Uncertainty in multitask learning: joint representations for probabilistic MR-only radiotherapy planning

Felix J.S. Bragman[1], Ryutaro Tanno[1], Zach Eaton-Rosen[1],
Wenqi Li[1], David J. Hawkes[1], Sebastien Ourselin[2]
Daniel C. Alexander[1,3], Jamie R. McClelland[1] and M. Jorge Cardoso[2,1]

[1] Centre for Medical Image Computing, University College London, UK
[2] Biomedical Engineering and Imaging Sciences, King's College London, UK
[3] Clinical Imaging Research Centre, National University of Singapore, Singapore

Abstract. Multi-task neural network architectures provide a mechanism that jointly integrates information from distinct sources. It is ideal in the context of MR-only radiotherapy planning as it can jointly regress a synthetic CT (synCT) scan and segment organs-at-risk (OAR) from MRI. We propose a probabilistic multi-task network that estimates: 1) *intrinsic* uncertainty through a heteroscedastic noise model for spatially-adaptive task loss weighting and 2) *parameter* uncertainty through approximate Bayesian inference. This allows sampling of multiple segmentations and synCTs that share their network representation. We test our model on prostate cancer scans and show that it produces more accurate and consistent synCTs with a better estimation in the variance of the errors, state of the art results in OAR segmentation and a methodology for quality assurance in radiotherapy treatment planning.

1 Introduction

Radiotherapy treatment planning (RTP) requires a magnetic resonance (MR) scan to segment the target and organs-at-risk (OARs) with a registered computed tomography (CT) scan to inform the photon attenuation. MR-only RTP has recently been proposed to remove dependence on CT scans as cross-modality registration is error prone whilst extensive data acquisition is labourious. MR-only RTP involves the generation of a synthetic CT (synCT) scan from MRI. This synthesis process, when combined with manual regions of interest and safety margins provides a deterministic plan that is dependent on the quality of the inputs. Probabilistic planning systems conversely allow the implicit estimation of dose delivery uncertainty through a Monte Carlo sampling scheme. A system that can sample synCT and OAR segmentations would enable the development of a fully end-to-end uncertainty-aware probabilistic planning system.

Past methods for synCT generation and OAR segmentation stem from multi-atlas propagation [1]. Applications of convolutional neural networks (CNNs) to CT synthesis from MRI have recently become a topic of interest [2,3]. Conditional generative adversarial networks have been used to capture fine texture

details [2] whilst a CycleGAN has been exploited to leverage the abundance of unpaired training sets of CT and MR scans [3]. These methods however are fully deterministic. In a probabilistic setting, knowledge of the posterior over the network weights would enable sampling multiple realisations of the model for probabilistic planning whilst uncertainty in the predictions would be beneficial for quality control. Lastly, none of the above CNN methods segment OARs. If a model were trained in a multi-task setting, it would produce OAR segmentations and a synCT that are anatomically consistent, which is necessary for RTP.

Past approaches to multi-task learning have relied on uniform or hand-tuned weighting of task losses [4]. Recently, Kendall et al. [5] interpreted homoscedastic uncertainty as task-dependent weighting. However, homoscedastic uncertainty is constant in the task output and unrealistic for imaging data whilst yielding non-meaningful measures of uncertainty. Tanno et al. [6] and Kendall et al. [7] have raised the importance of modelling both *intrinsic* and *parameter* uncertainty to build more robust models for medical image analysis and computer vision. *Intrinsic* uncertainty captures uncertainty inherent in observations and can be interpreted as the irreducible variance that exists in the mapping of MR to CT intensities or in the segmentation process. *Parameter* uncertainty quantifies the degree of ambiguity in the model parameters given the observed data.

This paper makes use of [6] to enrich the multi-task method proposed in [5]. This enables modelling the spatial variation of *intrinsic* uncertainty via heteroscedastic noise across tasks and integrating *parameter* uncertainty via dropout [8]. We propose a probabilistic dual-task network, which operates on an MR image and simultaneously provides three valuable outputs necessary for probabilistic RTP: (1) synCT generation, (2) OAR segmentation and (3) quantification of predictive uncertainty in (1) and (2) (Fig.2). The architecture integrates the methods of uncertainty modelling in CNNs [6,7] into a multi-task learning framework with hard-parameter sharing, in which the initial layers of the network are shared across tasks and branch out into task-specific layers (Fig.1). Our probabilistic formulation not only provides an estimate of uncertainty over predictions from which one can stochastically sample the space of solutions, but also naturally confers a mechanism to spatially adapt the relative weighting of task losses on a voxel-wise basis.

