# Enhance Multi-Modal and Multi-Center Whole Heart Segmentation using Data Augmentation and Model Calibration

Charlie Tran[1], Andy Li[2], Aaron Espinoza[3], Sayem Kamal[4], Anoushka Samuel[5], Charles Jiang[6], Jian Zhuang[7], Yiyu Shi[6], and Xiaowei Xu[7]

[1] University of Florida, Gainesville, Florida 32611, USA
[2] University of Texas at Austin, Austin, Texas 78712, USA
[3] Texas State University, San Marcos, Texas 78666, USA
[4] Columbia University, New York City, New York 10027, USA
[5] California Polytechnic State University, San Luis Obispo, California 93410, USA
[6] University of Notre Dame, Notre Dame, Indiana 46556, USA
[7] Guangdong Provincial People's Hospital, Guangzhou 510000, China

**Abstract.** Accurate whole heart segmentation can enhance the modeling and analysis of cardiovascular disease treatment. However, real-world imaging analysis faces challenges in generalization due to data inhomogeneity, which arises from factors such as low-quality image acquisition, variations across multiple modalities, differences in scanner vendors, and cardiac motion. In this paper, to tackle the problem of data inhomogeneity, we use data augmentation and model calibration for whole heart segmentation enhancement. We first adopted the state-of-the-art MedNeXt transformer-based architecture as the baseline. Then, we enhance the baseline in model robustness through an ensemble strategy with strong data augmentation and out-of-distribution model calibration techniques. Comprehensive experiments have been conducted on the dataset from the CARE2024 Task 5 Whole Heart Segmentation (WHS++) challenge, in which 7 target structures from CT and MRI images acquired from multiple centers are considered. Results show that data augmentation and model calibration can effectively improve the segmentation performance across various modalities and centers. We highlight that our team ranks first on the validation leaderboard with average dice scores of 0.9440 (CT) and 0.8956 (MRI).

**Keywords:** Deep learning · Whole heart segmentation

## 1 Introduction

Cardiovascular disease is the leading cause of death globally [14]. Accurate delineation of cardiovascular structures from medical imaging is essential for guiding treatment options. However, manual segmentation by experts is time-consuming due to the vast volume of data, variations in shape, and inconsistencies in data quality. Deep learning (DL) models offer a promising approach to streamlining

image segmentation due to their efficient feature extraction qualities from big data. Integration of artificial intelligence (AI) approaches for clinical usage is often limited due to the heavy burden of patient risk and safety. Desirable AI models must achieve high performance and robustness for the diverse conditions in real world clinical practice.

The CARE2024 Task 5 Whole Heart Segmentation (WHS++) challenge addresses real-world image analysis problems in semantic segmentation. This challenge evolves from prior work [18, 16, 17, 2], featuring a broader selection of data, centers, and scanner vendors. In the prior challenge [16], methods showcased architectures like fully convolutional networks (FCN) and U-Net, along with techniques such as region of interest (ROI) localization and image registration. Recent advancements have introduced more sophisticated techniques and architectures that address challenges in medical image segmentation. In particular, innovations in data augmentation and model calibration are crucial for overcoming performance degradation across diverse datasets.

In this work, we choose the MedNeXt transformer-based architecture for its state-of-the-art performance on medical image segmentation. To enhance model robustness against domain shifts and data inhomogeneity, we utilize stronger data augmentation strategies and DOMINO++ out-of-distribution model calibration. Our ensemble of three MedNeXt configurations ranks first on the leaderboard with dice scores of 0.9440 for CT and 0.8956 for MRI.

## 2 Methods

### 2.1 CARE2024 Challenge Overview

The multi-modality, multi-center Whole Heart Segmentation (WHS++) challenge was established to foster the development of robust automated models for enhancing the treatment of cardiovascular diseases. The dataset features expert annotations for seven key anatomical structures: (1) Left Ventricular Blood Cavity (LV), (2) Right Ventricular Blood Cavity (RV), (3) Left Atrial Blood Cavity (LA), (4) Right Atrial Blood Cavity (RA), (5) Myocardium of the Left Ventricle (Myo), (6) Ascending Aorta (AO), and (7) Pulmonary Artery (PA). The dataset includes 40 + 46 (CT + MRI) training samples from five distinct centers (A, B, C, D, E), 30 + 20 (CT + MRI) validation samples from four centers (A, B, C, D), and 34 + 36 (CT + MRI) testing samples from five centers (A, B, C, D, F), with Center F being previously unseen. Data acquisition from each center varied utilizing different scanners, namely from Philips, Siemens, and GE.

