Domain Adaptation of Echocardiography Segmentation Via Reinforcement Learning

Arnaud Judge[1], Thierry Judge[1,2], Nicolas Duchateau[2,3], Roman A. Sandler[4], Joseph Z. Sokol[4], Olivier Bernard[2], and Pierre-Marc Jodoin[1]

[1] Department of Computer Science, University of Sherbrooke, QC, Canada
[2] University of Lyon, CREATIS, CNRS UMR5220, Inserm U1294, INSA-Lyon
[3] Institut Universitaire de France (IUF)
[4] iCardio.ai

Abstract. Performance of deep learning segmentation models is significantly challenged in its transferability across different medical imaging domains, particularly when aiming to adapt these models to a target domain with insufficient annotated data for effective fine-tuning. While existing domain adaptation (DA) methods propose strategies to alleviate this problem, these methods do not explicitly incorporate human-verified segmentation priors, compromising the potential of a model to produce anatomically plausible segmentations. We introduce RL4Seg, an innovative reinforcement learning framework that reduces the need to otherwise incorporate large expertly annotated datasets in the target domain, and eliminates the need for lengthy manual human review. Using a target dataset of 10,000 unannotated 2D echocardiographic images, RL4Seg not only outperforms existing state-of-the-art DA methods in accuracy but also achieves 99% anatomical validity on a subset of 220 expert-validated subjects from the target domain. Furthermore, our framework's reward network offers uncertainty estimates comparable with dedicated state-of-the-art uncertainty methods, demonstrating the utility and effectiveness of RL4Seg in overcoming DA challenges in medical image segmentation.

Keywords: Domain Adaptation · Reinforcement Learning · Self-supervised · Echocardiography · Segmentation.

1 Introduction

Image segmentation using deep neural networks is accurate and reliable on many medical applications, including 2D echocardiography [11,1]. However, knowledge acquired from one domain (e.g. high quality segmentations on one dataset) does not confer to easy transferability to another without fine-tuning. To this end, domain adaptation (DA) aims to bridge the gap between datasets by limiting (sometimes removing) the amount of annotations required on a new dataset [4]. Leveraging unlabeled data is essential for DA methods as the collection of such data is inexpensive compared to the time needed for their labeling.

Many methods use pseudo-labels to learn from unlabeled data on the target domain [12]. Pseudo-labels are obtained from the predictions of a pre-trained

model on the target domain and can be used either by a second model dedicated to the target domain [25] or by fine-tuning the same source model[19]. However, this may introduce inaccurate information into the training process. Confidence based pseudo-labeling was recently introduced to handle this issue [3,20]. These methods integrate an additional sub-network to assess the quality of the generated pseudo-labels and weight their influence during training. Other DA techniques include image-to-image translation [7], namely matching the distributions of images from the source to the target domain before training. However, these methods do not explicitly consider anatomical correctness, resulting in segmentations with reasonable Dice but poor anatomical validity.

Reinforcement learning (RL) is widely used for a variety of tasks requiring an intelligent agent. Notably, RL from human feedback (RLHF) is used in language processing to obtain outputs aligned with human preferences [26,21,14]. ChatGPT is a popular example of this methodology's success. However, applications of RL to image segmentation remain mostly limited to accessory tasks such as hyper-parameter tuning or region of interest detection [6].

In this paper, we propose *RL4Seg*, a novel DA framework orthogonal to all previous works. The framework uses RL to bridge the gap between source and target domains and ensure high rates of anatomical validity of the target segmentations on a large dataset of 10,000 unannotated images. Taking inspiration from ChatGPT's protocol in learning how to output text aligned with human preferences, our model learns to output segmentations aligned with anatomical validity metrics, eliminating the need for expert interaction and annotations thus making the method self-supervised. In addition to image segmentation, our framework optimizes an error prediction network which is shown to be competitive with standalone state-of-the-art uncertainty methods[5].

