Domain adaptation for contrast-agnostic CT volumetric kidney segmentation

Ramon Correa-Medero [1 2] **Sam Fathizadeh** [2] **Fatima Al Khafaji** [2] **Haidar Abdul-Muhsin** [2] **Bhavik Patel** [2]
Imon Banerjee [1 2]

Abstract

The efficacy of segmentation models for CT volumes is limited to the contrast phase they were trained on and often do not work for the non-contrast images. We introduce a domain adaptation approach leveraging a single latent space discriminator to train a robust segmentation model for segmenting CT volume irrespective of the contrast dose. Our model is trained on two publicly available non-contrast and arterial phase image datasets, and validated on both public and private datasets. Evaluation of internal and external tests demonstrates improved segmentation quality while leveraging less data than baseline models.

1. Introduction

Computed tomography imaging (CT) is an essential modality for evaluating the genitourinary system by providing high-quality structural and anatomical information. The utility of CT imaging has been further augmented through artificial intelligence (AI) applications, leveraging automatic/manual segmentation of the kidney environment for the estimation of kidney volume, tumor classification, and disease staging (Gardan et al., 2018; Demirjian et al., 2022). However, AI models are often trained on images from a single contrast phase. Each of which, as seen in Figure 1, has a uniquely different appearance, particularly for soft tissue. Furthermore, imaging appearance within each phase is subject to variations based on patient physiology and image acquisition protocol. Therefore, models trained on single institution datasets tend to be more sensitive toward patient population, acquisition device, and imaging protocol. Curating representative datasets from multiple institutions with expert annotations would be cost-prohibitive. In prior literature, two key solutions have been studied to improve the robustness of segmentation models — (i) curation of application-specific datasets and (ii) domain adaptation to bridge the gap in image appearance.

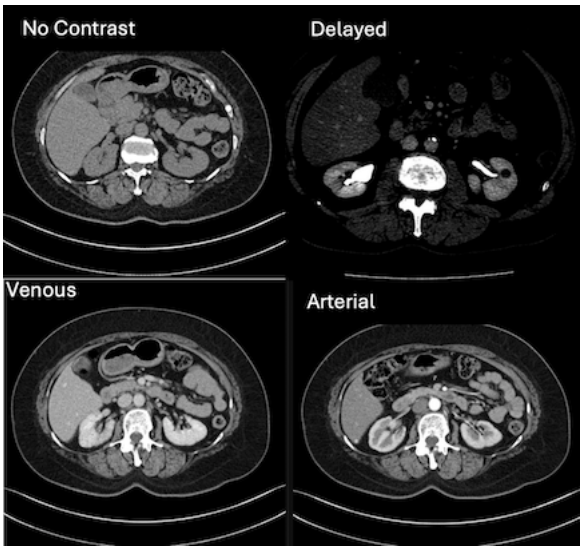

Figure 1. Abdominal CT scan of a single patient representing the appearance of kidney and surrounding organs under different contrast phase acquisitions - No contrast (top left), delayed (top right), venous (bottom left) and arterial (bottom right) phases

Curation of datasets primarily focuses on collecting and releasing datasets for specific contrast doses, which supports robust algorithmic development. Segmentation of the kidney has been a critical component in multi-organ segmentation challenges. The KiTs segmentation challenge provides a well-curated benchmark for developing kidney-specific segmentation models on 300 arterial phase CT images (Heller et al., 2023). Li et al. developed a dataset of 257 CT scans from Quanzhou Hospital to develop a kidney segmentation model able to detect kidney stones on non-contrast CT reliably scans (Li et al., 2022). Analysis of best-performing models has found that 3D U-net architectures with residual connections perform best for segmenting kidneys and related structures (Heller et al., 2023). These datasets provide access to expert kidney annotations but are limited to a single contrast phase with a carefully selected patient popu-

*Equal contribution [1]School of Computing and Augmented Intelligence, Arizona State University,Tempe, Arizona, United States of America [2]Department of Radiology Mayo Arizona, Scottsdale ,Arizona,United States of America. Correspondence to: Ramon Correa-Medero <correamedero.ramon@mayo.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

lation. Limiting efficiency of models when applied to other phases or patient populations.