2 Methods

We propose a probabilistic dual-task CNN algorithm which takes an MRI image, and simultaneously estimates the distribution over the corresponding CT image and the segmentation probability of the OARs. We use a heteroscedastic noise model and binary dropout to account for *intrinsic* and *parameter* uncertainty, respectively, and show that we obtain not only a measure of uncertainty over prediction, but also a mechanism for data-driven spatially adaptive weighting of task losses, which is integral in a multi-task setting. We employ a patch-based approach to perform both tasks, in which the input MR image is split into smaller overlapping patches that are processed independently. For each input patch $\mathbf{x}$,
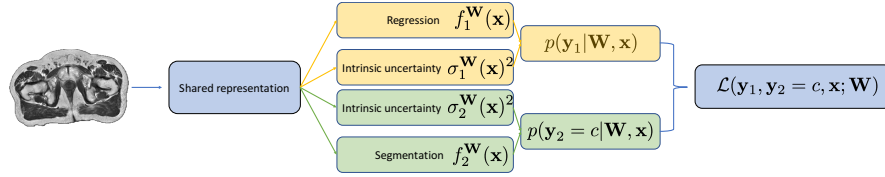
Fig. 1: Multi-task learning architecture. The predictive mean and variance $[f_i^{\mathbf{W}}(\mathbf{x}), \sigma_i^{\mathbf{W}}(\mathbf{x})^2]$ are estimated for the regression and segmentation. The task-specific likelihoods $p(\mathbf{y}_i|\mathbf{W}, \mathbf{x})$ are combined to yield the multi-task likelihood $p(\mathbf{y}_1, \mathbf{y}_2|\mathbf{W}, \mathbf{x})$.

our dual-task model estimates the conditional distributions $p(\mathbf{y}_i|\mathbf{x})$ for tasks $i = 1, 2$ where $\mathbf{y}_1$ and $\mathbf{y}_2$ are the Hounsfield Unit and OAR class probabilities. At inference, the probability maps over the synCT and OARs are obtained by stitching together outputs from appropriately shifted versions of the input patches.

Dual-task architecture. We perform multi-task learning with hard-parameter sharing [9]. The model shares the initial layers across the two tasks to learn an invariant feature space of the anatomy and branches out into four task-specific networks with separate parameters (Fig.1). There are two networks for each task (regression and segmentation). where one aims to performs CT synthesis (regression) or OAR segmentation, and the remaining models *intrinsic* uncertainty associated to the data and the task.

The rationale behind shared layers is to learn a joint representation between two tasks to regularise the learning of features for one task by using cues from the other. We used a high-resolution network architecture (HighResNet) [10] as the shared trunk of the model for its compactness and accuracy shown in brain parcellation. HighResNet is a fully convolutional architecture that utilises dilated convolutions with increasing dilation factors and residual connections to produce an end-to-end mapping from an input patch ($\mathbf{x}$) to voxel-wise predictions ($\mathbf{y}$). The final layer of the shared representation is split into two task-specific compartments (Fig. 1). Each compartment consists of two fully convolutional networks which operate on the output of representation network and together learn task-specific representation and define likelihood function $p(\mathbf{y}_i|\mathbf{W}, \mathbf{x})$ for each task $i = 1, 2$ where $\mathbf{W}$ denotes the set of all parameters of the model.

Task weighting with heteroscedastic uncertainty. Previous probabilistic multitask methods in deep learning [5] assumed constant intrinsic uncertainty per task. In our context, this means that the inherent ambiguity present across synthesis or segmentation does not depend on the spatial locations within an image. This is a highly unrealistic assumption as these tasks can be more challenging on some anatomical structures (e.g. tissue boundaries) than others. To capture potential spatial variation in intrinsic uncertainty, we adapt the *heteroscedastic* (data-dependent) noise model to our multitask learning problem.

