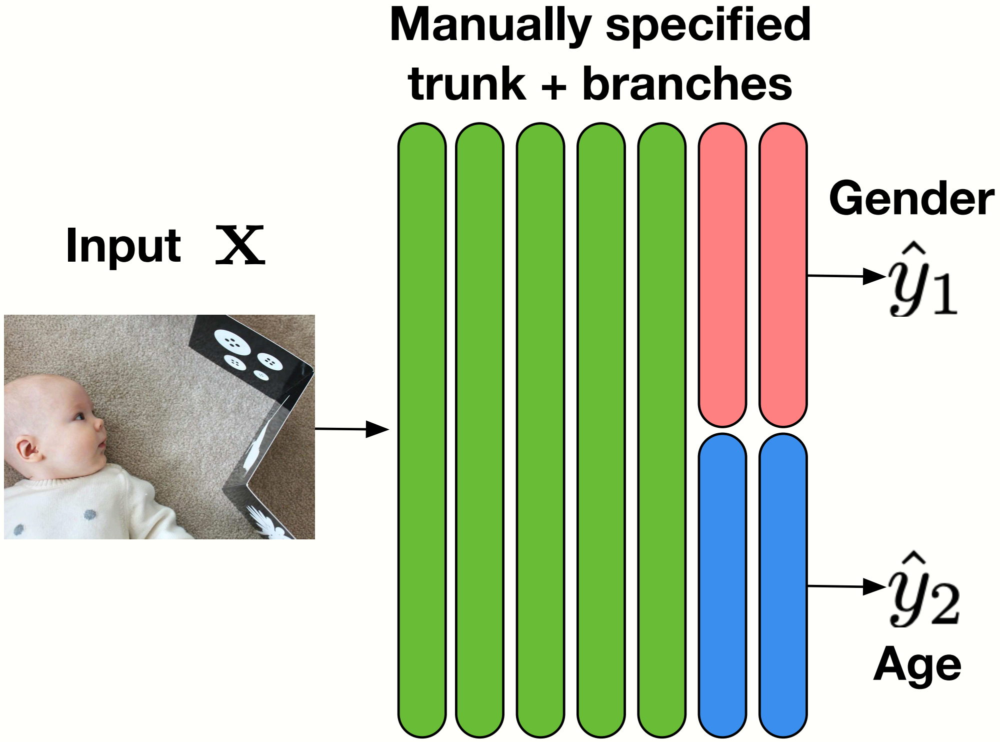
Stochastic Filter Groups for Multi-Task CNNs: Learning Specialist and Generalist Convolution Kernels