Models are formally evaluated using the Dice Score (DSC), Hausdorff Distance (HD), and Average Symmetric Surface Distance (ASSD). For simplicity, we use the Dice Score as our primary metric for analysis.

### 2.2 Model Architectures and Implementation

MedNeXt [10] serves as our primary architecture due to its state-of-the-art performance on various medical segmentation tasks [6, 10]. The MedNeXt architecture resembles a 3D U-Net [9], featuring transformer-like qualities through a

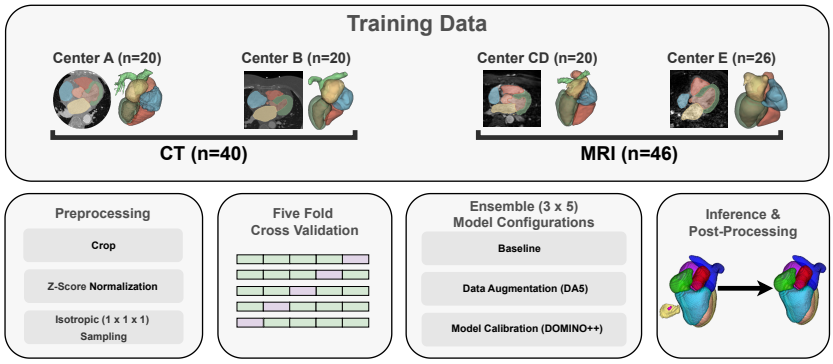


**Fig. 1.** Our MedNeXt pipeline, consisting of an ensemble with data augmentation and model calibration configurations.

ConvNeXt [7] backbone. To enhance the preservation of rich semantic information at lower resolutions, residual inverted blocks replace traditional upsampling and downsampling layers. Additionally, compound scaling [12] is employed to create different model configurations (S/B/M/L), simultaneously scaling both the number of network blocks (model depth) and the expansion ratio (model width). MedNeXt also allows larger kernel sizes (model receptive field) utilizing Upkern initialization from a trained $3 \times 3 \times 3$ kernel onto a $5 \times 5 \times 5$ model through trilinear interpolation to mitigate performance saturation (see Supplementary Section 5.2).

MedNeXt operates on top of the nnU-Net [4] framework with few modifications. Data pre-processing includes cropping, intensity normalization, and $1 \times 1 \times 1$ isotropic spacing. Models are trained according to five-fold cross-validation and ensembled at test time. Training is conducted for 1000 epochs with 250 iterations per epoch. At inference, a 50% patch overlap is used. For post-processing, an automatically determined connected components analysis-based object removal is applied to each class.

We opt for a simple approach for data compilation as each subject does not contain both CT and MRI images. The data from each modality are combined for the training set (partitioned according to five-fold cross-validation) and normalized using z-score normalization (nnU-Net's *nonCT* setting) which is a strategy seen by nnU-Net on alternative challenges [5]. Aside from architectural differences, we highlight several design choices that may enhance MedNeXt's performance over the conventional nnU-Net.

- MedNeXt uses the AdamW [8] optimizer with a learning rate of 0.001 and weight decay of $\lambda = 3e - 5$ in contrast to nnU-Net's stochastic gradient descent with learning rate 0.01.
- Group Normalization [15] is utilized by MedNeXt over Instance Normalization [13] used by nnU-Net, both of which are motivated as alternatives to the weakness of Batch Normalization [3] on small batches.

