

Extending nnU-Net is all you need

Fabian Isensee^{*1,2}, Constantin Ulrich ^{*1}, Tassilo Wald ^{*1,2}, and Klaus Maier-Hein^{1,2}

¹ Division of Medical Image Computing, German Cancer Research Center (DKFZ)
² Helmholtz Imaging

Abstract. Semantic segmentation is one of the most popular research areas in medical image computing. Perhaps surprisingly, despite its conceptualization dating back to 2018, nnU-Net continues to provide competitive out-of-the-box solutions for a broad variety of segmentation problems and is regularly used as a development framework for challenge-winning algorithms. Here we use nnU-Net to participate in the AMOS2022 challenge, which comes with a unique set of tasks: not only is the dataset one of the largest ever created and boasts 15 target structures, but the competition also requires submitted solutions to handle both MRI and CT scans. Through careful modification of nnU-net’s hyperparameters, the addition of residual connections in the encoder and the design of a custom postprocessing strategy, we were able to substantially improve upon the nnU-Net baseline. Our final ensemble achieves Dice scores of 90.13 for Task 1 (CT) and 89.06 for Task 2 (CT+MRI) in a 5-fold cross-validation on the provided training cases.

Keywords: Abdominal Multi-Organ Segmentation · nnU-Net · AMOS22.

1 Introduction

Automated delineation of all anatomical structures and pathologies in medical images is a long-standing goal in medical image computing. Due to the need of expert annotators and the time-intensive nature of 3D annotations, datasets have so far required careful balancing between the number of target structures and the number of training cases. Consequently, existing methods are either trained on many images and can robustly segment few structures or are trained on few images can segment many structures with reduced robustness. Thus, whenever a holistic perspective on a patients anatomy is required, multiple expert models must be pooled together from multiple sources. Not only does this increase the inference time, but it also creates new issues such as potentially conflicting predictions. Finally and perhaps most importantly, with each expert model being trained independently, label synergies cannot be exploited, thus potentially decreasing the label efficiency of the models as well as their robustness. In this context, the Abdominal Multi Organ Segmentation 2022 (AMOS2022) challenge [5], is set out to catalyze the development of holistic segmentation

^{*} equal contribution

methods. It comes with an unprecedented number of training images and annotated target structures: 15 organs of interest were labeled in 500 CT and 100 MRI scans, distributed into 200+40 (CT + MRI) training, 100+20 validation and 200+40 test images. The challenge poses two tasks: Task 1 is a classic multi-organ segmentation problem on just the CT images whereas Task 2 includes the MRI images and expects submitted methods to handle both modalities.

Within the context of medical image segmentation, nnU-Net [3] has stood the test of time. It consistently delivers state-of-the-art results on new segmentation datasets as they are released, despite its fully automated out-of-the-box nature. Moreover, nnU-Net was successfully used as a basis for task-specific method optimization, enabling not just us [4,1] but also many other teams [6] to win highly contested challenges. Thus, it seems only natural to use nnU-Net for our participation in the AMOS2022 challenge as well.

2 Method

nnU-Net is a framework that automatically configures and trains U-Net [7] based segmentation pipelines. Through rigorous analysis of the target dataset, nnU-Net makes automated adaptations to the patch size, batch size, preprocessing, network topology and more. For a full description of nnU-Net, we refer to [3].

In this section we propose several modifications to nnU-net’s automatically generated pipeline to maximize segmentation performance on the AMOS2022 challenge. Throughout method development we apply all modifications to both tasks with the sole difference being the intensity normalization scheme. For Task 1 we utilize nnU-Net’s ‘CT’ scheme (data-driven clipping and normalization) and for Task 2 we use simple z-scoring of all images (nnU-Net’s ‘nonCT’ setting).