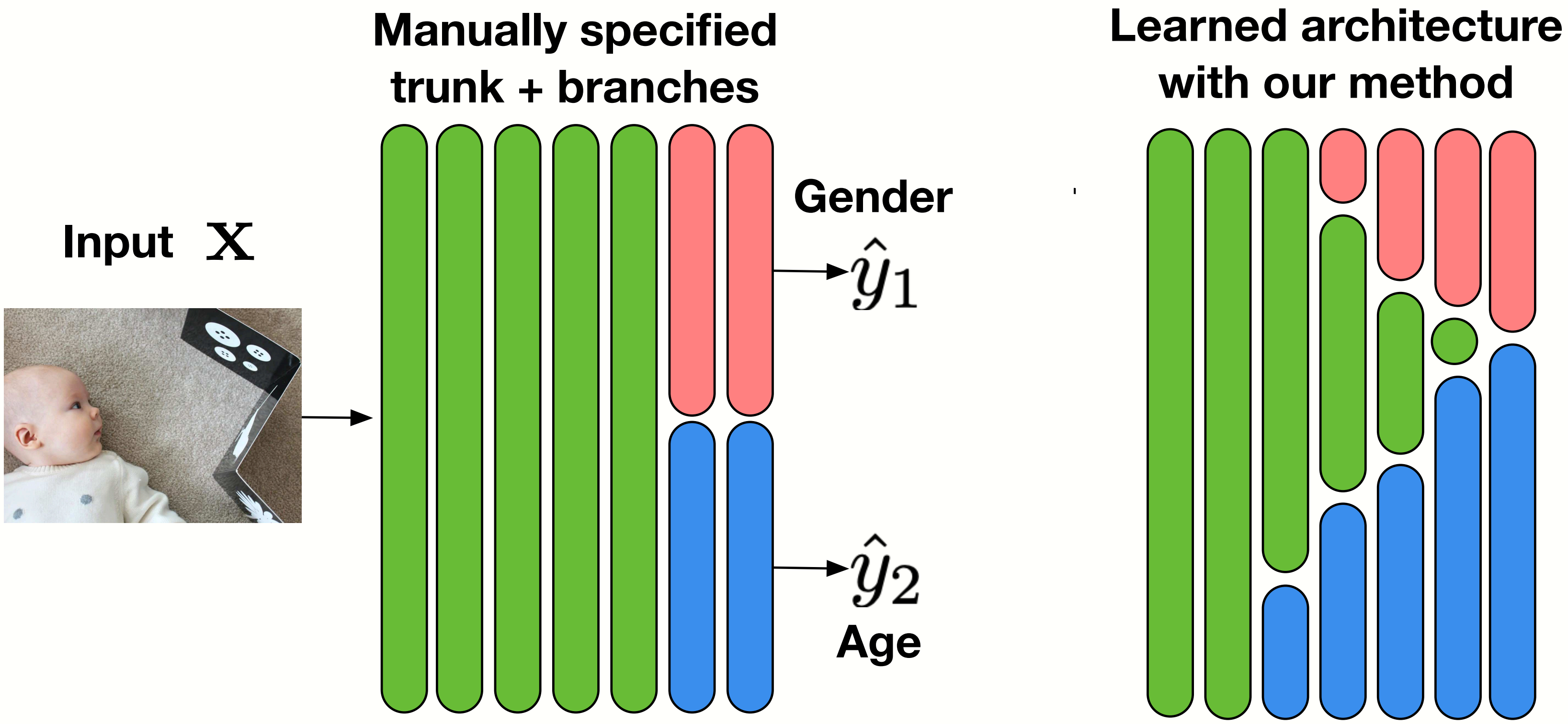
Felix J.S. Bragman*, Ryutaro Tanno*
Sébastien Ourselin, Daniel C. Alexander and M. Jorge Cardoso

Multi-task learning

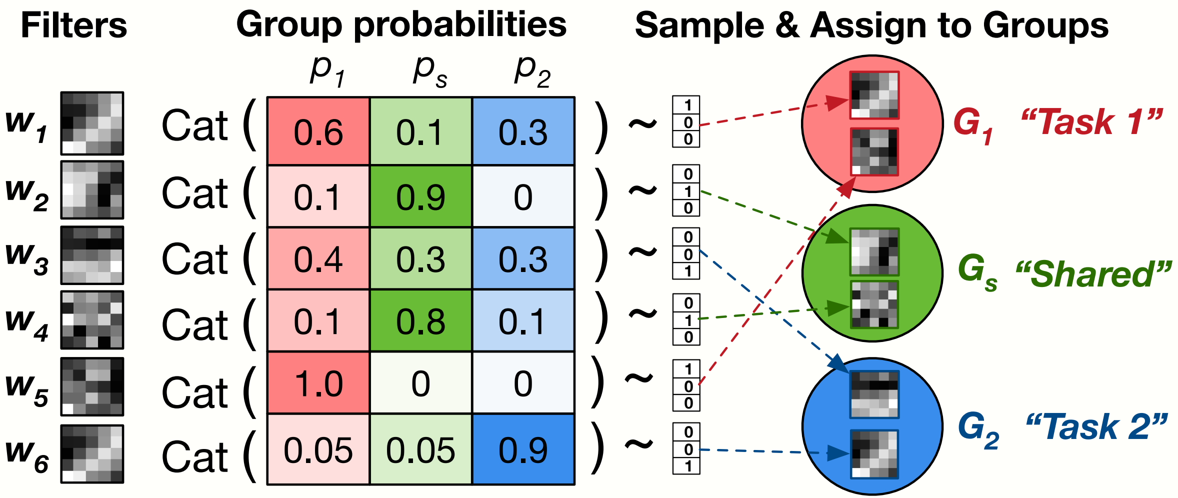


- The benefits of multi-task learning (MTL) depend on the structure of **feature sharing**
- Hand-crafted architecture with *a priori* knowledge on parameter sharing
- Number of sharing combinations combinatorial in layers and tasks
- Feature sharing mechanisms proposed: *Misra et al. 2016, Ruder et al. 2018, Meyerson et al. 2018 etc.*