2 Method

Considering a target dataset $\mathcal{D}_T = \{x_T^{(1)} ... x_T^{(n)}\}$ containing only images and a fully annotated source dataset $\mathcal{D}_S = \{(x_S^{(1)}, y_S^{(1)}) ... (x_S^{(m)}, y_S^{(m)})\}$, our method uses RL to optimize a neural network for segmenting images from a target domain. We illustrate the efficiency of our method on 2D echocardiographic images, one of the most challenging modalities for segmentation. Please note that our framework is generic and can fit other modalities and applications.

2.1 Reinforcement Learning

In the typical RL context, problems are posed as trajectories of states and actions following a Markov Decision Process [22]. At each time step t in the trajectory, a reward $r(s_t, a_t)$ is obtained for the action a_t taken by a policy π at the current state s_t and a new state s_{t+1} is reached. The policy is optimized with regards to the reward in order to maximize expected returns along a trajectory.

[5] Code is available at https://github.com/arnaudjudge/RL4Seg

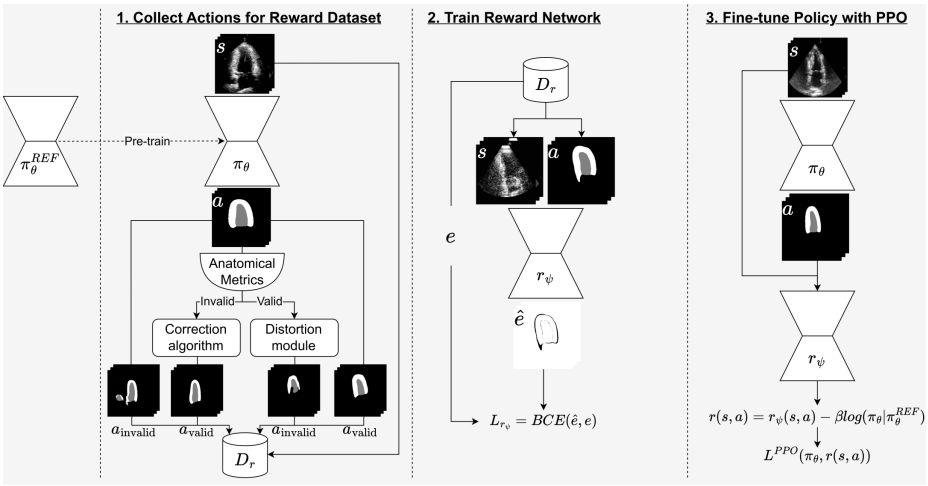

Fig. 1. RL4Seg, divided in three sections: [left] the reward dataset creation, [mid] the reward network training, and [right] the fine-tuning of the policy.

In the RL formalism, the value function V^π corresponds to the expected total reward the current policy π shall cumulate from state s_t until the end of the trajectory. It considers all possible actions that could be taken at state s_t by π, and can be expressed through Bellman's value function:

$$V^\pi(s_t) = \mathop{\mathbb{E}}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} [r(s_t, a_t) + \gamma V^\pi(s_{t+1})], \tag{1}$$

where $\gamma \in [0, 1]$ lowers the reward of actions further in time. Another key element is the Q function, which represents the total current and future reward for taking action a_t at state s_t, considering the current policy π. Bellman's equation for this function is the expectation under state transition probabilities of the current reward, plus the discounted Q values for all subsequent state-action pairs:

$$Q^\pi(s_t, a_t) = \mathop{\mathbb{E}}_{s_{t+1} \sim P(\cdot|s_t, a_t)} [r(s_t, a_t) + \gamma Q^\pi(s_{t+1}, a_{t+1})]. \tag{2}$$

Finally, an advantage function is defined as: $A(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)$. It describes the quality of the action taken given all possible actions the policy could take at that state, assuming the same policy dictates all future actions.

2.2 Segmentation RL

Through the lens of RL, the image segmentation problem involves trajectories of length one. The state ($s = s_0$) is the input image, while the action ($a = a_0$) corresponds to the predicted output of a segmentation model, i.e. the policy π. The reward $r(s, a)$ is the pixel-wise accuracy of the segmentation map a (the action) of a given image s (the state).

