QUEST for MEDISYN: Quasi-norm based Uncertainty ESTimation for MEDical Image SYNthesis

Uddeshya Upadhyay [*1] **Viswanath P. Sudarshan** [*1 2] **Suyash P. Awate** [1]

Abstract

Uncertainty quantification in medical imaging is critical for clinical translation of deep learning-based methods. Modality propagation within the context of medical imaging is a problem of interest, both across as well as within modalities. For magnetic resonance imaging (MRI), often, multi-contrast MRI images are acquired for improved diagnosis and prognosis. In this work, we focus on the synthesis of T2w MRI images from T1w MRI images. Prior works used generative adversarial networks (GANs), but lack (i) uncertainty quantification, (ii) evaluating the robustness of the network to out-of-distribution data (common in medical imaging). We propose a robust GAN framework that incorporates uncertainty quantification using quasi-norm based penalties, and also show the efficacy of the method on unseen systemic and physiological perturbations on a large publicly available multimodal MRI dataset.

1. Introduction

Medical image synthesis refers to the task of mapping a given source modality to a target modality. The task of automated generation of target modality from a source, greatly benefits applications such as super-resolution (Jog et al., 2014) and quality enhancement and reconstruction (Ye et al., 2013; Ralph & Matthias, 2015; Upadhyay & Awate, 2019; Sudarshan et al., 2020), with the aim to improve patient throughput (lesser number of scans). Magnetic resonance imaging (MRI) is a widely used non-ionizing, non-invasive, *in vivo* medical imaging modality which provides both structural and functional images at resolutions of < 1 mm. Depending on the acquisition protocol, it is possible to generate multiple contrasts in MRI. Routine clinical protocols acquire multiple contrasts of MR images, e.g. T1 and T2 weighted (T1w and T2w respectively) images, for improved diagnostics. However, multicontrast acquisitions increase scan times and reduce patient throughput. We can consider the multicontrast images as random variables that are not entirely independent of each other as they originate from the same underlying anatomy. Hence, it is desirable to reduce the number of scans required per patient by exploiting such statistical dependencies.

Related Work. Prior works using conditional generative adversarial networks (GANs) and its variants have shown synthesis of realistic medical images in the general context of cross-modality medical image synthesis (Nie et al., 2018) and multicontrast image synthesis in MRI (Dar et al., 2019). The work (Lee & Fujita, 2020) provides a detailed review of deep learning (DL) based medical image synthesis.

However, these prior works do not discuss the performance of their networks to perturbations in the test data. In the context of medical imaging, perturbations could be systemic (scanner-related) and/or physiological. Hence, beyond reliable image synthesis during inference, quantification of uncertainty in the predicted images is crucial for clinical translation of synthesis frameworks (Wang et al., 2019; Nair et al., 2020). Quasi-norm based loss functions for DL make the network robust to anomalies (Shah et al., 2018; Upadhyay & Awate, 2019). However, these networks choose the quasi-norm (q in l_q norm) based on empirical analysis. This work proposes to automatically learn the parameter q.

We propose a novel quasi-norm based formulation for quantification of uncertainty to make our network robust to perturbations in test data. Results on image-noise related perturbations to input data at test time show that our model produces realistic images even at a reduced signal to noise ratio (SNR) compared to state-of-the-art methods. Additionally, we demonstrate that our network provides a reliable uncertainty map that can potentially act as a proxy for residual maps during inference. We analyze the uncertainty maps in an out-of-distribution scenario, where a synthetic lesion was simulated in the input test data. Our results show that such uncertainty maps can indicate potential risks in the predicted image for an incoming test data, given a fixed model and training data (Kendall & Gal, 2017).

[*]Equal contribution [1]Department of Computer Science and Engineering, Indian Institute of Technology-Bombay, India [2]Department of Electrical and Computing Systems, Monash University, Australia. Correspondence to: Uddeshya Upadhyay <uddeshya.upa@gmail.com>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