For the CT synthesis task, we define our likelihood as a normal distribution $p(\mathbf{y}_1|\mathbf{W}, \mathbf{x}) = \mathcal{N}(f_1^{\mathbf{W}}(\mathbf{x}), \sigma_1^{\mathbf{W}}(\mathbf{x})^2)$ where mean $f_1^{\mathbf{W}}(\mathbf{x})$ and variance $\sigma_1^{\mathbf{W}}(\mathbf{x})^2$ are

modelled by the regression output and uncertainty branch as functions of the input patch $\mathbf{x}$ (Fig.1). We define the task loss for CT synthesis to be the negative log-likelihood (NLL) $\mathcal{L}_1(\mathbf{y}_1, \mathbf{x}; \mathbf{W}) = \frac{1}{2\sigma_1^{\mathbf{W}}(\mathbf{x})^2}||\mathbf{y}_1 - f_1^{\mathbf{W}}(\mathbf{x})||^2 + \log\sigma_1^{\mathbf{W}}(\mathbf{x})^2$. This loss encourages assigning high-uncertainty to regions of high errors, enhancing the robustness of the network against noisy labels and outliers, which are prevalent at organ boundaries especially close to the bone.

For the segmentation, we define the classification likelihood as softmax function of scaled logits i.e. $p(\mathbf{y}_2|\mathbf{W}, \mathbf{x}) = \text{Softmax}\left(f_2^{\mathbf{W}}(\mathbf{x})/2\sigma_2^{\mathbf{W}}(\mathbf{x})^2\right)$ where the segmentation output $f_2^{\mathbf{W}}(\mathbf{x})$ is scaled by the uncertainty term $\sigma_2^{\mathbf{W}}(\mathbf{x})^2$ before softmax (Fig.1). As the uncertainty $\sigma_2^{\mathbf{W}}(\mathbf{x})^2$ increases, the Softmax output approaches a uniform distribution, which corresponds to the maximum entropy discrete distribution. We simplify the scaled Softmax likelihood by considering an approximation in [5], $\frac{1}{\sigma^2}\sum_{c'}\exp(\frac{1}{2\sigma_2^{\mathbf{W}}(\mathbf{x})^2}f_{2,c'}^{\mathbf{W}}(\mathbf{x})) \approx \left(\sum_{c'}\exp(f_{2,c'}^{\mathbf{W}}(\mathbf{x}))\right)^{1/2\sigma_2^{\mathbf{W}}(\mathbf{x})^2}$ where c' denotes a segmentation class. This yields the NLL task-loss of the form $\mathcal{L}_2(\mathbf{y}_2 = c, \mathbf{x}; \mathbf{W}) \approx \frac{1}{2\sigma_2^{\mathbf{W}}(\mathbf{x})^2}\text{CE}(f_2^{\mathbf{W}}(\mathbf{x}), \mathbf{y}_2 = c) + \log\sigma_2^{\mathbf{W}}(\mathbf{x})^2$, where CE denotes cross-entropy.

The joint likelihood factorises over tasks such that $p(\mathbf{y}_1, \mathbf{y}_2|\mathbf{W}, \mathbf{x}) = \prod_i^2 p(\mathbf{y}_i|\mathbf{W}, \mathbf{x})$. We can therefore derive the NLL loss for the dual-task model as

$$\mathcal{L}(\mathbf{y}_1, \mathbf{y}_2 = c, \mathbf{x}; \mathbf{W}) = \frac{||\mathbf{y}_1 - f_1^{\mathbf{W}}(\mathbf{x})||^2}{2\sigma_1^{\mathbf{W}}(\mathbf{x})^2} + \frac{\text{CE}(f_2^{\mathbf{W}}(\mathbf{x}), \mathbf{y}_2 = c)}{2\sigma_2^{\mathbf{W}}(\mathbf{x})^2} + \log\left(\sigma_1^{\mathbf{W}}(\mathbf{x})^2\sigma_2^{\mathbf{W}}(\mathbf{x})^2\right)$$

where both task losses are weighted by the inverse of heteroscedastic intrinsic uncertainty terms $\sigma_i^{\mathbf{W}}(\mathbf{x})^2$, that enables automatic weighting of task losses on a per-sample basis. The log-term controls the spread.

Parameter uncertainty with approximate Bayesian inference. In data-scarce situations, the choice of best parameters is ambiguous, and resorting to a single estimate without regularisation often leads to overfitting. Gal et al.[8] have shown that dropout improves the generalisation of a NN by accounting for *parameter* uncertainty through an approximation of the posterior distribution over its weights $q(\mathbf{W}) \approx p(\mathbf{W}|\mathbf{X}, \mathbf{Y_1}, \mathbf{Y_2})$ where $\mathbf{X} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$, $\mathbf{Y}_1 = \{\mathbf{y}_1^{(1)}, ..., \mathbf{y}_1^{(N)}\}$, $\mathbf{Y}_2 = \{\mathbf{y}_2^{(1)}, ..., \mathbf{y}_2^{(N)}\}$ denote the training data. We also use binary dropout in our model to assess the benefit of modelling parameter uncertainty in the context of our multitask learning problem.