2.1 Optimization of nnU-Net’s segmentation pipeline to AMOS2022

Starting from the default ‘3d_fullres’ configuration provided by nnU-Net we explore multiple improvements. We experimented with replacing the default encoder of the U-Net with a residual encoder (based on [2]). We furthermore optimized the preprocessing, specifically the batch size, patch size and target spacing. Advances in GPU memory capacity and processing speed allowed for larger models and batch sizes than the standard nnU-Net. Method development is performed by running 5-fold cross-validation on the provided training cases. All models are trained from scratch using the nnU-Net framework. Our experimentation resulted in three well-performing candidates for Task 1 and two candidates for Task 2, all of which are summarized in table 1. Fig. 1 shows the segmentation architectures used by our final configurations. They share the same topology but their feature map sizes differ due to nnU-Net’s automatic configuration of convolutional strides and kernel sizes as a function of the patch size.

Table 1. Final configurations used in our submission. Table highlights changes to the nnU-Net defaults.

Task	Name	Patch Size	Spacing [mm]	Data	Aug. batch size	norm.	Arch.
-	'3d_fullres'	[64,160,160]	[2,0.69,0.69]	default	2	CT/z-score	-
1	Configuration 1	[128,192,192]	[1.5,1,1]	DA5	5	CT	A2
	Configuration 2	[80,224,192]	[2,0.69,0.69]	default	6	CT	A1
	Configuration 3	[128,192,192]	[1.5,1,1]	default	5	CT	A2
2	Configuration 4	[80,224,192]	[2,0.69,0.69]	default	6	z-score	A1
	Configuration 5	[128,192,192]	[1.5,1,1]	default	5	z-score	A2

2.2 Inference strategy

Prediction is carried out with the nnU-Net defaults (sliding window). For validation and test set prediction we use ensembling, using both the 5 models from our cross-validation as well as multiple different configurations (3 configurations for Task 1: $3 \times 5 = 15$ models in the ensemble). Ensembling is implemented as simple averaging of softmax outputs.

2.3 Postprocessing

The postprocessing offered by nnU-net was designed to cover a wide variety of use-cases. We believe that additional performance can be gained on AMOS2022 by specifically analyzing and targeting failure cases of our method. To identify those we generated a confusion matrix on our cross-validation results (see supplementary information fig. 2) and performed rigorous visual inspection of our predicted segmentation maps.

Left-right confusion. Sometimes parts of the left kidney were classified as right kidney and vice-versa (same for adrenal glands). Here, removing disconnected small components is suboptimal because then parts of the kidneys would be labeled as background. Thus, we pooled the final kidney predictions into a joint class and use a connected component analysis to determine connected kidney regions. For each connected region we calculate its position within the image and assign its label (left/right) accordingly.

Connected component filtering. Just like the default nnU-Net we explore filtering of connected components. Specifically we determine whether removing all but the largest connected component can improve the Dice score.

Organ size constraints. Driven by the human anatomy, organs are expected to have a certain volume. This property can be exploited to filter small false positive predictions, for example in images in which the target organ is not present but false positive areas are. We use a parameter called 'rate' in combination with the minimum organ size observed in the training set. The rate is intended to be a safety margin: a rate of 0.75 indicates that components smaller than 75% of the minimum organ size are removed.

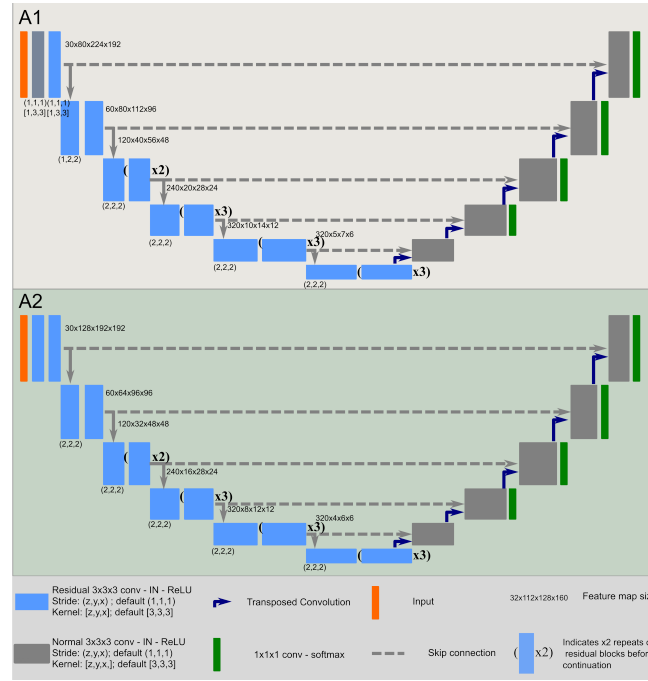


Fig. 1. The two architectures used by our final configurations. Due to nnU-Net’s automatic configuration of kernel sizes and strides as a function of the patch size, their feature map sizes differ. Both architectures make use of residual connections in the encoder.