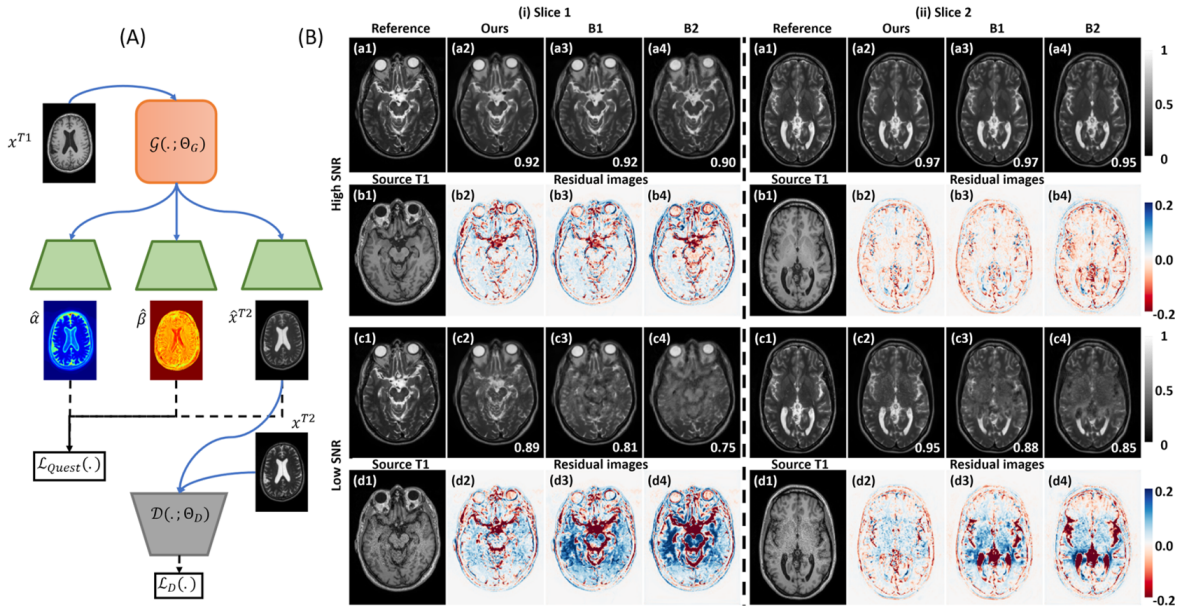

Figure 1. **(A) Schematic of the framework. (B) Performance of all methods at high SNR (NL0, same as training set) and low SNR (NL4, highest unseen noise-level) for two slices: (i) and (ii). (a1)–(b1) and (c1)–(d1)** T2w and T1w MRI images at high and low SNR respectively. **(a2)–(a4) and (c2)–(c4)** Predicted images using all methods (Ours, B1, and B2) at high and low SNR respectively. **(b2)–(b4) and (d2)–(d4)** Residual images for all methods at high and low SNR respectively. **(a1, c1)** are repeated for readers' convenience.

2. Methods

Let the image pair $(X^{\text{T1}}, X^{\text{T2}})$ model the pair of a randomly selected T1w MRI (X^{T1}) and its corresponding T2w MRI image (X^{T2}). Let x^{T1} and x^{T2} represent an instance of the T1w (source) and T2w (target) MRI image respectively. We learn the mapping from source to target modality using T pairs of co-registered training data: $\mathcal{X} := \{(x_t^{\text{T1}}, x_t^{\text{T2}})\}_{t=1}^{T}$.

2.1. Datasets and Models

We use the set of T1w and T2w MRI images available as part of IXI dataset (https://brain-development.org/ixi-dataset/), that consists of multicontrast MR images from several healthy subjects, collected across three different scanners. In this work, we use T1w and T2w contrast images which were co-registered for every subject using ANTS (Klein et al., 2009). The resolution of the MR images is $\sim 1\text{mm}^3$ isotropic. We use training, validation, and test split of 250, 50, and 100 non-overlapping subjects respectively. In this work, we use axial slices from the mid-brain region (~ 70 slices per subject). This work, inspired by conditional GANs, proposes a novel GAN-based framework which (i) employs a learned quasi-norm loss, and (ii) estimates uncertainty maps, as described below. Early works on image synthesis, modeled as a regression problem, do not account for heteroscedastic uncertainty in their modeling. Recent works model the heteroscedastic uncertainty by assuming that the variance is data-dependent and the residual (between the predicted and reference images) follows an isotropic Gaussian probability distribution function (PDF). However, the isotropic Gaussian PDF does not account for anomalies in data (outliers) that may be captured by heavy-tailed PDFs.