The Policy $\pi_\theta : \mathbb{R}^{H\times W} \to [0,1]^{K\times H\times W}$ is a neural network (U-Net) with parameters θ that outputs an action a (segmentation) given an input state s (image). K is the number of classes, and $H \times W$ is the image size. π_θ outputs a probability distribution over all possible actions (segmentations) via a *Softmax* function. This distribution is categorical over each pixel. During training, the actions are sampled from this distribution to explore the action space.

The Reward $r_\psi : \mathbb{R}^{2\times H\times W} \to [0,1]^{H\times W}$ is a second neural network (U-Net) with parameters ψ that estimates the reward for a given state s and action a (an image/segmentation pair). The reward is a pixel-wise error map of the given segmentation a. r_ψ has a *Sigmoid* output activation function. It is trained on the reward dataset $\mathcal{D}_r$ (Sec. 2.3), with a binary cross-entropy (BCE) loss function.

The Q, Value and Advantage Functions: in this single timestep context, the Q and value functions do not contain any future state-action pairs and state transition probabilities: $V^\pi(s) = \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}[r(s,a)]$, and $Q^\pi(s,a) = r(s,a)$. The Q function approximates the reward exactly while the value function approximates the expected reward at state s under the current policy. The advantage becomes $A(s,a) = r(s,a) - V^\pi(s)$, which estimates the quality of the segmentation action a compared to the average segmentation action the policy can take.

The Value operator $V_\phi^\pi : \mathbb{R}^{H\times W} \to [0,1]^{H\times W}$ is a third neural network (U-Net) with parameters ϕ that approximates the value function. Its input is the state s (the image), and its output is the anticipated reward map given the policy π_θ. It has a *Sigmoid* output activation function, as the possible rewards are in $[0,1]$.

Given the Reward and the Value networks r_ψ and V_ϕ^π, the advantage is computed by subtracting their predictions : $A(s,a) = r_\psi - V_\phi^\pi$.

2.3 RL4Seg

In the spirit of ChatGPT, our RL framework consists of three steps (Fig 1):

0. (Initialization) Before starting the RL loop, the segmentation neural network π_θ^{REF} is pre-trained on the fully annotated source dataset $\mathcal{D}_S$. π_θ^{REF} will stand as the first version of the target policy π_θ.
1. The policy π_θ first segments a subset of N images from the unannotated target dataset $\mathcal{D}_T$. Following a procedure described below, these segmentation maps are then post-processed and stored in a reward dataset $\mathcal{D}_r$.
2. The reward network $r_\psi(s,a)$ is trained on the reward dataset $\mathcal{D}_r$ to predict the error map e of a segmentation mask a associated to an image s.
3. A copy of the policy is stored in π_θ^{old}. Then, π_θ and V_ϕ^π are optimized with the newly trained reward model r_ψ and the target dataset $\mathcal{D}_T$ using the PPO RL algorithm (see below).

Steps 1 to 3 are repeated, each time with new samples from the unannotated target data to improve the policy π_θ and the reward network r_ψ. This goes on until every image of the target dataset $\mathcal{D}_T$ has been segmented.

Reward Dataset $\mathcal{D}_r$ It consists of pairs of images and segmentation masks (s^i, a^i) as well as their corresponding error map e^i. $\mathcal{D}_r$ is used to train the reward network r_ψ, which aims to predict e^i given (s^i, a^i).

First empty, $\mathcal{D}_r$ is populated during step 1 of the RL procedure. At each iteration, the policy π_θ segments a subset of N images from the target dataset, resulting in correct segmentation maps for some images and incorrect segmentations for others. Since these images are unannotated, the anatomical validity of the segmentation maps is determined based on prior knowledge about the segmented organ. In our case, we assess the validity of the segmentation maps with 10 cardiac anatomical metrics inspired by [16] (cf. Supplementary Material).

Each segmentation map a^i containing an anatomical error is post-processed with a dedicated warping system [15]. This system implements a variational autoencoder (VAE) that warps an anatomically invalid shape towards its closest valid shape (see [15] for more details). The post-processed mask $\hat{a}^i$, its associated image s^i, the invalid mask a^i and the pixel-wise difference e^i between the corrected and invalid masks are then stored in the reward dataset $\mathcal{D}_r$.