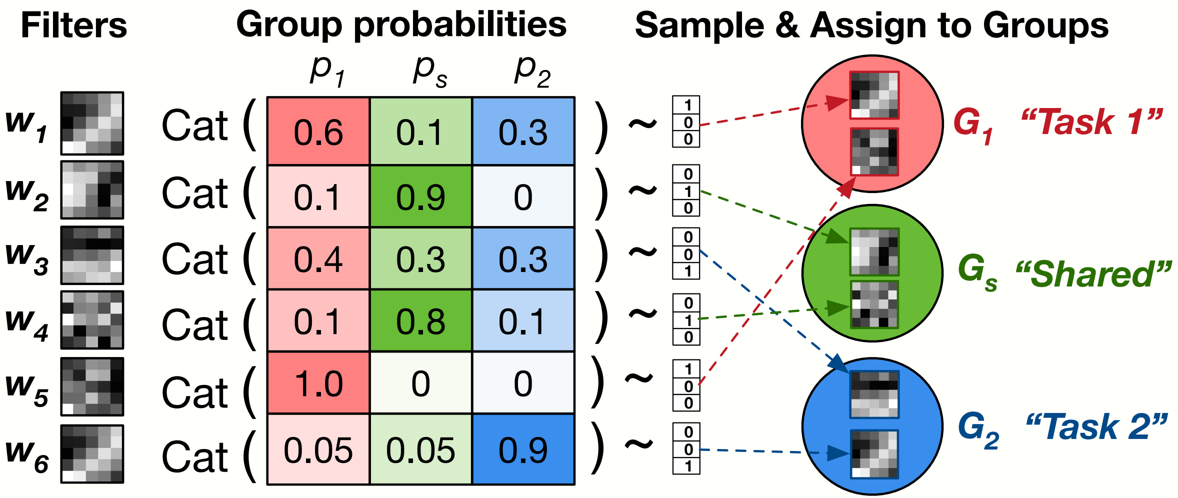
Main contribution



Stochastic Filter Groups (SFG)



Stochastic Filter Groups (SFG)



Group definitions

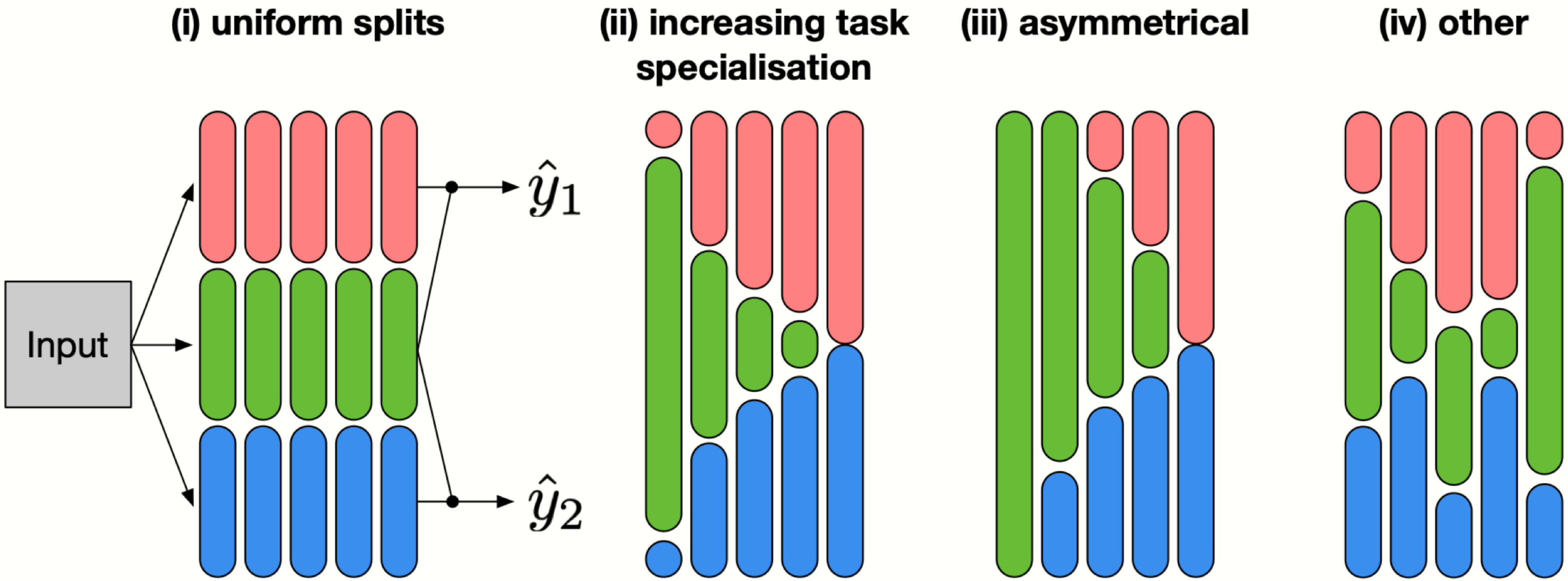
G_1 := filters updated w.r.t task-1 loss

