp, P. Bourgeat, 3d cgan based cross-modality mr image synthesis for brain tumor segmentation, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 626–630. URL:https://arxiv.org/abs/1810.11654.

[13] Q. Yang, N. Li, Z. Zhao, X. Fan, E.I.-C. Chang, and Y. Xu, "Mri cross-modality image-to-image translation," Scientific Reports, vol. 10, no. 1, p. 3753, 2020. [Online]. Available: URL:https://doi.org/10.1038/s41598-020-60520-6.

[14] J. Rubin, S.M. Abulnaga, "Ct-to-mr conditional generative adversarial networks for ischemic stroke lesion segmentation," in, in: 2019 IEEE International Conference on Healthcare Informatics (ICHI), 2019, pp. 1–7.

[15] J. Jiang, Y.-C. Hu, N. Tyagi, P. Zhang, A. Rimner, G.S. Mageras, J. Deasy, H. Veeraraghavan, Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation, Med. Image Comput Comput. Assist Interv. 11071 (2018) 777–785.

[16] E. Vorontsov, P. Molchanov, W. Byeon, S.D. Mello, V. Jampani, M. Liu, S. Kadoury, and J. Kautz, "Boosting segmentation with weak supervision from image-to-image translation," arXiv e-prints, vol. abs/1904.01636, 2019. [Online]. Available: URL: http://arxiv.org/abs/1904.01636.

[17] P. Klages, I. Benslimane, S. Riyahi, J. Jiang, M. Hunt, J.O. Deasy, H. Veeraraghavan, and N. Tyagi, "Patch-based generative adversarial neural network models for head and neck mr-only planning," IEEE Transactions on Medical Imaging, vol. XX, no. X, 2020. [Online]. Available: URL:https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13927.

[18] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "Medgan: Medical image translation using gans," Computerized Medical Imaging and Graphics, vol. 79, p. 101684, 2020. [Online]. Available: URL: https://www.sciencedirect.com/science/article/pii/S0895611119300990.

[19] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," 2017.

[20] S. Pandey, P.R. Singh, and J. Tian, "An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation," Biomedical Signal Processing and Control, vol. 57, p. 101782, 2020. [Online]. Available: URL: https://www.sciencedirect.com/science/article/pii/S1746809419303635.

[21] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241. URL:https://www.arxiv.org/abs/1505.04597.

[22] McConnell Brain Imaging Centre, Montreal Neurological Institute, "MNI human atlas," URL:http://nist.mni.mcgill.ca/?page_id=714, 2009, accessed 08/10/2020.

[23] M. Jenkinson, C.F. Beckmann, T.E. Behrens, M.W. Woolrich, and S.M. Smith, "Fsl," NeuroImage, vol. 62, no. 2, pp. 782–790, 2012, 20 YEARS OF fMRI. [Online]. Available: URL:https://www.sciencedirect.com/science/article/pii/S1053811911010603.

[24] University of Pennsylvania, Image Computing & Science Lab, "Advanced Normalization Tools (ANTs)," URL:http://picsl.upenn.edu/software/ants, 2014, accessed 08/10/2020.

[25] B. Fischl, D.H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A.M. Dale, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," Neuron, vol. 33, no. 3, pp. 341–355, 2002. [Online]. Available: URL:http://www.sciencedirect.com/science/article/pii/S089662730200569X.

[26] J. Zopes, M. Platscher, S. Paganucci, and C. Federau, "Multi-modal segmentation of 3d brain scans using neural networks," 2020, URL:https://www.arxiv.org/abs/2008.04594.

[27] A.U. Hirte, M. Platscher, T. Joyce, J.J. Heit, E. Tranvinh, and C. Federau, "Diffusion-weighted magnetic resonance brain images generation with generative adversarial networks and variational autoencoders: A comparison study," arXiv e-prints, 2020, URL:https://www.arxiv.org/abs/2006.13944.

[28] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2017. URL:https://www.arxiv.org/abs/1611.07004.

[29] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798–8807. URL:https://www.arxiv.org/abs/1711.11585.

[30] J.M. Wolterink, T. Leiner, M.A. Viergever, I. Išgum, Generative adversarial networks for noise reduction in low-dose ct, IEEE Transactions on Medical Imaging 36 (12) (2017) 2536–2545.

[31] H. Choi, D.S. Lee, and A.D.N. Initiative, "Generation of structural mr images from amyloid pet: Application to mr-less quantification," Journal of nuclear medicine: official publication, Society of Nuclear Medicine, vol. 59, no. 7, pp. 1111–1117, 07 2018. [Online]. Available: URL:https://pubmed.ncbi.nlm.nih.gov/29217736.