During training, for each input (or minibatch), network weights are drawn from the approximate posterior $w' \sim q(\mathbf{W})$ to obtain the multi-task output $\mathbf{f}^{w'}(\mathbf{x}) := [f_1^{w'}(\mathbf{x}), f_2^{w'}(\mathbf{x}), \sigma_1^{w'}(\mathbf{x})^2, \sigma_2^{w'}(\mathbf{x})^2]$. At test time, for each input patch $\mathbf{x}$ in an MR scan, we collect output samples $\{\mathbf{f}^{w^{(t)}}(\mathbf{x})\}_{t=1}^{\text{T}}$ by performing T stochastic forward-passes with $\{w^{(t)}\}_{t=1}^{\text{T}} \sim q(\mathbf{W})$. For the regression, we calculate the expectation over the T samples in addition to the variance, which is the *parameter* uncertainty. For the segmentation, we compute the expectation of class probabilities to obtain the final labels whilst *parameter* uncertainty in the segmentation is obtained by considering variance of the stochastic class

probabilities on a class basis. The final predictive uncertainty is the sum of the *intrinsic* and *parameter* uncertainties.

Implementation details We implemented our model within the NiftyNet framework [11] in TensorFlow. We trained our model on randomly selected 152×152 patches from 2D axial slices and reconstructed the 3D volume at test time. The representation network was composed of a convolutional layer followed by 3 sets of twice repeated dilated convolutions with dilation factors $[1, 2, 4]$ and a final convolutional layer. Each layer (l) used a 3×3 kernel with features $f_R = [64, 64, 128, 256, 2048]$. Each task-specific branch was a set of 5 convolutional layers of size $[256_{l=1,2,3,4}, n_{i,l=5}]$ where $n_{i,l=5}$ is equal to 1 for regression and σ and equal to the number of segmentation classes. The first two layers were 3×3 kernels whilst the final convolutional layers were fully connected. A Bernouilli drop-out mask with probability $p = 0.5$ was applied on the final layer of the representation network. We minimised the loss using ADAM with a learning rate 10^{-3} and trained up to 19000 iterations with convergence of the loss starting at 17500. For the stochastic sampling, we performed model inference 10 times at iterations 18000 and 19000 leading to a set of $T = 20$ samples.

3 Experiments and Results

Data We validated on 15 prostate cancer patients, who each had a T2-weighted MR (3T, $1.46 \times 1.46 \times 5\text{mm}^3$) and CT scan (140kVp, $0.98 \times 0.98 \times 1.5\text{mm}^3$) acquired on the same day. Organ delineation was performed by a clinician with labels for the left and right femur head, bone, prostate, rectum and bladder. Images were resampled to isotropic resolution. The CT scans were spatially aligned with the T2 scans prior to training [1]. In the segmentation, we predicted labels for the background, left/right femur head, prostate, rectum and bladder.

Experiments We performed 3-fold cross-validation and report statistics over all hold-out sets. We considered the following models: 1) baseline networks for regression/segmentation (M1), 2) baseline network with drop-out (M2a), 3) the baseline with drop-out and heteroscedastic noise (M2b), 4) multi-task network using homoscedastic task weighting (M3) [5] and 5) multi-task network using task-specific heteroscedastic noise and drop-out (M4). The baseline networks used only the representation network with $^1/_2 f_R$ and a fully-connected layer for the final output to allow a fair comparison between single and multi-task networks. We also compared our results against the current state of the art in atlas propagation (AP) [1], which was validated on the same dataset.

Model performance An example of the model output is shown in Fig 2. We calculated the Mean Absolute Error (MAE) between the predicted and reference scans across the body and at each organ (Tab. 1). The fuzzy DICE score between the probabilistic segmentation and the reference was calculated for the
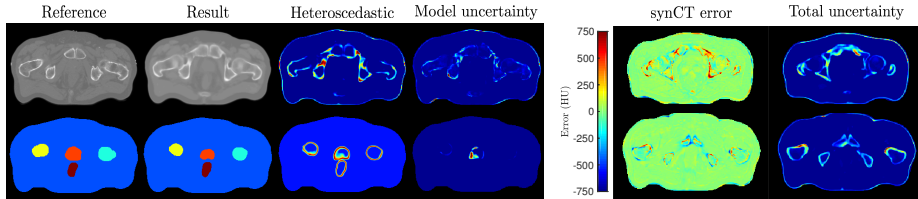
Fig. 2: Model output. *Intrinsic* and *parameter* uncertainty both correlate with regions of high contrast (bone in the regression, organ boundary for segmentation). Note the correlation between model error and the predicted uncertainty.

