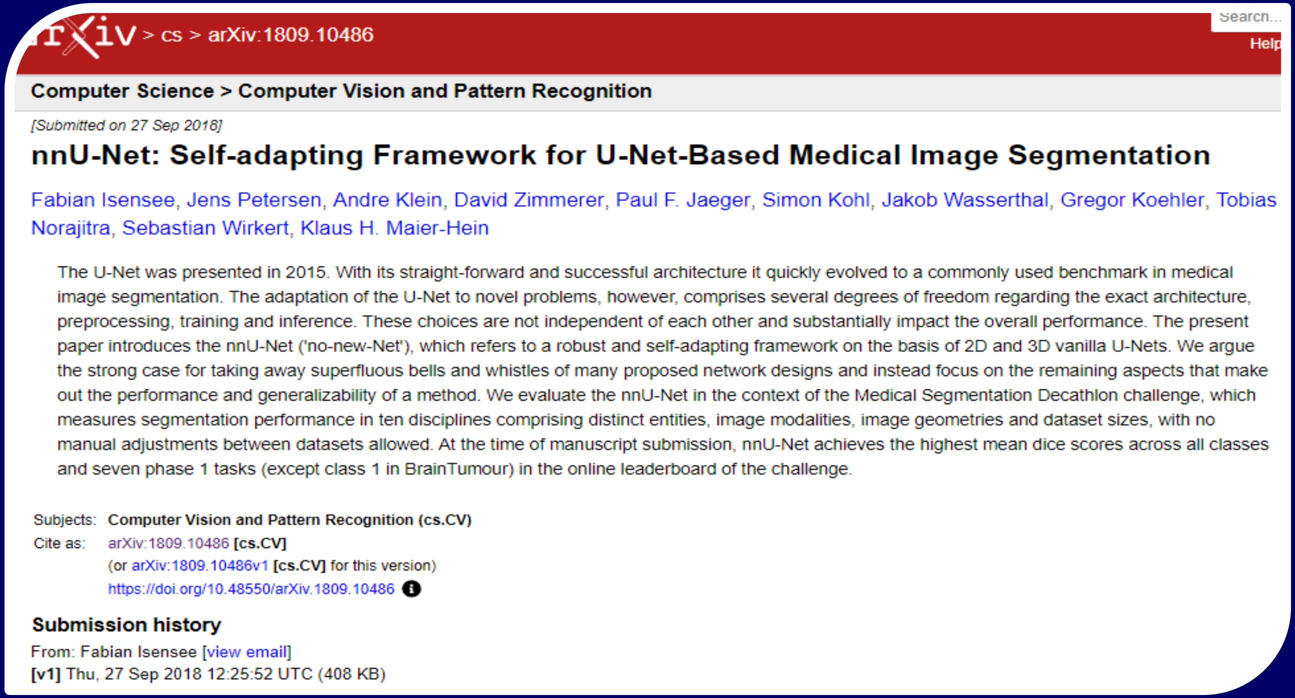
2024.07.26



Computer Science > Computer Vision and Pattern Recognition

[Submitted on 27 Sep 2018]

nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, Klaus H. Maier-Hein

The U-Net was presented in 2015. With its straight-forward and successful architecture it quickly evolved to a commonly used benchmark in medical image segmentation. The adaptation of the U-Net to novel problems, however, comprises several degrees of freedom regarding the exact architecture, preprocessing, training and inference. These choices are not independent of each other and substantially impact the overall performance. The present paper introduces the nnU-Net ('no-new-Net'), which refers to a robust and self-adapting framework on the basis of 2D and 3D vanilla U-Nets. We argue the strong case for taking away superfluous bells and whistles of many proposed network designs and instead focus on the remaining aspects that make out the performance and generalizability of a method. We evaluate the nnU-Net in the context of the Medical Segmentation Decathlon challenge, which measures segmentation performance in ten disciplines comprising distinct entities, image modalities, image geometries and dataset sizes, with no manual adjustments between datasets allowed. At the time of manuscript submission, nnU-Net achieves the highest mean dice scores across all classes and seven phase 1 tasks (except class 1 in BrainTumour) in the online leaderboard of the challenge.

Subjects:	**Computer Vision and Pattern Recognition (cs.CV)**
Cite as:	arXiv:1809.10486 **[cs.CV]**
	(or arXiv:1809.10486v1 **[cs.CV]** for this version)
	https://doi.org/10.48550/arXiv.1809.10486 ℹ

Submission history

From: Fabian Isensee [view email]
[v1] Thu, 27 Sep 2018 12:25:52 UTC (408 KB)

U-Net; Medical Image Segmentation;
self-adapting framework

Yeon Su Park

Pukyong National University
Division of Computer Engineering And Artificial Intelligence
Medical AI Lab.

Part 1,

Introduction

U-Net and Challenge

- Medical Image Segmentation is currently dominated by deep CNNs.
- U-Net commonly used benchmark in medical image segmentation.
- However, each segmentation benchmark seems to require specialized architectures and training scheme modifications to achieve competitive performance.
- This challenge is designed to address the problem. And finally participants create a segmentation algorithm that generalizes across 10 datasets corresponding to different entities of the human body.
- Hypothesis : some of the architectural modifications presented recently are in part overfitted to specific problems or could suffer from imperfect validation that results from sub-optimal reimplementations of the SOTA.

Introduction

Underestimated aspect of segmentation algorithm

- Even though the architecture is quite straight-forward, and even though the method is quite commonly used as a benchmark, the remaining interdependent choices regarding the exact architecture, preprocessing, training, inference and post-processing quite often cause the U-Net to underperform when used as a benchmark.