Let $\mathcal{G}(\cdot;\theta_G)$ and $\mathcal{D}(\cdot;\theta_D)$ represent our generator and the discriminator respectively. The input to the generator is $x^{\text{T1}} \sim X^{\text{T1}}$, the predicted image is $\hat{x}^{\text{T2}} := \mathcal{G}(x^{\text{T1}};\theta_G)$, and the ground-truth is $x^{\text{T2}} \sim X^{\text{T2}}$. Each image consists of K pixels. We denote j^{th} pixel in i^{th} image, say y, as y_{ij}.

We improve upon the assumptions made by prior works by modeling the residuals to follow a Generalized Gaussian distribution (GGD) with 0 mean, which is, $\frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} e^{-\left(\frac{|\epsilon-0|}{\alpha}\right)^{\beta}}$

Therefore, for the i^{th} image, the residual at pixel location j, ϵ_{ij} (between the predicted value $\hat{x}_{ij}^{\text{T2}}$ and ground-truth value x_{ij}^{T2}), follows GGD. In other words, $\hat{x}_{ij}^{\text{T2}} := x_{ij}^{\text{T2}} + \epsilon_{ij}$, that is, $\hat{x}_{ij}^{\text{T2}} \sim \frac{\beta_{ij}}{2\alpha_{ij}\Gamma(\frac{1}{\beta_{ij}})} e^{-\left(\frac{|\mathcal{G}(x^{\text{T1}};\theta_G)_{ij} - x_{ij}^{\text{T2}}|}{\alpha_{ij}}\right)^{\beta_{ij}}}$, which is capable of modelling heavy-tailed distributions including the Gaussian and Laplace PDFs ($\beta = 2$ and $\beta = 1$, respectively). Here $\alpha_j > 0$ is the scale parameter, $\beta_j > 0$ denotes the shape parameter, and $\Gamma(\cdot)$ is the standard gamma function. In our formulation all the ϵ_j's are independent but not necessarily identically distributed as α_j and β_j may vary spatially. Hence, the likelihood is,

$$P(\mathcal{X}|\Theta) := \prod_{i=1,j=1}^{i=T,j=K} \frac{\beta_{ij}}{2\alpha_{ij}\Gamma(\frac{1}{\beta_{ij}})} e^{\left(\frac{-|\mathcal{G}(x^{\text{T1}};\theta_G)_{ji} - x_{ij}^{\text{T2}}|}{\alpha_{ij}}\right)^{\beta_{ij}}} \tag{1}$$

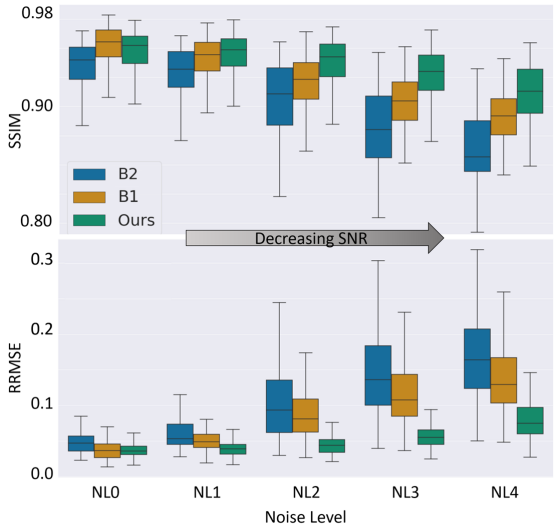

Figure 2. **SSIM and RRMSE values for all methods at all noise levels (NL0 to NL4).** At each NL, 70 mid brain slices from each subject (total 100) were evaluated (i.e., 7000 slices).

Therefore, log-likelihood is,

$$\log P(\mathcal{X}|\Theta) = \sum_{i=1,j=1}^{i=T,j=K} -\left(\frac{|\hat{x}_{ij}^{\mathrm{T2}} - x_{ij}^{\mathrm{T2}}|}{\alpha_{ij}}\right)^{\beta_{ij}} + \log\frac{\beta_{ij}}{2\alpha_{ij}} - \log\Gamma(\frac{1}{\beta_{ij}}) \quad (2)$$

where Θ represents the collection of network parameters. In this way, to improve robustness of the network, we predict (i) $\hat{x}^{\mathrm{T2}}$, (ii) $\hat{\alpha}$, and (iii) $\hat{\beta}$. Hence, the proposed robust quasi-norm based loss is

$$\mathcal{L}_U(\hat{X}^{\mathrm{T2}}, \{\hat{\alpha}_i\}_{i=1}^{i=T}, \{\hat{\beta}_i\}_{i=1}^{i=T}, X^{\mathrm{T2}}) := -\log P(\mathcal{X}|\Theta) \quad (3)$$