As for the anatomically valid segmentation maps a^i, many pairs of valid and invalid segmentations are created and added to $\mathcal{D}_r$, using small perturbations independently applied to the policy weights θ, the input image, and segmentations, aiming to simulate possible errors that a policy π_θ could produce.

Proximal Policy Optimization (PPO) Optimization of the policy π_θ is done according to the actor-critic style PPO algorithm [18]. In this paper, a two-term loss function is used : $L^{PPO} = L^{CLIP} + \alpha L^H$.

$L^{CLIP}(\theta) = \mathbb{E}_\theta[\min(\rho(\theta)A,\ \text{clip}(\rho(\theta), 1-\epsilon, 1+\epsilon)A)]$ is the *clipped surrogate loss* of the advantage function A pursuing two objectives. First, with the ratio $\rho(\theta) = \frac{\pi_\theta(a|s)}{\pi_\theta^{old}(a|s)}$[6], it favors an increase of the output probability of the policy π_θ compared to the old policy π_θ^{old} for high-reward segmentations and decreases probabilities for low-reward segmentations. Second, clipping $\rho(\theta)$ between $[1-\epsilon, 1+\epsilon]$ ensures that the policy updates remain reasonably small (we use $\epsilon = 0.2$), whereas the *min* operator allows for larger optimization steps in the direction of higher advantage when a previous update has led the policy to output actions with a worse outcome. The second loss term is $L_H = -\sum \pi_\theta \log(\pi_\theta)$, an entropy penalty on the policy's output distribution to ensure sufficient exploration.

The reward $r(s, a)$ used to calculate the advantage is obtained with the reward network r_ψ and a logarithmic penalty term to prevent the current policy π_θ from diverging from the reference π_θ^{REF} (the policy trained on $\mathcal{D}_S$):

$$r(s, a) = r_\psi(s, a) - \beta(\log\pi_\theta(a|s) - \log\pi_\theta^{REF}(a|s)). \tag{3}$$

where $\beta = 0.05$ in our experiments. When creating the reward dataset $\mathcal{D}_r$, the anatomically valid actions a^i are kept as gold standards for PPO. They are substituted into PPO in place of the actions taken by the policy, and their reward is set to 1 (maximum value) for all pixels. Thus, the PPO algorithm increases the probabilities that the policy will output such a segmentation map.

[6] $\pi_\theta(a|s)$ represents the probability of taking action a given state s, under the policy π_θ

Table 1. Results on the target data (average ± std. over 3 seeds) vs. the source dataset intra-expert variability. See Supplementary Material for an ablation study.

Method	Dice (%) ↑			Hausdorff (mm) ↓			Anatomical Validity (%) ↑
	ENDO	EPI	Avg.	ENDO	EPI	Avg.	
$\mathcal{D}_S$ intra-expert var.	94.4	95.4	94.9	4.3	5.0	4.6	100
Baseline (U-Net)	$89.9_{\pm 0.2}$	$93.7_{\pm 0.4}$	$91.8_{\pm 0.2}$	$7.0_{\pm 0.5}$	$9.4_{\pm 1.0}$	$8.2_{\pm 0.7}$	$91.5_{\pm 1.4}$
nnU-Net	$91.0_{\pm 0.1}$	$94.6_{\pm 0.0}$	$92.8_{\pm 0.0}$	$6.3_{\pm 0.2}$	$7.8_{\pm 0.4}$	$7.1_{\pm 0.3}$	$95.0_{\pm 0.7}$
UDAS [19]	$90.7_{\pm 0.3}$	$93.7_{\pm 0.1}$	$92.2_{\pm 0.2}$	$6.7_{\pm 0.3}$	$8.0_{\pm 0.5}$	$7.3_{\pm 0.4}$	$95.9_{\pm 1.0}$
TS-IT [3]	$90.5_{\pm 0.1}$	$93.6_{\pm 0.2}$	$92.0_{\pm 0.2}$	$6.1_{\pm 0.1}$	$8.2_{\pm 0.4}$	$7.1_{\pm 0.2}$	NA[7]
RL4Seg (ours)	$\mathbf{91.9}_{\pm 0.1}$	$\mathbf{94.7}_{\pm 0.1}$	$\mathbf{93.3}_{\pm 0.0}$	$\mathbf{4.9}_{\pm 0.1}$	$\mathbf{5.6}_{\pm 0.1}$	$\mathbf{5.3}_{\pm 0.1}$	$\mathbf{98.9}_{\pm 0.8}$

