Intra-Domain Adaptation for Robust Visual Guidance in Intratympanic Injections*

Luke M. Shepherd[1], Sophia Bano[1], Joseph G. Manjaly[2], Lukas Lindenroth[1], and Danail Stoyanov[1]

[1]*Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London*
[2]*University College London Hospitals Biomedical Research Centre, National Institute for Health Research*

INTRODUCTION

Intratympanic steroid injections are commonly used for the treatment of ear diseases. During this treatment, an expert Ear, Nose Throat (ENT) clinician deliver the drug by viewing through a large microscope that provides a close-up view of the anatomical landmarks on the middle ear. A steady hand and swift response to any patient movement is required to avoid improper placement of the needle. To assist the clinician during this treatment, a fluidic soft robot is proposed in [1] that can steer inside a lumen for providing steady guidance for the drug delivery. For robust visual guidance, stable anatomical landmarks (tympanic membrane, malleus, umbo) segmentation is required.

The current method [1] uses a clinical data pixel-wise annotated for the segmentation model training, which does not generalise to the phantom ear data currently use for the in-lab validation of the soft robot. The clinical data is taken from a high-resolution optical microscope from a fixed camera perspective with diffuse, even lighting. While the phantom images are recorded on a miniature digital camera endoscopically, from multiple perspectives and often with uneven, highly local illumination. The phantom is 3d-printed from patient scans [2] and coated with a pigmented silicone rubber (Dragon Skin™ 30, Smooth-On Inc., Easton, PA, US) to create a skin-like surface texture. Transparent silicone rubber is employed to create a membrane to resemble the tympanic membrane. Whilst being close to real patient anatomy, the phantom exhibits visual differences in terms of the coloring of the tissues and tympanic membrane, the translucence of the tympanic membrane as well as the overall visibility of the middle ear structure. This means that labelled clinical data and the previously labelled phantom data are not representative of the images that will be passed to the model during deployment (Fig. 1).

Due to the difference in training and deployment data, model predictions can be both inaccurate (incorrectly identifying the structures of the tympanic membrane) and unstable (predictions are discontinuous and noisy). Both are serious challenges for using such models as part of a robot control system – especially, the lack of prediction stability. In this work, we perform intra-domain adaptation to learn a generalised model that provides stable and consistent segmentation on unseen phantom data.

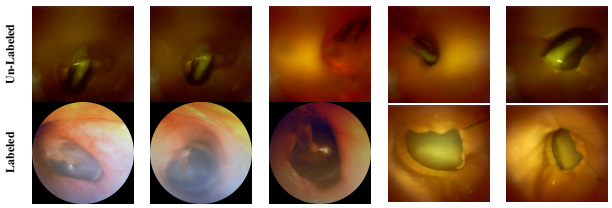

Fig. 1: Target domain (top) source domain (bottom)

MATERIALS AND METHODS

We propose using three segmentation models with DeeplabV3+ [3] architecture and three different backbones. We use transfer learning, initialising the models with ImageNet weights, and high amounts of data augmentation while using training techniques that encourage reduced convergence times for domain adaptation. These models are ensembled, with the output taking predictions of the composed logits for each class; taking advantage of each model whilst filtering out uncorroborated predictions. This results in both the stable and generalised predictions on unseen phantom data. Our approach takes advantage of smaller models' ability to learn smaller class features in fewer epochs and larger encoder's ability to output more stable predictions over a wider array of novel inputs.

Transfer Learning and Learning rate schemes: To compensate for the limited amount of labelled data and the need for the model to generalise to out of domain images, we used encoders pre-trained on Imagenet. To preserve useful, generalisable pattern and shape identifying convolutions shallower in the network, we utilise *discriminative learning rates* throughout the training - training shallower layers at a lower learning rate and deeper layers at a higher learning rate to increase the performance of the task specific segmentation. We also utilise *scheduled learning rates* – specifically the fit one cycle [4] which has been shown to both improve model performance and reducing convergence time. This approach increases the learning rate of the model for the first 30% of the total epochs; the learning rate is then annealed, returning to its original base value at the end of the training. Additionally, momentum is also adjusted to help with regularisation and ameliorate the

*This research was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145/Z/16/Z]; the Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P012841/1]; and the Royal Academy of Engineering Chair in Emerging Technologies Scheme.