– MedNeXt parameters are fixed, with isotropic $1 \times 1 \times 1$ spacing and patch size of $128 \times 128 \times 128$. This differs from nnU-Net, which uses auto-configured settings based on the dataset's median. For this dataset, nnU-Net employs a spacing of $1 \times 0.885 \times 0.885$ and a patch size of $96 \times 160 \times 160$. Work by the nnU-Net authors claims the isotropic spacing is a major contributor to MedNeXt's success [6].

### 2.3 Data Augmentation

Data augmentation is a practical strategy for encouraging model generalization. Owing to the compatibility of MedNeXt with the nnU-Net pipeline, we explored three different presets for data augmentation: (1) the default of MedNeXt/nnU-Net, (2) DA-M&M [1] developed for the Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge, and (3) DA5 [5] showcased in the AMOS2022 challenge. Augmentation strength (probability, scaling range, etc.) generally increases for each preset beyond the default. Note the M&M dataset utilizes only cardiac MRI, while the AMOS2022 dataset contains abdominal CT and MRI images.

**Default:** The augmentations consist of random rotations, random scaling, gamma correction, and mirroring.

**M&M:** The augmentations consist of random rotations, elastic deformations, random scaling, gamma corrections, and brightness changes. Note, the work [1] also studied the improvement of batch normalization [3] together with the M&M augmentations. MedNeXt opts for Group Normalization [15].

**DA5:** Significant modifications to the default include additional spatial orientations (90-degree rotations and transposition), median filtering, Gaussian blurring, Gaussian noise, brightness, contrast, low-resolution simulation, gamma corrections, rectangle blanketing, brightness gradient additions, local gamma adjustment, and sharpening.

### 2.4 DOMINO++ Model Calibration

Model calibration is a concept of refining a model's decision boundaries based upon aligning the predicted probability estimates with their true likelihood. Moreover, this approach is linked with improved generalization of out-of-distribution data (e.g. the unseen Center F). We utilize a recent technique, Domain-aware model calibration (DOMINO++) [11], originally developed for whole brain segmentation onto our whole heart segmentation task. We adapt this approach with modifications according to the five-fold cross-validation and our loss function. First, a baseline uncalibrated model is evaluated for its respective fold in the cross-validation set. Second, the model's output confusion matrix is obtained

and normalized across each row. Next, a penalty matrix is obtained by subtracting the normalized confusion matrix by an all ones matrix, $\mathbf{1}\mathbf{1}^T$ where the main diagonal generates no penalty and off-diagonal elements contain different penalty weights. Finally, we embed the penalty matrix as a loss function regularizer, penalizing a newly trained model more for misclassifications between semantically distinct classes than for those between semantically similar classes, thereby calibrating the model's decision boundaries. The regularized loss is given by Equation 1.

$$(1-\beta)\mathcal{L}\left(\mathbf{y},\hat{\mathbf{y}}\right)+\beta\mathbf{y}^T\left(s\mathbf{W}\right)\hat{\mathbf{y}},\quad \beta=1-\frac{\text{Iteration}_{\#}}{\text{Iteration}_{\text{Total}}} \tag{1}$$

The scalar s is a dynamic scaling set to be the same order of magnitude as the base loss term $\mathcal{L}(\mathbf{y},\hat{\mathbf{y}})$ (e.g. if the base loss is 15, then the regularization loss is scaled by $\mathbf{s}=10$). The loss function varies *adaptively* with the number of iterations as defined by $\beta$ where the base loss function $\mathcal{L}(\mathbf{y},\hat{\mathbf{y}})$ increases in weight and the regularization term decays throughout the training. Note that the base loss function of MedNeXt $\mathcal{L}(\mathbf{y},\hat{\mathbf{y}})$ is the deep supervision loss, which computes the weighted DiceCrossEntropy loss at five different output scales (the original output scale and at four additional downsampled resolutions). In accordance, we compute the DOMINO calibration at each individual weighted term of the deep supervision loss summed together to evaluate the total loss. Refer to Supplementary Figure 4 for an example of the obtained DOMINO++ matrix.