Uncertainty Estimation Once trained, the reward network r_ψ can serve as an uncertainty estimator by computing the complement to one of its output. In this way, high error probability areas have high uncertainty and vice-versa. For r_ψ to output calibrated uncertainty maps, temperature scaling [5] is applied during inference, using a scaling factor calculated with the validation set.

3 Experiments

Source Dataset: 500 echocardiography images (CAMUS dataset [11]) at end-diastole (ED) and end-systole (ES), in two- and four-chamber views, with left ventricle endocardium (ENDO) and epicardium (EPI) annotated by a cardiologist. The dataset was split into train-validation-test sets of 450-50-50 subjects.
Target Dataset: 10,000 unlabeled echocardiography images (at ED and ES) in two- and four-chamber views, from a heterogeneous private database, from various scanners and locations. A subgroup of 220 subjects were annotated and manually validated by two experts to be used as the test set for all experiments.
Pre-processing and Post-processing: All images in source and target datasets underwent identical preprocessing to bring the domains as close as possible. Images' contrast was increased locally through histogram equalization using *scikit-image*'s *exposure* package [23]. All output segmentations were post-processed to remove any disconnected regions.
Model Configuration: We used a U-Net with 7.8M parameters for all models and SOTA implementations (except nnU-Net). For 4 iterations ($N = 2500$ target images) of our framework, training time was 5 hours with a NVIDIA 3090 GPU.

Segmentation performance We compared our framework (Tab. 1) with two segmentation methods: a U-Net and the nnU-Net[8], and two DA methods:

[7] As the classes are segmented separately, anatomical validity cannot be computed reliably.

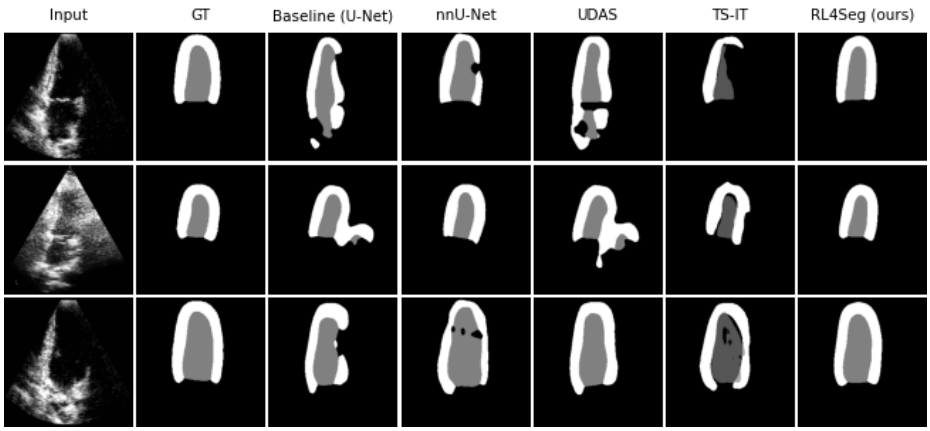

Fig. 2. Results for input images that the baseline was unable to segment correctly.

Self-Training of Early Features (UDAS)[19] and *Transformation-Invariant Self-Training* (TS-IT)[3]. UDAS trains the earliest layers of the network with pseudo-labels and a second segmentation head. As for TS-IT, it uses confidence masked pseudo-labels. All models were tested on the same expert-validated test set from the target dataset. Metrics are Dice, Hausdorff distance and anatomical validity.