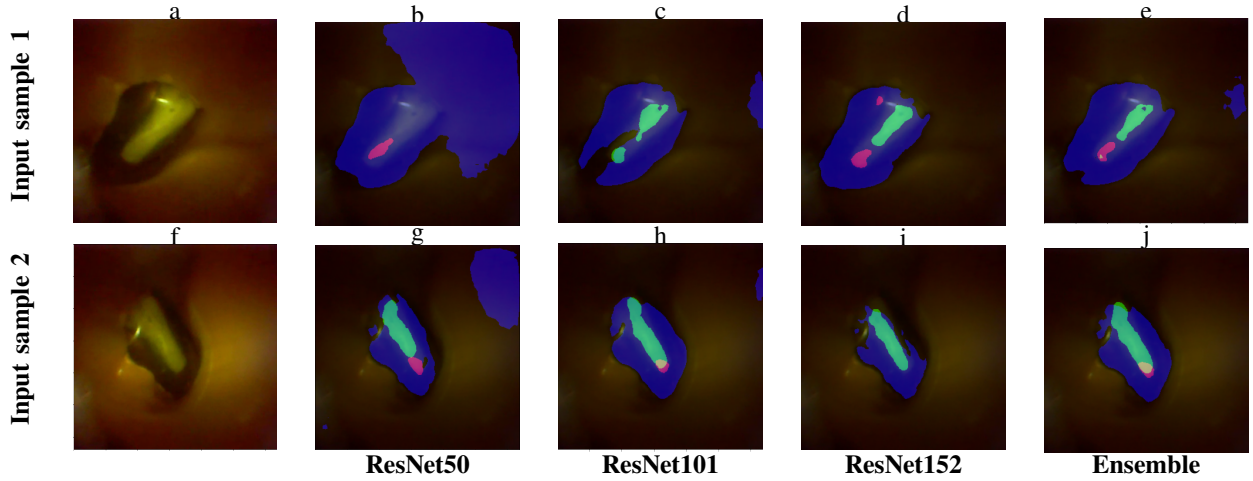

Fig. 2: Visual results showing unseen phantom input images and the model output when using ResNet50, ResNet101, ResNet101 as backbones for DeepLabV3+ [3] and the ensemble of all these models.

TABLE I: Qualitative comparison of different models using mean intersection over union (IoU) and F1 score.

Architecture	Encoder	IoU	F1 score
DeepLabV3+	ResNet50	0.8323	0.9085
DeepLabV3+	ResNet101	0.8277	0.9057
DeepLabV3+	ResNet152	0.8322	0.9084
Ensemble (proposed)	ResNet50/101	0.8353	0.9089

effects of the higher learning rate. Following the inverse of the learning rate schedule, momentum is reduced to its minimum value at 30% of total epochs, returning to its original value at the end of training. Learning rates were set using the *learning rate finder approach* [4].

Starting with a very low initial learning rate many batches are drawn, the loss calculated and recorded then optimised. With each mini batch the learning rate is incrementally increased until the loss explodes. This process allows us to efficiently and quickly estimate optimal learning rates without having to conduct large, slow and computationally expensive hyperparameter sweeps.

Data Augmentation: Data augmentation is an important part in stabilising and generalising model predictions. To help the model generalise to the unseen Phantom data we augment the data across three categories: image position (dihedral flips, rotation) perspective (warping, zooming) and lighting (brightness, saturation) Additionally, we also employed random resize crop which has been shown to both assist with generalisation and model performance.

Ensemble of Models: The ensemble mask prediction was created by composing the three models logits' for each class; The sum of the logits were taken, averaged and then passed through a threshold to create the ensemble's class prediction. For the malleus and umbo layers, the magnitude of the negative logits were reduced by 50% for the ResNet152 and ResNet101 encoders and 80% for the ResNet50 - This helps to reduce the effect of large negative logits from an unsure model cancelling out correct predictions. This is particularly necessary for small class features that can be easily obscured by false negative logits; conversely, the Tympanic membrane Benefits from taking the sum of (un-reduced) logits helping to stabilise boundary edges, especially with directional or off axis lighting.

RESULTS, DISCUSSION AND CONCLUSIONS

We observe from qualitative comparison that the ensembled model's prediction are both more stable and more robust even on off axis out of domain sequential images (videos). However, the improvement reported in Table I is marginal which is mainly because limited labelled data was available for the quantitative analysis that didnot capture variabilities that are observed in sequential data. Ensembling model of different backbone sizes creates outputs that have significantly more stable anatomical landmarks - specifically with the umbo which often is either not predicted (Fig. 2 (b), (c), (g)) or predicted in multiple locations (Fig. 2(d)). This can cause problems If the landmark is used As part of a positioning or typing system. Ensembling multiple different models has several advantages over alternative methods such as test time augmentation which while can effectively stabilise predictions severe off axis or novel lighting can insufficiently cause smaller, more marginal class features to be lost. However, increase in predictive power comes at the cost of greater prediction latency memory requirements.

REFERENCES

[1] L. Lindenroth, S. Bano, A. Stilli, J. G. Manjaly, and D. Stoyanov, "A fluidic soft robot for needle guidance and motion compensation in intratympanic steroid injections," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 871–878, 2021.

[2] D. Sieber, P. Erfurt, S. John, G. R. Dos Santos, D. Schurzig, M. S. Sørensen, and T. Lenarz, "Data descriptor: The openEar library of 3D models of the human temporal bone based on computed tomography micro-slicing," *Scientific Data*, vol. 6, pp. 1–9, 2019.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[4] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, 2019, p. 1100612.