Yu et al. trained a multi-organ nnU-net model using simultaneous contrast and non-contrast images. The authors found the model to achieve high dice scores for kidney segmentation with average DICE scores of 0.96 (Yu et al., 2022). Inspection of output segmentation found the model to underperform for non-contrast images as measured by lower quality assessment scores. Tang et al. took a similar approach where early and late arterial phase scans were used to train a patch-based network to segment renal structures (Tang et al., 2021). The group found the model to perform adequately on test data, with no distinction between late and early arterial phase performance. Lee et al. attempted to reduce the dependency of labeling multiple phases by using paired samples where only the contrast-enhanced volume was annotated (Lee et al., 2022). They were able to outperform existing models but were limited by the need for correct anatomical correspondence between scans. Domain adaptation techniques are another key strategy for building robust models. These techniques allow models to learn similar representations from diverse data, allowing them to perform robust segmentation tasks. Ananda et al. removed the dependency for paired contrast/non-contrast CT samples in liver segmentation by training a dual discriminator-based network penalizing model for retaining phase information while ensuring consistency of the augmentations. (Ananda et al., 2022). Dinsdale et al. implemented a similar multi-step approach to improve segmentation quality by improving resiliency to age-related physiological changes in Brain MRI segmentation. A discriminator concatenates features from the U-net bottleneck and decoder; it is then trained to predict patient ages. Confusion loss is introduced in a second training step to remove the encoding of age-related features that previously biased the segmentation model (Dinsdale et al., 2021). However, none of the prior works deal with the contrast discrepancy issue for segmenting normal and abnormal kidneys.

In this work, we aim to develop a novel phase-agnostic segmentation model that reliably segments normal and abnormally functioning kidneys across different contrast phases based on a data-efficient domain adaptation technique. [1]

2. Methods

2.1. Data Description

Two datasets with significantly different contrast phase images were used for model development. Contrast-enhanced images were obtained via the KiTS21 dataset (Heller et al., 2023), while non-contrast images were obtained from the dataset curated by Li et al., referred to as the STU dataset (Li et al., 2022). Both datasets were split into training, validation, and testing following a 75/19/6 % split. A mixed-phase dataset was also formed by combining corresponding splits of the KiTS21 and STU datasets. Additional external test sets were collected following institutional review board approval from Mayo Clinic. We collected CT scans of patients who underwent treatment for kidney dysfunction, and we leveraged venous phase and non-contrast CT scans as they were the most prevalent studies. Segmentation masks were annotated by expert urologists using visual assessment. Table 1 summarizes the training and testing splits of our data. All datasets were pre-processed to only consider the kidney tissue masks, and tumor and cyst masks were not included for this segmentation task. CT Volumes and kidney masks were resampled to a resolution of (0.75mm x 0.75mm x 0.75mm) using linear and nearest neighbor interpolation, respectively. CT image intensities were windowed to be in the range (-79, 304) and then scaled to the range (0 - 1).

Table 1. Train, validation, and test splits for each of the datasets. Mayo Clinic contrast and non-contrast datasets were only treated as external tests and were not included in the training and hyperparameter tuning.

Dataset	Number of Volumes	Train	Val	Test
KiTS21 (Arterial)	300	225 (75%)	56 (19%)	19 (6%)
STU (Non Contrast)	257	192 (75%)	48 (19%)	17 (6%)
Mayo Clinic (Venous)	63	-	-	63
Mayo Clinic (Non-Contrast)	42	-	-	42

2.2. Model Description

We design a model with a main segmentation task branch with DICE loss and an auxiliary branch from bottleneck for contrast phase identification with confusion loss that can simultaneously learn kidney segmentation and stay robust to image contrast appearance. The model was trained to segment voxel patches X where $x \epsilon R^{(1x128x128x64)}$.The U-net encoder produces latent representation $z = g_{enc}(X)$ with dimensionality $z \epsilon R^{64x16x16x16}$. The decoder produces segmentation mask $\hat{y} = g_{dec}(z)$. One auxiliary classifier ϕ was introduced to predict binary contrast phase $\hat{p}$ from U-net bottleneck features. The classifier was formed by four 3D convolution layers with relu activations followed by linear layers, producing the final binary phase classification.