[32] Y. Huo, Z. Xu, S. Bao, C. Bermudez, A. Plassard, Y. Yao, J. Liu, A. Assad, R. Abramson, B. Landman, Splenomegaly segmentation using global convolutional kernels and conditional generative adversarial networks 10574 (03) (2018) 8.

[33] H. Emami, M. Dong, S.P. Nejad-Davarani, and C.K. Glide-Hurst, "Generating synthetic cts from magnetic resonance images using generative adversarial networks," Medical Physics, vol. 45, no. 8, pp. 3627–3636, 2018. [Online]. Available: URL:https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13047.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, URL:https://www.arxiv.org/abs/1512.03385.

[35] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 702–716, URL:https://www.arxiv.org/abs/1604.04382.

[36] C. Ledig, L. Theis, F. Huszar, J. Caballero, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," 2017, pp. 4681–4690, URL:https://www.arxiv.org/abs/1609.04802.

[37] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251. URL:https://www.arxiv.org/abs/1703.10593.

[38] H. Yang, J. Sun, A. Carass, C. Zhao, J. Lee, Z. Xu, J. Prince, Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan, in: D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J.M.R. Tavares, A. Bradley, J.P. Papa, V. Belagiannis, J.C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, A. Madabhushi (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham, 2018, pp. 174–182.

[39] V. Sandfort, K. Yan, P.J. Pickhardt, and R.M. Summers, "Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks," Scientific reports, vol. 9, no. 1, p. 16884, November 2019, URL:https://europepmc.org/articles/PMC6858365.

[40] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2332–2341. URL:https://www.arxiv.org/abs/1903.07291.

[41] M. Abadi et al., "Tensorflow: A system for large-scale machine learning," in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp. 265–283. [Online]. Available: URL:https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

[42] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in Proceedings of the 34th International Conference on Machine Learning, Aug 2017, pp. 214–223, URL: https://www.arxiv.org/abs/1701.07875.
[43] R. Zhang, L. Zhao, W. Lou, J.M. Abrigo, V.C.T. Mok, W.C.W. Chu, D. Wang, L. Shi, Automatic segmentation of acute ischemic stroke from dwi using 3-d fully convolutional densenets, IEEE Transactions on Medical Imaging 37 (9) (2018) 2149–2160.
[44] F. Isensee, P.F. Jaeger, S. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature methods 18 (2) (2020) 203–211.
[45] GitHub, "MedPy," URL:https://github.com/loli/medpy/.
[46] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, Eds. Curran Associates Inc, 2012, pp. 1097–1105. [Online]. Available: URL:http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.
[47] L.S. Hesse, G. Kuling, M. Veta, and A.L. Martel, "Intensity augmentation for domain transfer of whole breast segmentation in mri," 2019.
[48] B. Billot, D. Greve, K.V. Leemput, B. Fischl, J.E. Iglesias, and A.V. Dalca, "A learning strategy for contrast-agnostic mri segmentation," 2020.
[49] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," arXiv e-prints, p. arXiv:2102.09508, Feb. 2021.
[50] T. Song, "Generative model-based ischemic stroke lesion segmentation," 2019.
[51] N.K. Subbanna, D. Rajashekar, B. Cheng, G. Thomalla, J. Fiehler, T. Arbel, and N.D. Forkert, "Stroke lesion segmentation in flair mri datasets using customized markov random fields," Frontiers in Neurology, vol. 10, p. 541, 2019. [Online]. Available: URL:https://www.frontiersin.org/article/10.3389/fneur.2019.00541.
[52] N. Tomita, S. Jiang, M.E. Maeder, and S. Hassanpour, "Automatic post-stroke lesion segmentation on mr images using 3d residual convolutional neural network," NeuroImage: Clinical, vol. 27, p. 102276, 2020. [Online]. Available: URL:https://www.sciencedirect.com/science/article/pii/S2213158220301133.
[53] S. Thrun, Lifelong Learning Algorithms. Boston, MA: Springer, US, 1998, pp. 181–209. [Online]. Available: URL:https://doi.org/10.1007/978-1-4615-5529-2_8.
[54] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2010) 1345–1359.
[55] N. Karani, K. Chaitanya, C. Baumgartner, E. Konukoglu, A lifelong learning approach to brain mr segmentation across scanners and protocols, in: A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham, 2018, pp. 476–484.
[56] D. Ulyanov, A. Vedaldi, and V.S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," arXiv e-prints, 2016, http://arxiv.org/abs/1607.08022.
[57] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A.C. Courville, "Improved training of wasserstein gans," arXiv e-prints, 2017, URL:http://arxiv.org/abs/1704.00028,.