The adversarial term depending upon $\mathcal{D}(\cdot;\theta_D)$ is defined in terms of binary cross-entropy (for true and predicted probability vectors, say Y, and $\hat{Y}$ respectively) as,

$$\mathcal{L}_{CE}(\hat{Y}, Y) = -\sum_i [Y_i \log(\hat{Y}_i) + (1 - Y_i)\log(1 - \hat{Y}_i)] \quad (4)$$

Hence, $\mathcal{G}(\cdot;\theta_G)$ minimizes the following loss:

$$\mathcal{L}_{Quest}(\hat{X}^{\mathrm{T2}}, \{\hat{\alpha}_i\}_{i=1}^{i=T}, \{\hat{\beta}_i\}_{i=1}^{i=T}, X^{\mathrm{T2}}) := \mathcal{L}_U(\hat{X}^{\mathrm{T2}}, \{\hat{\alpha}_i\}_{i=1}^{i=T}, \{\hat{\beta}_i\}_{i=1}^{i=T}, X^{\mathrm{T2}}) + \lambda\mathcal{L}_{CE}(\mathcal{D}(\hat{X}^{\mathrm{T2}};\theta_D), 1) \quad (5)$$

On the other hand, $\mathcal{D}(\cdot;\theta_D)$ minimizes,

$$\mathcal{L}_D(\hat{X}^{\mathrm{T2}}, X^{\mathrm{T2}}) := \mathcal{L}_{CE}(\mathcal{D}(X^{\mathrm{T2}};\theta_D), 1) + \mathcal{L}_{CE}(\mathcal{D}(\hat{X}^{\mathrm{T2}};\theta_D), 0) \quad (6)$$

The proposed network is trained using the strategy in (Nie et al., 2018; Goodfellow, 2016).

2.2. Training and Testing Scheme

All the networks used in this work were trained using Adam optimizer (Kingma & Ba, 2014) by sampling mini-batches of size 16. The initial learning rate was set to $2e^{-4}$ and cosine annealing was used to decay the learning rate with epochs. We used a U-net (Ronneberger et al., 2015) based generator for which λ (Equation 5) was set to $7e^{-4}$. For numerical stability, the proposed network produces $\frac{1}{\hat{\alpha}}$ instead of $\hat{\alpha}$. The positivity constraint on the output is enforced by applying the ReLU activation function at the end of the three output layers in the network (Figure 1-(A)). The aleatoric uncertainty is defined as $\sigma^2_{aleatoric} = \frac{\hat{\alpha}^2\Gamma(3/\hat{\beta})}{\Gamma(1/\hat{\beta})}$ and the epistemic ($\sigma_{epistemic}$) uncertainty is obtained by multiple (50) forward passes (Gal & Ghahramani, 2016).

2.3. Comparison and Evaluation Metrics

We evaluate our method against two other baselines: (i) GAN-based framework proposed in (Nie et al., 2018) (current state of the art; say B1), and (ii) the corresponding generator network without the discriminator (say B2). We use relative root mean squared errors (RRMSE) and structural similarity index (SSIM) to compare the performance of different networks. RRMSE is defined as $RRMSE(a,b) = \|a-b\|_F/\|a\|_F$, where $\|\cdot\|_F$ indicates Frobenius norm. We use the standard SSIM as in (Wang et al., 2004). We evaluated the robustness of all the networks to perturbations in the input T1w MRI images by (i) adding multiple levels of i.i.d Gaussian perturbations in a particular region of interest (of the input image), and (ii) simulating synthetic lesions (described in Section 3). For (i), apart from the original test set, a total of four additional noise-levels(NLs) were simulated. The simulated images at different NLs (NL0 to NL4) show a deviation of (0, 4, 6, 8, and 10) % respectively in terms of RRMSE with respect to the reference image. Both (i) and (ii) cater to medically relevant out-of-distribution scenarios.

3. Results and Discussion

Figure 1 shows the predicted and residual images using the proposed method and two other baselines (B1 and B2) at high SNR (same as the training set) and low SNR (NL4, highest unseen noisy test data), for two representative slices ((i) and (ii)). The RRMSE between the source image at NL0 and NL4 (Figure 1 (b1) and (d1)) is 10%.