Within each minibatch, the model is trained using three learning steps alternating between learning segmentation loss and removing phase information from the U-net encoder. The first step minimizes the DICE and cross-entropy (CE) loss calculated using difference between output seg-

[1] https://github.com/ramon349/domainadapt_segmentation

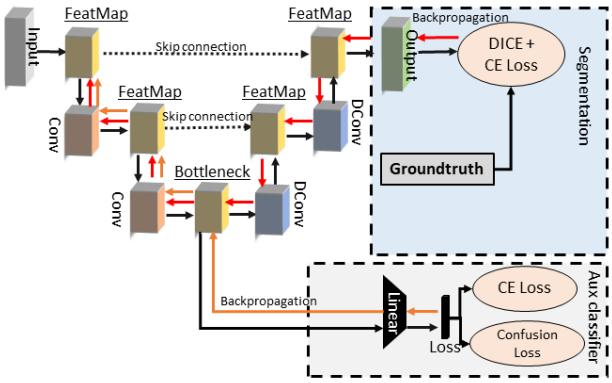

Figure 2. Proposed model diagram for performing domain adaptaton on the basis of contrast phase

mentation $\hat{y}$ and ground truth segmentation y, updating the U-net model. A second step updated the auxiliary classifier by minimizing the phase CE loss, allowing the phase classifiers to learn which model features indicate phase status. The phase loss calculation was constrained to only include patches where the kidney mask was present due to contrast information being uncertain in patches from non kidney regions. The third step updated the backbone in favor of features agnostic to phase information. The outputs of the auxiliary classifiers ($\hat{p}$) are fed to a KL divergence against the uniform distribution(Tzeng et al., 2015). Back-propagation of the error across the entire U-Net backbone encouraged the model to learn features independent of the contrast phase. The constraint based on the presence of a kidney mask was also applied for this loss propagation. The DICE+CE loss is also included during this final step to guide the model updates toward being phase agnostic while preserving kidney segmentation performance. The total optimization loss of our model is expressed as $L = L_{CE+DICE}(y, \hat{y}) + \lambda L_{CE}(p, \hat{p}) + \alpha L_{conf}(\hat{p})$, where λ and α are the regularization parameters.

2.3. Training Details

Two separate baseline 3D SegResNet models were trained for the internal datasets STU and KiTS21 (Myronenko et al., 2024). A mixed-phase model was trained using a combination of both datasets. The mixed phase and domain adaptation models were trained using the mixed dataset. To evaluate the data efficiency of the domain adaptation technique, separate models were trained with a mixed dataset that was x% KiTS21 and only y% STU data. Training for all models followed the same training scheme in terms of pre-processing, augmentation, and initial learning parameters.

Each minibatch saw two volumes randomly sampled, where eight 3D voxels of size (124,124,64) were randomly extracted. Patches were selected such that 50% were centered around a kidney mask while the remainder were centered around non-kidney regions. Each patch was then randomly augmented through random rotations, affine, and intensity changes. Hyperparameters λ and α were selected through a grid search via the optuna framework(Akiba et al., 2019). The search ran for 100 trials over 50% of the training data, and the average dice score on validation determined the optimal set of parameters. In our study, the best values were $\lambda = 0.134$ $\alpha = 0.084$, obtaining an average DICE of 0.747. Models were trained for 500 epochs using Adam optimizer with an initial learning rate of 0.001 and linear learning rate decay. Training was halted early if no improvement was seen for 20 epochs.

3. Results

Table 2 demonstrates comparative model performance evaluated using the DICE score, and auto bootstrap sampling was used to obtain 95% confidence intervals. Studying the single-phase models, i.e., trained on data from one phase (0/100 or 100/0 in Table 2), we found that they performed optimally on their corresponding phase, achieving DICE scores above 0.8 while achieving dice below 0.6 on other phases. In evaluating the external data from Mayo Clinic, the STU-trained model (100/0) has a significant drop in performance on both contrast data. This is likely due to wide variations between non-contrast to contrast tissue characteristics. The model trained on the KiTS21 only dataset (0/100) also observe a drop in model performance; however, it is less severe than the STU models. The trend is then swapped when observing the non-contrast data from Mayo Clinic, where the STU-based model performs best.