## 3 Results and Discussion

### 3.1 Model Selection: Five-Fold Cross-Validation

Five-fold cross-validation is a traditional benchmark for determining optimal models on held-out validation/test sets [6]. We confirm consistent findings of the MedNeXt architecture outperforming the historically powerful nnU-Net architectures [6, 10] and conclude to make use of the MedNeXt-M architecture based on the higher baseline performances (Table 1). Our comparison also includes a comparison with the more recent nnU-Net with Residual Encoders [5, 6] (M).

**Table 1.** Five-fold cross-validation dice scores without post-processing. The results are further partitioned according to the modality and center for further analysis.

| Model | CT | $\text{CT}_\text{A}$ | $\text{CT}_\text{B}$ | MRI | $\text{MRI}_{\text{C,D}}$ | $\text{MRI}_\text{E}$ | AVG |
|---|---|---|---|---|---|---|---|
| | n = 40 | n = 20 | n = 20 | n = 46 | n = 20 | n = 26 | n = 86 |
| nnU-Net | 0.9395 | 0.9277 | 0.9514 | 0.8820 | 0.8680 | 0.8927 | 0.9087 |
| nnU-Net-ResEnc-M | 0.9405 | 0.9278 | 0.9532 | 0.8842 | 0.8692 | 0.8958 | 0.9104 |
| MedNeXt-S-k3 | 0.9404 | 0.9270 | 0.9538 | 0.8851 | 0.8700 | 0.8967 | 0.9108 |
| MedNeXt-B-k3 | 0.9408 | 0.9276 | 0.9540 | 0.8838 | 0.8691 | 0.8951 | 0.9103 |
| MedNeXt-M-k3 | 0.9414 | 0.9280 | 0.9549 | 0.8845 | 0.8715 | 0.8945 | 0.9110 |
| MedNeXt-L-k3 | 0.9410 | 0.9281 | 0.9538 | 0.8835 | 0.8700 | 0.8939 | 0.9102 |

### 3.2 Data Augmentation, Calibration, and Ensembling

Following the selection of MedNeXt-M-k3 as our baseline model, we experiment with additional data augmentation (DA5, MMDA) and model calibration (DOMINO++) techniques. The results are summarized in Table 2 as an ablation study of five-fold cross-validation performances. The MedNeXt-M model attains a baseline dice score of 0.9110. Pairing configurations together, we observe the DA5 yields larger improvements to the ensemble followed by DOMINO++. Our final model consists of a triple ensembling of the baseline MedNeXt-M, DA5, and DOMINO++ for an average dice score of 0.9129 and 0.9132 with connected components post-processing. Our validation analysis from Supplementary Table 7 suggests this data augmentation and model calibration strategy enhance the model performance more than ensembling across different model complexities (e.g. B & M & L), and over the individual baseline M-k3 model itself.

**Table 2.** Ablation study: 5-fold cross-validation results for our model configurations and their ensembles. Our final submission model is bolded.

| **Configuration** | **Dice CV** |
|---|---|
| 1) MedNeXt-M-k3 | 0.9110 |
| 2) MedNeXt-M-k3-DOMINO++ | 0.9104 |
| 3) MedNeXt-M-k3-DA5 | 0.9101 |
| 4) MedNeXt-M-k3-MMDA | 0.9096 |
| Config (1,2) | 0.9118 |
| Config (1,3) | 0.9126 |
| Config (1,4) | 0.9117 |
| Config (2,3) | 0.9125 |
| Config (2,4) | 0.9114 |
| Config (3,4) | 0.9118 |
| **ENSEMBLE** (1,2,3) | 0.9129 |
| + postprocessing | 0.9132 |

### 3.3 Model Validation

Our submission consists of a three-configuration (15 models) ensemble of MedNeXt-M-k3 configurations (M/M-DA5/M-DOMINO++). Our CT validation scores are 0.9440 (DSC), 12.8562 (HD), and 0.6984 (ASSD). Our MRI validation scores are 0.8956 (DSC), 17.8029 (HD), and 1.2127 (ASSD). Further comparisons are provided in Table 3 between the five-fold cross-validation and validation performance, and all validation attempts considered by our team in Supplementary Table 7 (showcasing the benefits of our DA5/DOMINO++ ensembling strategy). Notably, validation performances exhibit greater performance and reduced variance as a consequence of a complete 15 model ensembling while five-fold cross-validation incorporates only 3 models (1 per configuration) for each fold.