At NL0 (high SNR), the predicted image using our method (Figure 1 (a2)) is closer in structure and contrast to the reference image (Figure 1 (a1)) and shows residuals with least magnitude (Figure 1 (b2)) compared to that of B1 and B2 (Figure 1 (b3) and (b4)). For B1, the predicted image (Figure 1 (a3)) as well as the residual image (Figure 1 (b3)) are comparable to the proposed method. The predicted image using B2 (Figure 1 (a4)) is blurred than that of B1 and the proposed method.

At NL4 (low SNR), the output of our method (Figure 1 (c2)) is still very close to the reference image (Figure 1 (a1)) with-

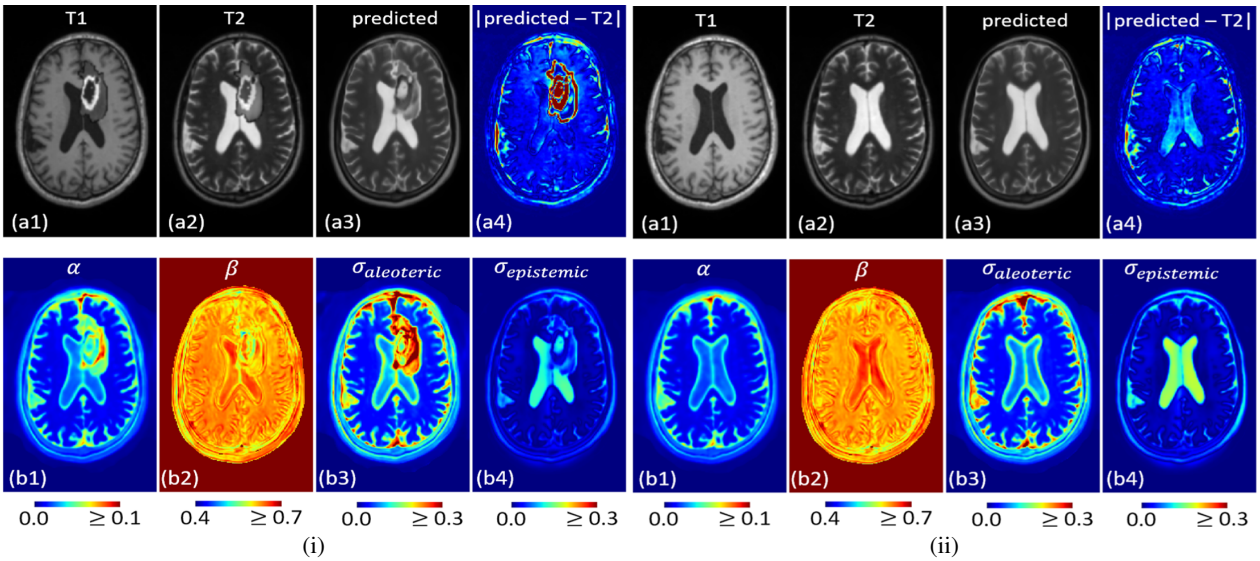

Figure 3. **Uncertainty quantification on out-of-distribution data (synthetic lesion added to test data). Subfigures (i) and (ii)** show results from two representative slices. **(a3)-(a4)** Predicted image ($\hat{x}^{T2}$) and absolute error map. **(b1)-(b2)** The learned scaling ($\hat{\alpha}$) and shape ($\hat{\beta}$) parameter maps. **(b3)-(b4)** Aleatoric ($\sigma_{aleatoric}$) and epistemic ($\sigma_{epistemic}$) uncertainty maps.

out significant loss of contrast and other textural features. On the other hand, both B1 and B2 (Figure 1 (c3) and (c4)) suffer from severe blurring and loss of textural features. Our method shows residual images (Figure 1 (d2)) that are significantly better (lower magnitude) than the residual images of B1 and B2 (Figure 1 (d3) and (d4)). However, between B1 and B2, the residual images of B1 exhibit lower magnitudes than that of B2. Our method shows significantly higher SSIM scores (0.89 and 0.95 for the two slices) compared to B1 (0.81 and 0.88) and B2 (0.75 and 0.85) at NL4.

Figure 2 shows box-plots of RRMSE and SSIM values for the predicted images from all the models, at different noise-levels (NL0 to NL4). At each NL, we evaluated 70 mid-brain slices of each of the 100 subjects (i.e., 7000 slices). As noise increases, the proposed method shows substantially improved SSIM and RRMSE values compared to B1 and B2. At NL0, B1 is comparable (slightly better) to our model both in terms of SSIM and RRMSE. Nevertheless, our model demonstrates robustness to the simulated perturbations even at high noise levels (SSIM $> 90\%$). In all the noise-levels B2 shows sub-optimal performance compared to both the GAN-based models. Moreover, our method shows very little variation in terms of RRMSE at all noise-levels compared to the other two methods.