As a baseline, a mixed-phase model trained using an equal proportion (50/50) of contrast and non-contrast images demonstrates improved performance for STU and KiTS21 test sets compared to single-phase models. Performance in Mayo Clinic also improves, achieving a DICE score above 0.8 for both phases. The mixed model trained using a smaller proportion of non-contrast studies (10/90) noticeably showed a slight decrease in performance in the non-contrast studies. The model trained with one auxiliary classifier performs similarly to the mixed-phase model. While, the model with auxiliary classifier trained with a smaller proportion of non-contrast studies (10/90), improves the segmentation performance of all the contrast studies compared to the baseline approaches. Achieving the highest performance on the venous phase external data of Mayo Clinic with a dice score of 0.88. Performance on Mayo Clinic's non-contrast images is higher (0.7987) than the mixed phase model (0.7854) trained on the same proportion of data.

We also evaluated the performance of a model using two aux-

Model	%non Contrast/ %Arterial	KiTS Arterial (N=19)	STU Non Contrast (N=17)	Mayo Clinic Venous (N=63)	Mayo Clinic Non Contrast (N=42)	All Datasets
3D SegResNet	100/0	0.5850 (0.5088,0.6612)	0.8968 (0.8399,0.9537)	0.5032 (0.4929,0.5136)	0.8000 (0.7082,0.8917)	0.6171 (0.4612,0.7731)
	0/100	0.9573 (0.9363,0.9783)	0.5880 (0.4954,0.6805)	0.8325 (0.7061,0.9590)	0.5718 (0.4999,0.6438)	0.7966 (0.6278,0.9654)
	10/90	0.9543 (0.9328,0.9758)	0.9172 (0.8579,0.9765)	0.8573 (0.7387,0.9760)	0.7854 (0.6688,0.9019)	0.8697 (0.7505,0.9889)
	50/50	0.9541 (0.9241,0.9840)	**0.9541 (0.9079,1.0000)**	0.8523 (0.7483,0.9563)	0.8260 (0.7344,0.9177)	0.8885 (0.7943,0.9827)
3D SegResNet+ 1 Aux Classifiers	10/90	0.9619 (0.9494,0.9745)	0.9214 (0.8414,1.0000)	**0.8804 (0.7915,0.9694)**	0.7987 (0.6844,0.9130)	**0.8892 (0.7751,1.0000)**
	50/50	0.9511 (0.9327,0.9695)	0.8872 (0.8215,0.9529)	0.8290 (0.7410,0.9170)	**0.8322 (0.7439,0.9205)**	0.8654 (0.7756,0.9553)
3D SegResNet+ 2 Aux classifiers	10/90	**0.9636 (0.9492,0.9780)**	0.9261 (0.8641,0.9880)	0.8479 (0.7231,0.9728)	0.8115 (0.7098,0.9132)	0.8773 (0.7705,0.9841)
	50/50	0.9632 (0.9529,0.9735)	0.8928 (0.8121,0.9735)	0.8173 (0.7157,0.9188)	0.8212 (0.7224,0.920	0.8685 (0.7689,0.9680

Table 2. Model performance was evaluated on each of our test sets using the DICE metric. 95% confidence intervals are reported. Performance measured in DICe score. The best-performing model is highlighted in **BOLD**.

iliary classifier branches, one leveraging the bottleneck features and another using decoder outputs. Similarly, we find that training on a skewed proportion (10/90) of non-contrast and contrast scans improved performance over training on a balanced proportion.

4. Conclusion

Often, the primary challenge for developing a robust segmentation model for CT exams is the access to diverse and balanced training data with representation of all contrast phases. This work proposes a novel data-efficient domain adaptation technique to design a robust segmentation model with an auxiliary branch for handling variations in contrast dosages. Experimental results for kidney segmentation demonstrated that training a U-Net model on equal data representation from two contrast phases (arterial and non-contrast) improves performance on unseen phases from external dataset with poorly functioning kidneys. However, similar or even better performance can be achieved on both normal and poorly functioning kidneys by domain-adaptation with auxiliary branch even with a smaller amount of data from non-contrast phase. Additionally, we compared the benefits of training a U-Net segmentation model with only a single latent auxiliary classifier vs. an auxiliary latent classifier and a mask discriminator. We found that the one discriminator model obtained the best performance, which leads us to the conclusion that a robust encoder is essential for domain adaptation over the penalizing decoder component.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Ananda, S., Iwamoto, Y., Han, X., Lin, L., Hu, H., and Chen, Y.-W. Dual discriminator-based unsupervised domain adaptation using adversarial learning for liver segmentation on multiphase CT images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1552–1555, 2022. doi: 10.1109/EMBC48229.2022.9871188. ISSN: 2694-0604.