**Table 3.** Five-fold CV results compared to the validation performance after post-processing. The results are written as the mean (standard deviation) for each metric.

| Method | Five-Fold CV | | | Validation | | |
|---|---|---|---|---|---|---|
| | DSC(↑) | HD(↓) | ASSD(↓) | DSC(↑) | HD(↓) | ASSD(↓) |
| CT | 0.9427 (0.0248) | 10.6493 (6.1063) | 0.7116 (0.3293) | 0.9440 (0.0185) | 12.8562 (5.5664) | 0.6984 (0.2486) |
| MRI | 0.8876 (0.0469) | 15.8149 (11.8186) | 1.5146 (1.6735) | 0.8956 (0.0266) | 17.8029 (3.5774) | 1.2127 (0.3889) |

### 3.4 Performance Analysis

We discuss generalizations of our model performances from five-fold cross-validation and validation according to the modality, center, and class.

**Modality (CT vs. MRI):** Our results suggest segmentation quality of CT images is consistently higher than MRI images. On the validation set, we obtained dice scores of 0.9440 (CT) compared to 0.8956 (MRI). This discrepancy is consistent with that reported in the prior challenge [17], and related to the stronger clarity (image quality) of anatomical structures in the CT images. A worst and median case analysis is shown in Supplementary Figure 5 which suggests our method provides strong segmentation results with a minimal whole heart validation dice score of 0.9125 (CT) and 0.8429 (MRI).

**Center:** A summary of our performance on each center is shown in Table 4. On the CT centers, we obtain validation performances of 0.9351 (Center A) and 0.9618 (Center B). The increase in Center B is consistent with the five-fold cross-validation (0.9224 Center A vs. 0.9559 Center B). Performance comparisons with Center E are not available in the validation set. On Centers C&D, we observe a larger gap between training and validation dice scores of 0.8749 and 0.8956 than similar comparisons derived from Center A and Center B. This suggests both a larger diversity in MRI instances, as well as the benefits of our enhanced ensembling strategy at inference time.

**Class:** For each class, we list the five-fold cross-validation scores in Supplementary Table 5 and showcase the box-plot of dice scores for the validation set in Figure 2. Our results indicate that our enhanced ensembling strategy enabled stable segmentation with relatively low variance for most classes. Segmentation results for the pulmonary artery (PA) are consistently worse for both modalities (0.8867 and 0.8221 for CT and MRI respectively). This is exemplified by the pulmonary artery variance in the CT evaluation and outlier in the MRI evaluation.

We showcase 2D visualizations of the training set in Figure 3 (with ground truth) and 3D visualizations of the validation set in Supplementary Figure 5. The

**Table 4.** Five-fold cross-validation results for each center on both the five-fold cross-validation and validation set after post-processing. The results are written as the mean (standard deviation) for each metric.

| **Method** | **Five-Fold CV** | | | **Validation** | | |
|---|---|---|---|---|---|---|
| | DSC(↑) | HD(↓) | ASSD(↓) | DSC(↑) | HD(↓) | ASSD(↓) |
| Center A | 0.9294 (0.0215) | 11.6240 (4.2989) | 0.8830 (0.3187) | 0.9351 (0.0133) | 14.0226 (5.4484) | 0.8023 (0.20302) |
| Center B | 0.9559 (0.0208) | 9.6745 (7.4871) | 0.5402 (0.2432) | 0.9618 (0.0141) | 10.1345 (5.1189) | 0.4562 (0.1635) |
| Center C&D | 0.8749 (0.0446) | 20.8940 (14.7855) | 1.2611 (0.3074) | 0.8956 (0.0266) | 17.8029 (3.5774) | 1.2127 (0.3889) |
| Center E | 0.8974 (0.0472) | 11.9079 (6.9834) | 1.7097 (2.2087) | - | - | - |