Component filtering with size constraints. We combine the two previous techniques and remove instances below a volume threshold unless its the only connected region.

Composition of post-processing steps. In order to find the optimal post-processing scheme we create compositions of the previously mentioned post-processing schemes, namely:

1. PP1: Left right confusion (only kidneys & adrenal glands) followed by organ size constraints
2. PP2: Left right confusion (only kidneys & adrenal glands) followed by component filtering with size constraints
3. PP3: Left right confusion (only kidneys & adrenal glands) followed by connected component filtering.
4. PP4: Only left right confusion (only kidneys & adrenal glands).

We use the predictions of our ensemble on the 5-fold cross-validation (training images) to optimize the postprocessing scheme. Each organ is optimized independently and the best performing postprocessing strategy is retained. Note that we test multiple possible values for the ‘rate’ parameter where applicable. We refer to supplementary tables 4 and 5 for the final postprocessing configurations.

3 Results

The proposed modifications to the default nnU-Net pipeline substantially improved the results both on the training set cross-validation as well as the official validation set. We perform ablation studies to highlight the contributions of each component, see table 2. Overall, we were able to increase the Dice for configuration 3 (see table 1) by about 1% w.r.t. the nnU-Net baseline by modifying nnU-Net’s automated configuration (specifically, target spacing, patch size) and using residual connections in the encoder of the U-Net. Increasing the batch size yielded further improvements.

Table 2. Ablation results.

Method	Task 1		Task 2	
	5-fold CV	Val	5-fold CV	Val
	Dice	mean score (Dice score)	Dice	mean score (Dice score)
nnU-Net default	88.64	86.52 (90.31)	-	-
+ configuration improvements	89.08	-	-	-
+ residual encoder	89.45	-	88.43	-
+ increased batch size (bs 5) (corresponds to configs 3 & 5)	89.57	87.71 (91.43)	88.68	87.72 (90.97)

For the sake of brevity we do not show detailed results of the remaining configurations. We summarize the cross-validation performance of all our configurations and their ensembles in table 3. As expected, ensembling provided a substantial gain in segmentation performance, as did our postprocessing.

Table 3. 5-fold cross-validation results of all our configurations and their ensembles.

Task 1		Task 2	
Config.	Dice	Config.	Dice
Config. 1	89.60	Config. 4	88.56
Config. 2	89.59	Config. 5	88.69
Config. 3	89.57		
Ensemble	89.92	Ensemble	88.94
+ postprocessing	90.13	+ postprocessing	89.06

4 Discussion

In this paper we performed task-specific optimizations of the default nnU-Net pipeline to maximize segmentation performance on the AMOS2022 challenge. Changing crucial hyperparameters such as the patch size, batch size and target spacing for resampling yielded substantial gains relative to the default configuration, as did the addition of residual connections in the encoder of the U-Net. Our final submission consists of three configurations for Task 1 and two for Task 2. Since each configurations was trained as a 5-fold cross-validation, our ensembles consist of 15 and 10 models, respectively. At the time of submission, we ranked third in Task 1 and first in Task 2, although we should add that none of these submissions are reflected in this paper since we decided on short notice to deviate from our original plan and use larger ensembles instead (which according to our testing should take approx. 60h on Task 1 and 50h on Task 2 for inference on an RTX 3090). This decision was made after the validation set submission was closed and could thus only be evaluated on the training set cross-validation. Nonetheless, we are confident in the capabilities of our solution and hope for the improved performance we measured to translate into the test set. Naturally, the source code for training our models as well as the inference dockers will be made publicly available after the competition.