All DA methods improve the segmentation results compared to simple supervised learning on the source domain (baseline). Our method, however, stands out with higher Dice scores, lower Hausdorff distances and notably, a higher rate of anatomical validity. Looking specifically at the Hausdorff distance, our method achieves substantially lower scores. This reflects the fact that the output segmentations have smoother borders with less variability. Errors are smaller in cases where the segmentation may be inaccurate. Also, holes and protrusions are almost nonexistent, which is reflected in the anatomical validity scores.

Representative segmentation results from the different methods are presented in Fig. 2. Selected images were poorly segmented by the baseline model, therefore examples show the improvement provided by these methods. This confirms the observations from Tab. 1. SOTA methods' overall coverage of the valid segmented areas is greater than the baseline, but anatomical inconsistencies remain. RL4Seg conserves anatomically valid shapes while improving segmentation quality.

While nnU-Net is also trained exclusively on the source domain, its usage of common voxel spacing and patches may help it generalize better [8], thus performing better on the target domain as well. UDAS helps deal with the domain shifts of high-level features present in the images, but not on possible differences of underlying structures between domains. The TS-IT method relies on reliable pixels present in pseudo-labels. Confidence may be underestimated in some regions, leading to holes in the output segmentations. In addition, TS-IT is binary in nature, so the coherence of the overall mask is lacking, as the left-ventricle and myocardium segmentation masks were computed separately.

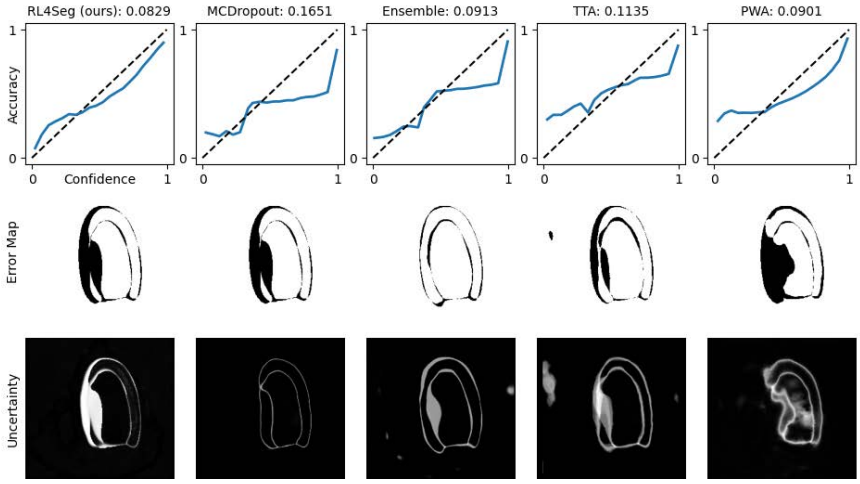

Fig. 3. Top: Reliability diagrams and ECE for five uncertainty methods, evaluated on the entire target test set. The dashed line represents perfect calibration. Middle and bottom: Examples of error map and corresponding uncertainty map for each method for segmentations from SOTA models on the target test set.

In general, these methods have reasonable performance on the target domain, but they remain limited by the inconsistencies in the output segmentations. While sufficient data is available for the model to learn anatomical constraints implicitly in the source domain, this is not the case for the target domain. Ignoring the underlying nature of the segmented structure leads to many incoherent segmentations corresponding to images for which the domain shift is the largest. Our method addresses this issue and provides highly consistent outputs, with results approaching intra-expert variability from the source dataset.

Uncertainty We compared our reward network with two epistemic uncertainty methods, Monte-Carlo Dropout (MCDropout) [2] and model ensembling [10], as well as two aleatoric methods, test-time augmentations (TTA) [24] and pixel-wise aleatoric uncertainty (PWA) [9]. Results for uncertainty estimation of predictions on the target domain are presented in Fig. 3. The best calibrated method is our reward network, with the lowest expected calibration error (ECE) [17]. Also, the reliability diagrams [13] show that our network is the most consistently calibrated through the entire range of output probabilities.