G_S := filters updated w.r.t task-1 loss and task-2 loss

G_2 := filters updated w.r.t task-2 loss

Stochastic Filter Groups (SFG)

Possible grouping patterns



SFG optimisation method

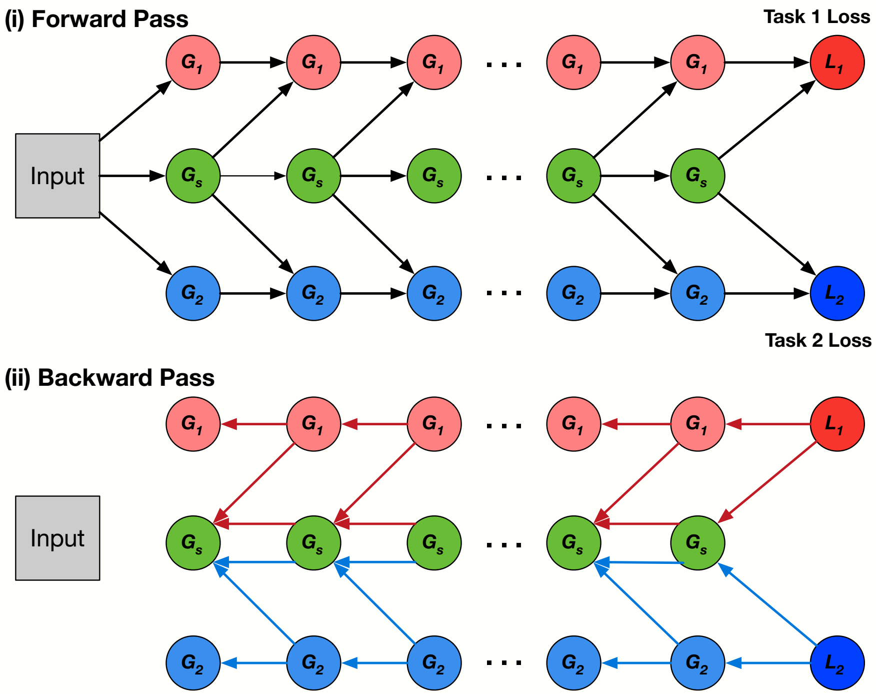
1. **Structured routing of features** to ensure desired flow of gradients
   - Multi-task extension of filter groups [Ioannou et al. 2016]

2. View filter assignment as ***T+1 way drop-out***
   - Cast learning of grouping probabilities and filter weights as variational inference [Gal et al. 2018]

3. Continuous relaxation using **Gumbel-Softmax** [Jang et al. 2016]
   - Learn categorical distribution over filter group assignments