Acknowledgements

Part of this work was funded by Helmholtz Imaging, a platform of the Helmholtz Incubator on Information and Data Science.

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: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 238–249. Springer (2020)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2016-December, pp. 770–778. IEEE Computer Society (dec 2016)
3. 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)(2), 203–211 (2021)
4. Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnu-net for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 118–132. Springer (2020)
5. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023* (2022)
6. Ma, J.: Cutting-edge 3d medical image segmentation methods in 2020: Are happy families all alike? *arXiv preprint arXiv:2101.00232* (2021)

7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, Springer International Publishing (2015)

Appendix

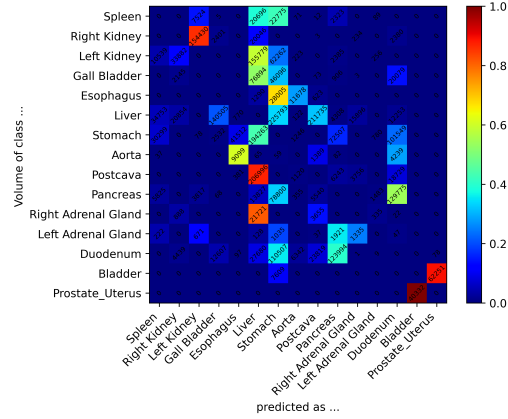


Fig. 2. Confusion of organs. The color encodes the relative percentage per row while the number represents the absolute volume over the entire cross validation dataset. We used this as an indicator which error cases exist and could be solved through postprocessing. The diagonal was removed to not influence percentage calculations.

Table 4. Postprocessing results for the different organs. We exemplarily show the results for our final ensemble on Task 1 as calculated on our cross validation. As previously explained, the rate indicates at which percentage of the minimum volume of the organ, a connected prediction of that organ gets removed.

Organ	Min. Organ Vol. [mm ³]	best PP method	rate	Dice	
				before PP	after PP
Spleen	14514.5	PP1	0.9	97.104	97.624
R. Kidney	32323.3	PP1	0.10	96.516	96.753
L. Kidney	10084.4	PP3	-	95.443	96.277
Gall Bladder	1242.1	PP1	0.25	86.014	87.360
Esophagus	418.9	PP1	0.25	85.480	85.481
Liver	653406.2	PP3	-	97.880	97.944
Stomach	30.6	PP1	0.9	91.580	91.579
Aorta	26481.4	PP1	0.75	95.901	96.036
Postcava	31401.8	PP1	0.10	91.745	91.755
Pancreas	15213.3	-	-	86.734	86.734
R. Adrenal Gland	1076.1	-	-	79.689	79.689
L. Adrenal Gland	663.7	PP1	0.1	80.996	81.010
Duodenum	19244.0	-	-	83.789	83.789
Bladder	14665.4	PP1	0.25	92.285	92.304
Prostate/Uterus	7937.5	PP1	0.25	87.610	87.610
Average				89.918	90.130

Table 5. Postprocessing results for the different organs. Equivalent for Task 2 to table 4.

Organ	Min. Organ Vol. [mm ³]	best PP method	rate	Dice	
				before PP	after PP
Spleen	14514.5	PP1	0.10	96.915	97.325
R. Kidney	32323.3	-	-	96.408	96.408
L. Kidney	10084.4	-	-	95.606	95.606
Gall Bladder	1242.1	PP1	0.5	85.981	86.387
Esophagus	418.9	PP1	0.95	82.792	82.798
Liver	653406.2	PP3	-	97.865	97.963
Stomach	30.6	PP1	0.25	91.161	91.161
Aorta	26481.4	PP2	0.9	95.392	95.495
Postcava	31401.8	PP1	0.10	91.252	91.273
Pancreas	15213.3	-	-	86.660	86.660
R. Adrenal Gland	1076.1	-	-	77.002	77.002
L. Adrenal Gland	663.7	-	-	78.007	78.007
Duodenum	19244.0	PP1	0.10	81.935	81.948
Bladder	14665.4	PP1	0.50	90.430	90.795
Prostate/Uterus	7937.5	PP1	0.25	86.658	87.121
Average				88.938	89.064