segmentation (Tab. 1). Best performance was in our presented method (M4) for the regression across all masks except at the bladder. Application of the multi-task heteroscedastic network with drop-out (M4) produced the most consistent synCT across all models with the lowest average MAE and the lowest variation across patients (43.3 ± 2.9 versus 45.7 ± 4.6 [1] and 44.3 ± 3.1 [5]). This was significant lower when compared to M1 ($p < 0.001$) and M2 ($p < 0.001$). This was also observed at the bone, prostate and bladder ($p < 0.001$). Whilst differences at $p < 0.05$ were not observed versus M2b and M3, the consistent lower MAE and standard deviation across patients in M4 demonstrates the added benefit of modelling heteroscedastic noise and the inductive transfer from the segmentation task. We performed better than the current state of the art in atlas propagation, which used both T1 and T2-weighted scans [1]. Despite equivalence with the state of the art (Tab. 1), we did not observe any significant differences between our model and the baselines despite an improvement in mean DICE at the prostate and rectum (0.70 ± 0.06 and 0.74 ± 0.12) versus the baseline M1

Table 1: Model comparison. Bold values indicate when a model was significantly worse than M4 $p < 0.05$. No data was available for significance testing with AP. M2b was statistically better $p < 0.05$ than M4 in the prostate segmentation.

| Models	All	Bone	L femur	R femur	Prostate	Rectum	Bladder
		Regression - synCT - Mean Absolute Error (HU)					
M1	**48.1(4.2)**	**131(14.0)**	78.6(19.2)	**80.1(19.6)**	**37.1(10.4)**	63.3(47.3)	**24.3(5.2)**
M2a	**47.4(3.0)**	**130(12.1)**	78.0(14.8)	77.0(13.0)	**36.5(7.8)**	67(44.6)	**24.1(7.5)**
M2b [7]	44.5(3.6)	128(17.1)	75.8(20.1)	74.2(17.4)	31.2(7.0)	56.1(45.5)	17.8(4.7)
M3 [5]	44.3(3.1)	126(14.4)	74.0(19.5)	73.7(17.1)	29.4(4.7)	58.4(48.0)	18.2(3.5)
AP [1]	45.7(4.6)	125(10.3)	-	-	-	-	-
M4 (ours)	43.3(2.9)	121(12.6)	69.7(13.7)	67.8(13.2)	28.9(2.9)	55.1(48.1)	18.3(6.1)
		Segmentation - OAR - Fuzzy DICE score					
M1	-	-	0.91(0.02)	0.90(0.04)	0.67(0.12)	0.70(0.15)	0.92(0.05)
M2a	-	-	0.85(0.03)	0.90(0.04)	0.66(0.12)	0.69(0.13)	0.90(0.07)
M2b [7]	-	-	0.92(0.02)	0.92(0.01)	0.77(0.07)	0.74(0.13)	0.92(0.03)
M3 [5]	-	-	0.92(0.02)	0.92(0.02)	0.73(0.07)	0.76(0.10)	0.93(0.02)
AP [1]	-	-	0.89(0.02)	0.90(0.01)	0.73(0.06)	0.77(0.06)	0.90(0.03)
M4 (ours)	-	-	0.91(0.02)	0.91(0.02)	0.70(0.06)	0.74(0.12)	0.93(0.04)
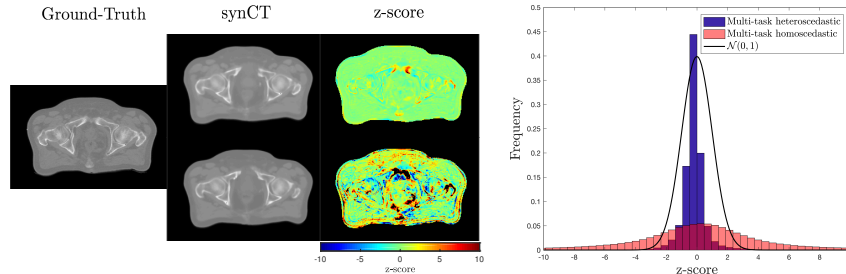
Fig. 3: Analysis of uncertainty estimation. a) synCTs and z-scores for the a subject between M4 (top) and M3 (bottom) models. b) z-score distribution of all patients (15) between both models.