SFG optimisation method

1. **Structured routing of features** to ensure desired flow of gradients
   - Multi-task extension of filter groups [Ioannou et al. 2016]

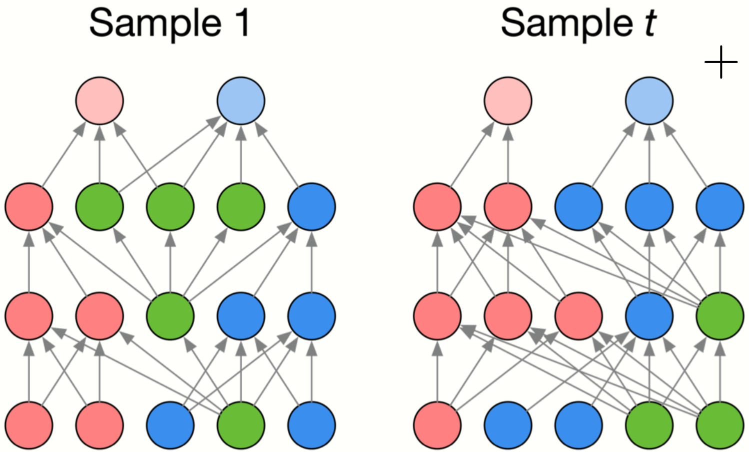


SFG optimisation method

2. View filter assignment as ***T+1 way drop-out***
   - Cast learning of grouping probabilities and filter weights as variational inference [Gal et al. 2018]

$$\mathcal{L}_{\mathrm{MC}}(\phi) = -\frac{N}{M}\sum_{i=1}^{M}\left[\log p\left(y_i^{(1)}|\mathbf{x}_i, \mathcal{W}_i\right) + \log p\left(y_i^{(2)}|\mathbf{x}_i, \mathcal{W}_i\right)\right]$$

$$+\lambda_1 \cdot \sum_{l=1}^{L}\sum_{k=1}^{K_l} ||\mathbf{M}^{(l),k}||^2 - \lambda_2 \cdot \sum_{l=1}^{L}\sum_{k=1}^{K_l} \mathcal{H}(\mathbf{p}^{(l),k})$$



SFG optimisation method

2. View filter assignment as ***T+1 way drop-out***
   - Cast learning of grouping probabilities and filter weights as variational inference [Gal et al. 2018]

$$\mathcal{L}_{\mathrm{MC}}(\phi) = -\frac{N}{M}\sum_{i=1}^{M}\Big[\log p\Big(y_i^{(1)}|\mathbf{x}_i, \mathcal{W}_i\Big) + \log p\Big(y_i^{(2)}|\mathbf{x}_i, \mathcal{W}_i\Big)\Big]$$

$$+ \lambda_1 \cdot \sum_{l=1}^{L}\sum_{k=1}^{K_l} ||\mathbf{M}^{(l),k}||^2 - \lambda_2 \cdot \sum_{l=1}^{L}\sum_{k=1}^{K_l} \mathcal{H}(\mathbf{p}^{(l),k})$$

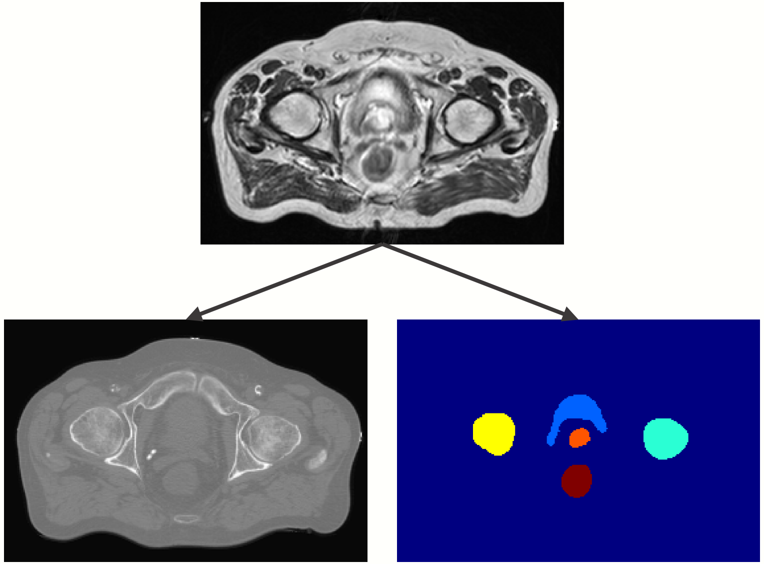
3. Continuous relaxation using **Gumbel-Softmax** [Jang et al. 2016]

$$z = \mathrm{Softmax}\left([g_i + \log p_i]/\tau\right) \quad g \sim \mathrm{Gumbel}(0,1)$$