Uncertainty quantification. We study the importance of providing uncertainty maps under scenarios where the test image has a lesion. Having trained the network on scans from healthy individuals, we evaluated the performance of our model, that predicts the mean, and uncertainty maps ($\hat{\sigma_{aleatoric}}$ and $\hat{\sigma_{epistemic}}$), on the test set with simulated lesions (unseen or out-of-distribution data). We simulated synthetic lesion in the images by using segmentation masks available from the BRATS Challenge 2020 (https://www.med.upenn.edu/cbica/brats2020/).

Figure 3 (a1) and (a2) correspond to input and the reference slice with synthetically added lesions. The outputs from our proposed network are: the mean, $\hat{\mathbf{x}}^{T2}$(Figure 3 (a3)), the scale of generalized Gaussian distribution parameter, $\hat{\alpha}$ (Figure 3 (b1)), and the corresponding shape parameter, $\hat{\beta}$. (Figure 3 (b2)). Figure 3 (b3) and (b4) shows the aleatoric ($\sigma_{aleatoric}$) and epistemic ($\sigma_{epistemic}$) uncertainty maps as defined in 2.2. Figure 3 (a4) shows the absolute error between the prediction and the ground-truth image. We see that our $\sigma_{aleatoric}$ map agrees with the error map (Figure 3 (a4)). The uncertainty map shows peak values that are localized around the lesion. This is along the expected direction as the model was not trained on pathological cases. *which are not available due to the absence of ground-truth images*. Moreover, several textural features are evident in our uncertainty maps. Hence, during inference, our uncertainty maps provide a close approximation of the error maps.

4. Conclusion

In this work, we have proposed a Bayesian deep learning framework that estimates uncertainty using learned quasi-norms. In addition to substantial improvements in synthesized images over the state-of-the-art networks (which do not quantify uncertainty), we demonstrate the utility of our network by evaluating performance on two experiments with out-of-distribution test-data: (i) perturbations in the noise-level, and (ii) providing uncertainty maps that act as a surrogate for residual maps (not available during inference).

References

Dar, S. U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., and Çukur, T. Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE transactions on medical imaging*, 38(10):2375–2388, 2019.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Jog, A., Carass, A., and Prince, J. Improving magnetic resonance resolution with supervised learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2014)*, pp. 987–990, 2014.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pp. 5574–5584, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Klein, A., Andersson, J., Ardekani, B., Ashburner, J., Avants, B., Chiang, M., Christensen, G., Collins, D., Gee, J., Hellier, P., et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, 46(3):786–802, 2009.

Lee, G. and Fujita, H. *Deep learning in medical image analysis : challenges and applications*. Advances in experimental medicine and biology ; Volume 1213. 1st ed. 2020. edition, 2020. ISBN 3-030-33128-8.

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q., and Shen, D. Medical image synthesis with deep convolutional adversarial networks. *IEEE Transactions on Biomedical Engineering*, 65(12):2720–2730, 2018.

Ralph, W. and Matthias, N. Advancing biomedical imaging. *Proceedings of the National Academy of Sciences*, 112 (47):14424–14428, 2015. ISSN 0027-8424.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Shah, M. P., Merchant, S., and Awate, S. P. Abnormality detection using deep neural networks with robust quasi-norm autoencoding and semi-supervised learning. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 568–572. IEEE, 2018.

Sudarshan, V. P., Egan, G. F., Chen, Z., and Awate, S. P. Joint pet-mri image reconstruction using a patch-based joint-dictionary prior. *Medical Image Analysis*, pp. 101669, 2020.

Upadhyay, U. and Awate, S. P. A mixed-supervision multilevel gan framework for image quality enhancement. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 556–564, Cham, 2019. Springer International Publishing.

Upadhyay, U. and Awate, S. P. Robust super-resolution gan, with manifold-based and perception loss. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1372–1376, 2019.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Imag. Proc.*, 13(4):600–12, 2004.

Ye, D. H., Zikic, D., Glocker, B., Criminisi, A., and Konukoglu, E. Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 606–613. Springer, 2013.