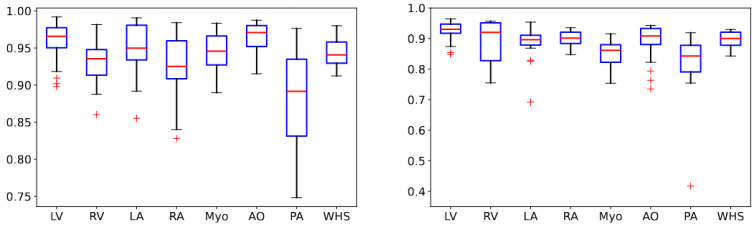


**Fig. 2.** Box-plot of validation dice scores for CT (left) and MRI (right) for each class.

predominant error occurs due to natural ambiguity near class boundaries. Another error of concern is related to preciseness in labeling. In Figure 3, we reveal from the training set labels that the lengths of the great vessels are vulnerable to inconsistency. This leads to under/over-estimation of the model performance across images, specifically when concerned with the great vessels.

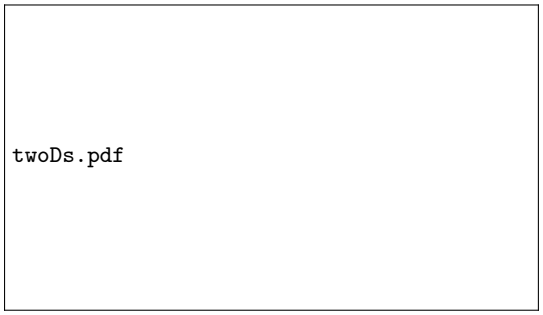


**Fig. 3.** 2D plots of the worst cases of Centers B, C&D, and E. They contain the centers' predictions, ground truth, and an overlay to highlight the errors of our model.

## 4 Conclusion

We present the leading validation submission of the CARE2024 Whole Heart Segmentation Challenge by modifying the MedNeXt framework. In summary, we incorporated the baseline MedNeXt-M-k3 architecture and embedded stronger data augmentation techniques and DOMINO++ model calibration techniques as a three-configuration-based ensemble with 15 models total obtained from five-fold cross-validation. These modifications achieved a validation dice score of 0.9440 and 0.8956 on CT and MRI, respectively. Our strategy is justified by thoroughly benchmarking the five-fold cross-validation protocol, as well as demonstrating improvement over the (highly anticipated) nnU-Net designs.

Our extended analysis provides insights into the strengths and weaknesses of our model. We showcase stronger performance on CT images than on MRI, which can be partly explained by the visual image quality (contrast, signal-to-noise ratio, etc.). Segmentation performances are highest for the four chambers (dice score $> 0.9$) and are vulnerable to decline for the myocardium (Myo) and the great vessels (AO, PA). While greatly improved, these weaknesses are overall consistent with that of the prior challenge [17]. One concern worthy of discussion can be understood by our benchmarking results on the five-fold cross-validation and validation dataset, demonstrating an evident performance saturation. This saturation may stem from several factors including small data limitations, data inhomogeneity, labeling errors, etc. Rigorous studies for improving model robustness, including those incorporating data augmentation, domain adaptation, model uncertainty, and model calibration are encouraged for future approaches. Nevertheless, we confidently present our solution onto the test set through enhanced MedNeXt ensembling with data augmentation and model calibration.

## References

1. Full, P.M., Isensee, F., Jäger, P.F., Maier-Hein, K.: Studying robustness of semantic segmentation under domain shift in cardiac mri. In: Puyol Anton, E., Pop, M., Sermesant, M., Campello, V., Lalande, A., Lekadir, K., Suinesiaputra, A., Camara, O., Young, A. (eds.) Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges. pp. 238–249. Springer International Publishing, Cham (2021)
2. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. Medical Image Analysis **89**, 102889 (2023)
3. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015)
4. 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**(2), 203–211 (Dec 2020). https://doi.org/10.1038/s41592-020-01008-z, http://dx.doi.org/10.1038/s41592-020-01008-z
5. Isensee, F., Ulrich, C., Wald, T., Maier-Hein, K.H.: Extending nnu-net is all you need (2022), https://arxiv.org/abs/2208.10791

6. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. arXiv preprint arXiv:2404.09556 (2024)
7. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
10. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
11. Stolte, S.E., Volle, K., Indahlastari, A., Albizu, A., Woods, A.J., Brink, K., Hale, M., Fang, R.: DOMINO++: Domain-Aware Loss Regularization for Deep Learning Generalizability, p. 713–723. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-43901-8_68
12. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
13. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
14. Vaduganathan, M., Mensah, G.A., Turco, J.V., Fuster, V., Roth, G.A.: The global burden of cardiovascular diseases and risk: A compass for future health. Journal of the American College of Cardiology **80**(25), 2361–2371 (2022). https://doi.org/https://doi.org/10.1016/j.jacc.2022.11.005, https://www.sciencedirect.com/science/article/pii/S0735109722073120
15. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
16. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(12), 2933–2946 (2019)
17. Zhuang, X., Li, L., Payer, C., Stern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Medical image analysis **58**, 101537 (2019)
18. Zhuang, X., Shen, J.: Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. Medical Image Analysis **31**, 77–87 (2016)

# 5 Supplementary

## 5.1 Per-Class Five-Fold Cross-Validation Results

**Table 5.** Five-fold cross-validation dice scores for each class without post-processing with n = 86 (40 CT + 46 MRI).

| Model | LV | RV | LA | RA | Myo | AO | PA | AVG |
|---|---|---|---|---|---|---|---|---|
| MedNeXt-M-k3-DOMINO++ | 0.9455 | 0.9142 | 0.9318 | 0.9266 | 0.8963 | 0.8952 | 0.8633 | 0.9104 |
| MedNeXt-M-k3-DA-MM | 0.9458 | 0.9136 | 0.9303 | 0.9260 | 0.8951 | 0.8960 | 0.8605 | 0.9096 |
| MedNeXt-M-k3-DA5 | 0.9469 | 0.9118 | 0.9294 | 0.9241 | 0.8951 | 0.8967 | 0.8666 | 0.9101 |
| **ENSEMBLE** | 0.9477 | 0.9162 | 0.9323 | 0.9284 | 0.8982 | 0.9010 | 0.8663 | 0.9129 |

## 5.2 UpKern and External Validation Strategies

The UpKern technique in MedNeXt for training a $5 \times 5 \times 5$ kernel model from a preceding $3 \times 3 \times 3$ trained model was investigated. Training times for these models are knowingly expensive, nearing 700 seconds per epoch for 1000 epochs ($\approx$ 8 days) compared to a k3 model which measures nearly 250 seconds per epoch ($\approx$ 3 days). We applied the same techniques with a MedNeXt-M, DA5, and DOMINO++ model with UpKern and reported our results.

**Table 6.** Five-fold cross-validation dice scores for each class without post-processing with n = 86 (40 CT + 46 MRI).

| Model | LV | RV | LA | RA | Myo | AO | PA | AVG |
|---|---|---|---|---|---|---|---|---|
| MedNeXt-M-k5 | 0.9473 | 0.9141 | 0.9317 | 0.9269 | 0.8969 | 0.8972 | 0.8660 | 0.9115 |
| MedNeXt-M-k5-DA5 | 0.9464 | 0.9118 | 0.9297 | 0.9253 | 0.8925 | 0.9011 | 0.8691 | 0.9108 |
| MedNeXt-M-k5-DOMINO | 0.9528 | 0.9250 | 0.9389 | 0.9351 | 0.9081 | 0.9178 | 0.8832 | 0.9230 |
| **ENSEMBLE** | 0.9501 | 0.9194 | 0.9348 | 0.9315 | 0.9023 | 0.9065 | 0.8764 | 0.9173 |

For openness of our validation guidelines, we list our attempts used on the validation set in Table 7. We summarize our analysis as follows.

- The kernel 5 ensemble ranks as the absolute first on the leaderboard but to little gain over the kernel 3 ensemble (our final submission). Due to the significant increases in training time, we do not encourage the kernel 5 solution for this specific application. Moreover, there are indications of over-calibrating/over-fitting when comparing the training to the validation set.
- Ensembling MedNeXt-M configurations (M/M-DA5/M-DOMINO++) benefited more than ensembling across different model complexities (e.g. B/M/L).
- Ensembling provides an expected improvement when compared to individual model performance.