Experiments - datasets

- **Age regression** and **gender prediction** from face images (UTKFace)
- Multi-task VGG11 [Simonyan et al. 2015] with SFG

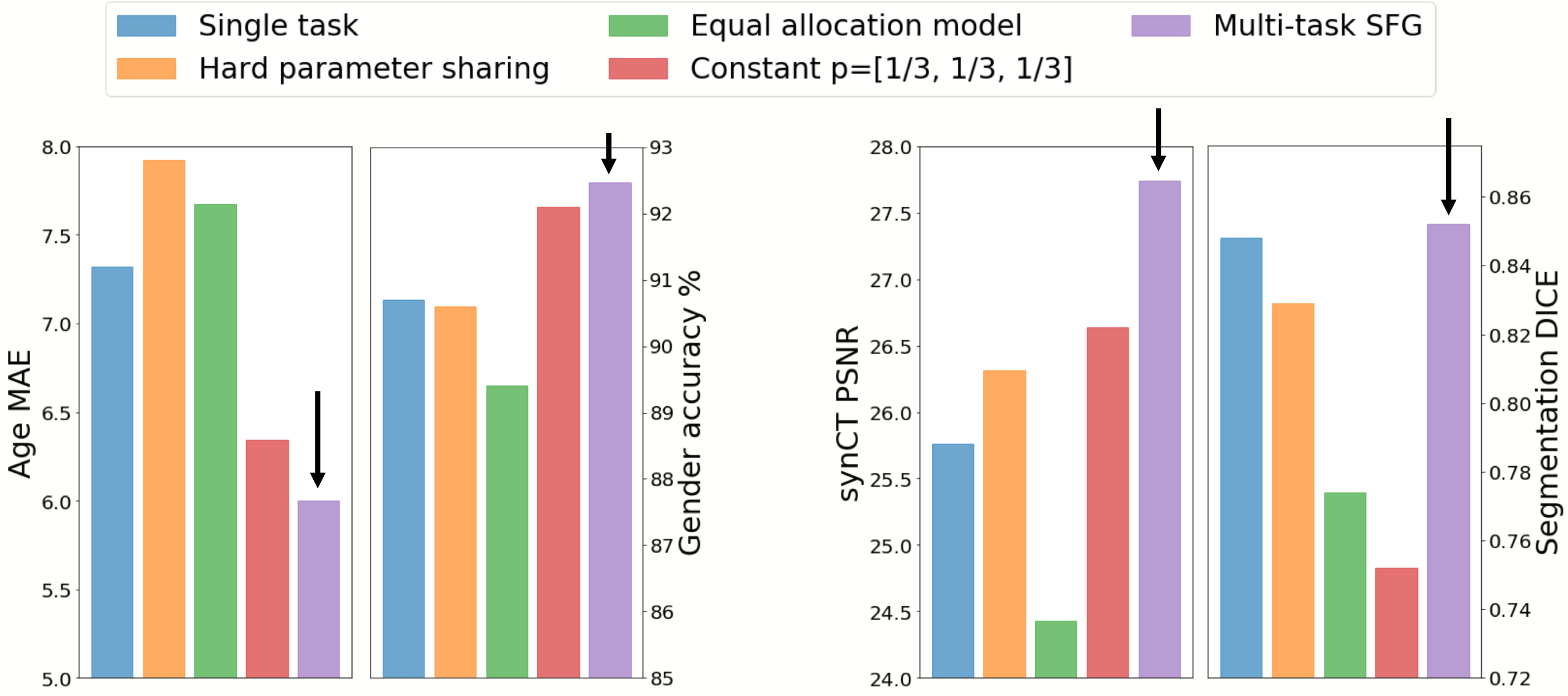
- **Organ segmentation** and **CT synthesis** from 3D prostate MRI scans
- Multi-task *High*ResNet [Li et al. 2018] with SFG