$(0.67 \pm 0.12, 0.70 \pm 0.15)$. The *intrinsic uncertainty* (Fig. 2) models the uncertainty specific to the data and thus penalises regions of high error leading to an under-segmentation yet with higher confidence in the result.

Uncertainty estimation for radiotherapy We tested the ability of the proposed network to better predict associated uncertainties in the synCT error. To verify that we produce clinically viable samples for treatment planning, we quantified the distribution of regression z-scores for the multi-task heteroscedastic and homoscedastic models. In the former, the total predictive uncertainty is the sum of the *intrinsic* and *parameter* uncertainties. This leads to a better approximation of the variance in the model. In contrast, the total uncertainty in the latter reduces to the variance of the stochastic test-time samples. This is likely to lead to a miscalibrated variance. A χ^2 goodness of fit test was performed, showing that the homoscedastic z-score distribution is not normally distributed $(0.82 \pm 0.54, p < 0.01)$ in contrast to the heteroscedastic model $(0.04 \pm 0.84, p > 0.05)$. This is apparent in Fig.3 where there is greater confidence in the synCT produced by our model in contrast the homoscedastic case.

The predictive uncertainty can be exploited for quality assurance (Fig. 4). There may be issues whereupon time differences have caused variations in bladder and rectum filling across MR and CT scans causing patient variability in the training data. This is exemplified by large errors in the synCT at the rectum (Fig. 4) and quantified by large localised z-scores (Fig. 4g), which correlate strongly with the *intrinsic* and *parameter* uncertainty across tasks (Fig. 2 and 4).

4 Conclusions

We have proposed a probabilistic dual-network that combines uncertainty modelling with multi-task learning. Our network extends prior work in multi-task learning by integrating heteroscedastic uncertainty modelling to naturally weight task losses and maximize inductive transfer between tasks. We have demonstrated the applicability of our network in the context of MR-only radiotherapy treatment planning. The model simultaneously provides the generation of
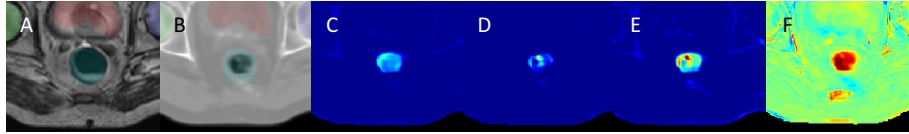
Fig. 4: Uncertainty in problematic areas. a) T2 with reference segmentation, b) synCT with localised error, c) *intrinsic* uncertainty, d) *parameter* uncertainty, e) total predictive uncertainty and g) error in HU (range [-750HU, 750HU]).

synCTs, the segmentation of OARs and quantification of predictive uncertainty in both tasks. We have shown that a multi-task framework with heteroscedastic noise modelling leads to more accurate and consistent synCTs with a constraint on anatomical consistency with the segmentations. Importantly, we have demonstrated that the output of our network leads to consistent anatomically correct stochastic synCT samples that can potentially be effective in treatment planning.

References

1. Burgos, N., et al.: Iterative framework for the joint segmentation and ct synthesis of mr images: application to mri-only radiotherapy treatment planning. Phys. Med. Biol. **62** (2017)
2. Nie, D., et al.: Medical image synthesis with context-aware generative adversarial networks. arXiv:1612.05362
3. Wolterink, J., et al.: Deep mr to ct synthesis using unpaired data. In: SASHIMI. (2017) 14–22
4. Moeskops, P., et al.: Deep learning for multi-task medical image segmentation in multiple modalities. In: MICCAI. (2016) 478–486
5. Kendall, A., et al.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR. (2018)
6. Tanno, R., et al.: Bayesian image quality transfer with cnns: Exploring uncertainty in dmri super-resolution. In: MICCAI. (2017) 611–619
7. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NIPS. (2017) 5580–5590
8. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML. (2016) 1050–1059
9. Caruana, R.: Multitask learning: A knowledge-based source of inductive bias. In: ICML. (1993)
10. Li, W., et al.: On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task. In: IPMI. (2017) 348–360
11. Gibson, E., et al.: Niftynet: a deep-learning platform for medical imaging. Comput. Methods Programs Biomed. **158** (2018)