Demirjian, N. L., Varghese, B. A., Cen, S. Y., Hwang, D. H., Aron, M., Siddiqui, I., Fields, B. K. K., Lei, X., Yap, F. Y., Rivas, M., Reddy, S. S., Zahoor, H., Liu, D. H., Desai, M., Rhie, S. K., Gill, I. S., and Duddalwar, V. CT-based radiomics stratification of tumor grade and TNM stage of clear cell renal cell carcinoma. *European Radiology*, 32 (4):2552–2563, 2022. ISSN 0938-7994, 1432-1084. doi: 10.1007/s00330-021-08344-4.

Dinsdale, N. K., Jenkinson, M., and Namburete, A. I. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage*, 228:117689, 2021. ISSN 10538119. doi: 10.1016/j.neuroimage.2020.117689.

Gardan, E., Jacquemont, L., Perret, C., Heudes, P.-M., Gourraud, P.-A., Hourmant, M., Frampas, E., and Limou, S. Renal cortical volume: High correlation with pre- and

post-operative renal function in living kidney donors. *European Journal of Radiology*, 99:118–123, 2018. ISSN 0720-048X. doi: https://doi.org/10.1016/j.ejrad.2017.12.013.

Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., Ma, J., Rickman, J., Dean, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Kaluzniak, H., Raza, S., Rosenberg, J., Moore, K., Walczak, E., Rengel, Z., Edgerton, Z., Vasdev, R., Peterson, M., McSweeney, S., Peterson, S., Kalapara, A., Sathianathen, N., Papanikolopoulos, N., and Weight, C. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge, 2023. URL http://arxiv.org/abs/1912.01054.

Lee, H. H., Tang, Y., Gao, R., Yang, Q., Yu, X., Bao, S., Terry, J. G., Carr, J. J., Huo, Y., and Landman, B. A. Pseudo-label guided multi-contrast generalization for non-contrast organ-aware segmentation, 2022.

Li, D., Xiao, C., Liu, Y., Chen, Z., Hassan, H., Su, L., Liu, J., Li, H., Xie, W., Zhong, W., and Huang, B. Deep segmentation networks for segmenting kidneys and detecting kidney stones in unenhanced abdominal CT images. *Diagnostics*, 12(8):1788, 2022. ISSN 2075-4418. doi: 10.3390/diagnostics12081788.

Myronenko, A., Yang, D., He, Y., and Xu, D. Automated 3D Segmentation of Kidneys and Tumors in MICCAI KiTS 2023 Challenge. In Heller, N., Wood, A., Isensee, F., Rädsch, T., Teipaul, R., Papanikolopoulos, N., and Weight, C. (eds.), *Kidney and Kidney Tumor Segmentation*, pp. 1–7, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-54806-2.

Tang, Y., Gao, R., Lee, H. H., Xu, Z., Savoie, B. V., Bao, S., Huo, Y., Fogo, A. B., Harris, R., de Caestecker, M. P., Spraggins, J. M., and Landman, B. A. Renal cortex, medulla and pelvicaliceal system segmentation on arterial phase CT images with random patch-based networks. In Landman, B. A. and Išgum, I. (eds.), *Medical Imaging 2021: Image Processing*, pp. 42. SPIE, 2021. ISBN 978-1-5106-4021-4 978-1-5106-4022-1. doi: 10.1117/12.2581101.

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. Simultaneous deep transfer across domains and tasks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4068–4076, 2015. doi: 10.1109/ICCV.2015.463.

Yu, C., Anakwenze, C. P., Zhao, Y., Martin, R. M., Ludmir, E. B., S.Niedzielski, J., Qureshi, A., Das, P., Holliday, E. B., Raldow, A. C., Nguyen, C. M., Mumme, R. P., Netherton, T. J., Rhee, D. J., Gay, S. S., Yang, J., Court, L. E., and Cardenas, C. E. Multi-organ segmentation of abdominal structures from non-contrast and contrast enhanced CT images. *Scientific Reports*, 12(1):19093, 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-21206-3.

Latinx Member Contribution

Ramon Luis Correa-Medero (Presenting Author) contributed to designing/execution of models and experiments in this work. He also contributed to curation of the Mayo Clinic datasets. Finally, he prepared an initial draft of the manuscript alongside the figures and tables.