Results – SFGs improve multi-task performance

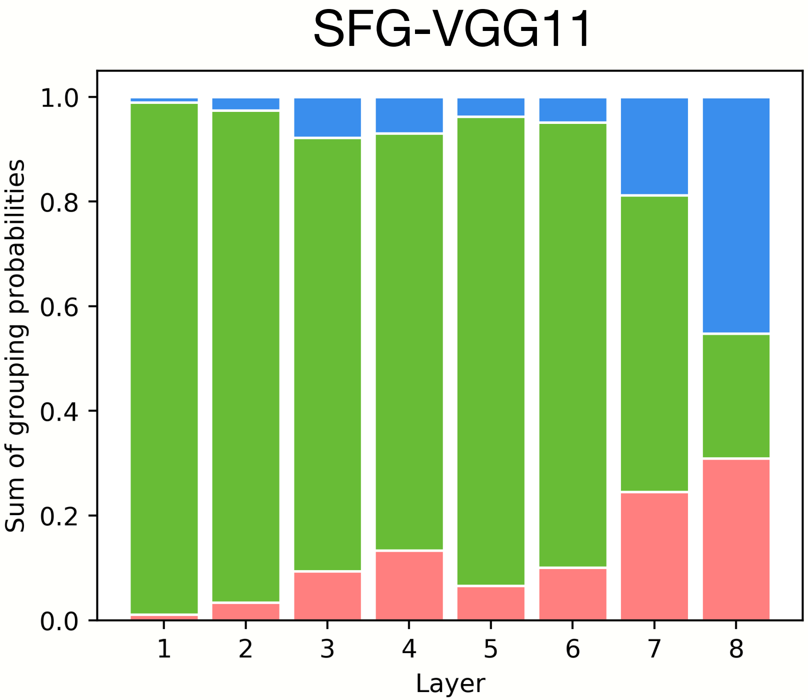
Age and gender prediction

Organ segmentation and CT synthesis

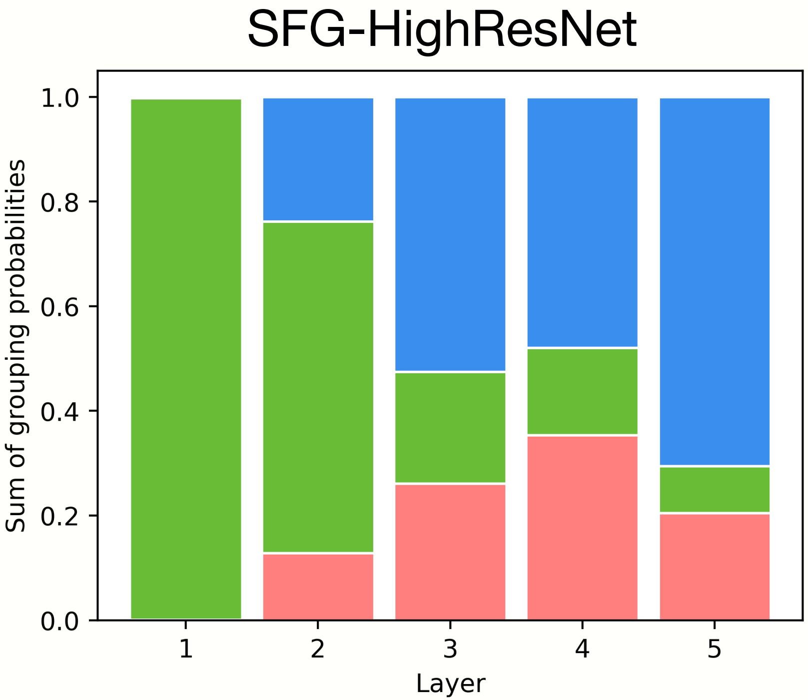


Results - filter group ratio across tasks

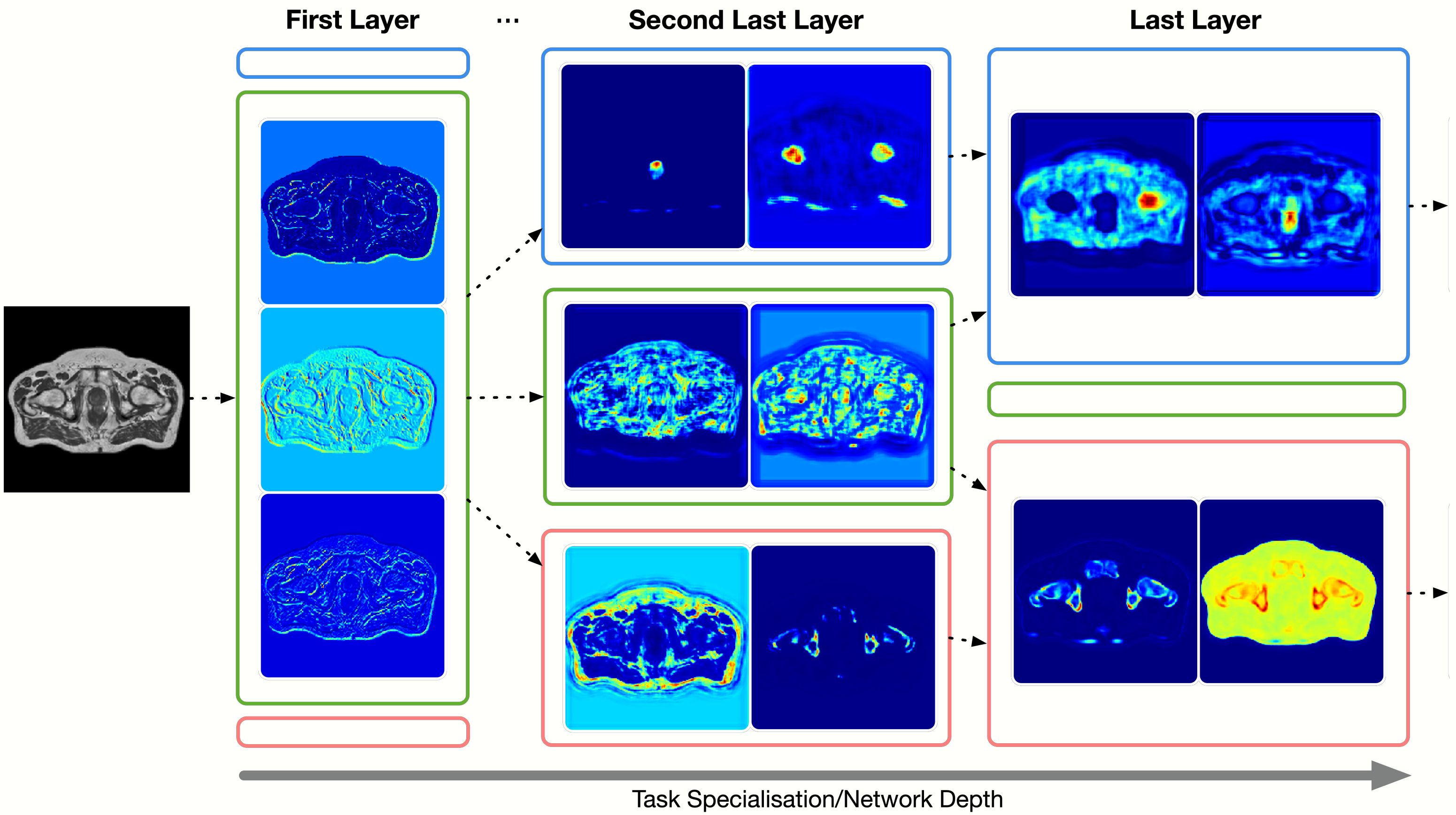
- Kernel allocation for age regression, shared and gender prediction



- Kernel allocation for CT synthesis, shared and organ segmentation



Results - visualising activations



Thank you!

Please visit poster #11 for more details and results