Perturbations introduced to both the input images and the model in the creation of the reward dataset allow our reward network to model both epistemic and aleatoric uncertainty, therefore performing better. The reward network stands out especially in cases where large errors occur, as many uncertainty methods model uncertainty along the border of the segmented structures.

4 Conclusion

We have presented RL4Seg, a novel domain adaptation framework using reinforcement learning. It produces both a strong segmentation model, and an accurate uncertainty estimation network without the need for any annotations on the target domain, which perform better than the state-of-the-art. In addition, our method not only limits the number of anatomical inconsistencies in the segmentations while improving metric scores, but also identifies erroneous or uncertain regions in the segmentation masks.

Acknowledgements We acknowledge support from the Natural Sciences and Engineering Research Council of Canada [551577754, RGPIN-2023-04584] and the Fonds de recherche du Québec en Nature et Technologies [334622], partial support from the French ANR (LABEX PRIMES [ANR-11-LABX-0063] of Université de Lyon, and from MIC-MAC [ANR-19-CE45-0005] and ORCHID [ANR-22-CE45-0029-01] projects).

Disclosure of Interests The authors have not competing interests to declare.

References

1. Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D.: Deep learning for cardiac image segmentation: A review. Frontiers in Cardiovascular Medicine **7**, 25 (2020)
2. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning (ICML). pp. 1050–1059 (2016)
3. Ghamsarian, N., Gamazo Tejero, J., Márquez-Neila, P., Wolf, S., Zinkernagel, M., Schoeffmann, K., Sznitman, R.: Domain adaptation for medical image segmentation using transformation-invariant self-training. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 331–341 (2023)
4. Guan, H., Liu, M.: Domain adaptation for medical image analysis: A survey. IEEE Transactions on Biomedical Engineering **69**, 1173–1185 (2022)
5. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning (ICML). pp. 1321–1330 (2017)
6. Hu, M., Zhang, J., Matkovic, L., Liu, T., Yang, X.: Reinforcement learning in medical image analysis: Concepts, applications, challenges, and future directions. Journal of Applied Clinical Medical Physics **24**, e13898 (2023)
7. Iacono, P., Khan, N.: Structure preserving Cycle-GAN for unsupervised medical image domain adaptation. arXiv (2023)
8. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203–211 (2021)
9. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
10. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
11. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., D'hooge, J., Lovstakken, L., Bernard, O.: Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Transactions on Medical Imaging **38**, 2198–2210 (2019)
12. Li, Y., Guo, L., Ge, Y.: Pseudo labels for unsupervised domain adaptation: A review. Electronics **12** (2023)
13. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. p. 625–632. ICML '05, Association for Computing Machinery (2005)
14. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. arXiv (2022)
15. Painchaud, N., Duchateau, N., Bernard, O., Jodoin, P.M.: Echocardiography segmentation with enforced temporal consistency. IEEE Transactions on Medical Imaging **41**, 2867–2878 (2022)
16. Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A., Jodoin, P.M.: Cardiac segmentation with strong anatomical guarantees. IEEE Transactions on Medical Imaging **39**, 3703–3713 (2020)

17. Pakdaman Naeini, M., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. Proceedings of the AAAI Conference on Artificial Intelligence **29**(1) (2015)
18. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv (2017)
19. Sheikh, R., Schultz, T.: Unsupervised domain adaptation for medical image segmentation via self-training of early features. In: International Conference on Medical Imaging with Deep Learning (MIDL). vol. 172, pp. 1096–1107 (2022)
20. Shen, Z., Cao, P., Yang, H., Liu, X., Yang, J., Zaiane, O.R.: Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. arXiv (2023)
21. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize with human feedback. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 3008–3021 (2020)
22. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction, 2nd ed. MIT Press, Cambridge (2018)
23. van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., the scikit-image contributors: scikit-image: image processing in Python. PeerJ **2**, e453 (2014)
24. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)
25. Zeng, X., Huang, R., Zhong, Y., Sun, D., Han, C., Lin, D., Ni, D., Wang, Y.: Reciprocal learning for semi-supervised segmentation, pp. 352–361 (2021)
26. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv (2019)