**Table 7.** All validation attempts by our team, all of lead at the top of the leaderboard against other teams, listed in descending order. Our final model is highlighted in red.

| Model | $CT_{DSC}$ | $CT_{HD}$ | $CT_{ASSD}$ | $MR_{DSC}$ | $MR_{HD}$ | $MR_{ASSD}$ |
|---|---|---|---|---|---|---|
| MedNeXt-M-k5/M-DA5/M-DOMINO++ | 0.9439 | 12.9392 | 0.6945 | 0.8959 | 17.283 | 1.2083 |
| MedNeXt-M-k3/M-DA5/DOMINO++ | 0.9440 | 12.8562 | 0.6984 | 0.8956 | 17.8029 | 1.2127 |
| MedNeXt-M-k3/DOMINO++/L | 0.9440 | 12.9787 | 0.7008 | 0.8951 | 17.9407 | 1.2228 |
| MedNeXt-M-k3 & L | 0.9440 | 12.9552 | 0.7002 | 0.8947 | 18.3741 | 1.2261 |
| MedNeXt-B-k3 & M & L | 0.9438 | 13.0632 | 0.7025 | 0.8950 | 17.9972 | 1.2227 |
| MedNeXt-M-k3-DOMINO++ | 0.9436 | 13.0978 | 0.7059 | 0.8950 | 17.8690 | 1.2261 |
| MedNeXt-M-k3-DA5 | 0.9427 | 13.0396 | 0.7114 | 0.8941 | 17.6121 | 1.2230 |
| MedNeXt-M-k5-DOMINO++ | 0.9430 | 13.3336 | 0.7063 | 0.8953 | 18.4014 | 1.2250 |
| MedNeXt-M-k3 | 0.9437 | 13.2161 | 0.7048 | 0.8944 | 18.3830 | 1.2279 |

## 5.3 DOMINO++

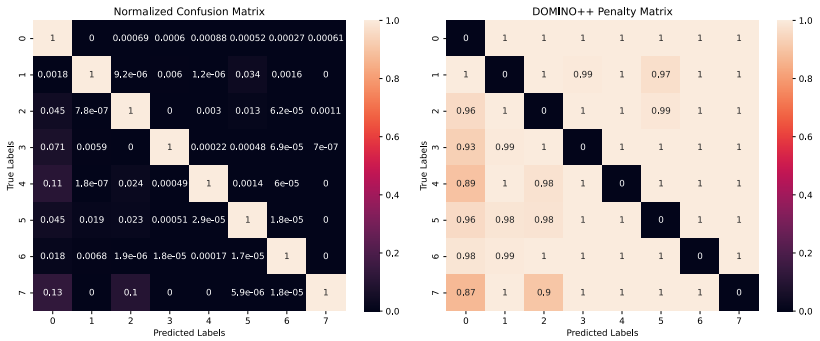


**Fig. 4.** Output confusion matrix from our MedNeXt-M-k3 model and DOMINO++ penalty matrix for a 7 (+1 background) class task of whole heart segmentation.

## 5.4 3D Segmentation Visualization

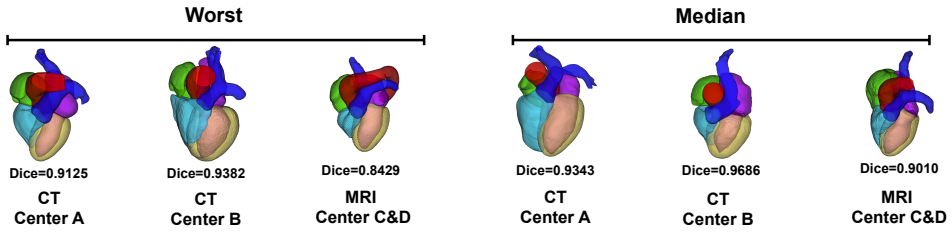


**Fig. 5.** 3D whole heart segmentation of the validation set. The ground truth is not known to the competitors.