- The influence of non-architectural aspects in segmentation methods is much more impactful, and also underestimated.

- In nnU-Net(no-new-Net) there is a set of three comparatively simple U-Net models that contain only minor modifications to the original U-Net.

- nnU-Net automatically adapts its architectures to the given image geometry.

- the nnU-Net framework thoroughly defines all the other steps around them like preprocessing, training, inference, potential post-processing.

Methods

Network architecture

- Because of Medical Image commonly made of 3D, the framework proposes three Basic U-Net which is 2D U-Net, 3D U-Net, and U-Net Cascade.

- The architectural modification is similar to the original U-Net and instead focuses on designing an automatic training pipeline for these models.

- What is U-Net Cascade?
  - In real world there are limitations when dealing with big 3D data like liver because GPU memory is small.
  - To address this problem 3D U-Net is first trained on downsampled images.
  - The segmentation results of this U-Net are then upsampled to the original voxel spacing and passed as additional (one hot encoded) input channels to a second 3D U-Net, which is trained on patches at full resolution.

Part 2,

Methods

Convolutional Block

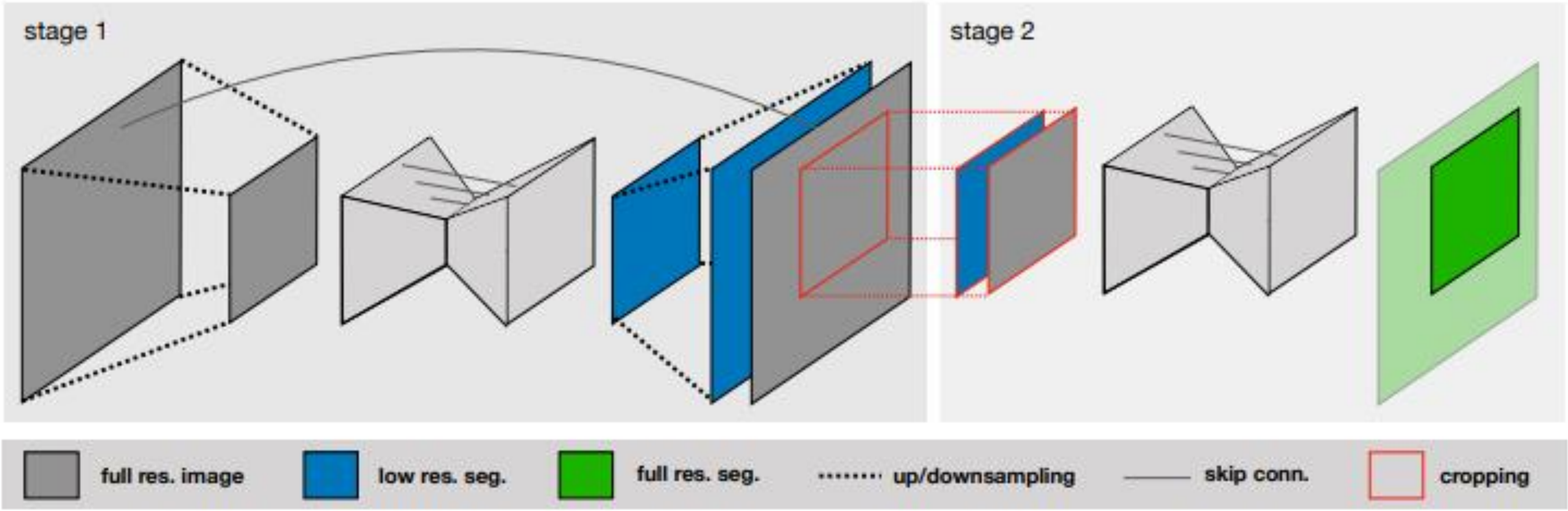


Fig. 1. U-Net Cascade (on applicable datasets only). Stage 1 (left): a 3D U-Net processes downsampled data, the resulting segmentation maps are upsampled to the original resolution. Stage 2 (right): these segmentations are concatenated as one-hot encodings to the full resolution data and refined by a second 3D U-Net.

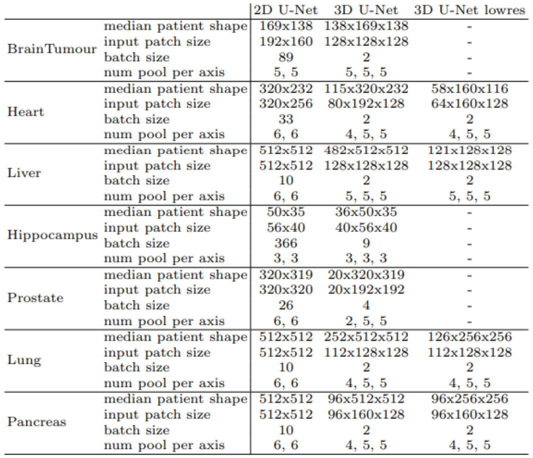
Part 2,

Methods

Dynamic adaptation of network topologies

- Due to the large differences in image size (median shape 482 × 512 × 512 for Liver vs. 36 × 50 × 35 for Hippocampus) the input patch size and number of pooling operations per axis must be automatically adapted for each dataset to allow for adequate aggregation of spatial information.

	2D U-Net	3D U-Net
input patch size	256^2	Below 128^3
batch size	42	2
feature maps in the highest layers	30	30



		2D U-Net	3D U-Net	3D U-Net lowres
BrainTumour	median patient shape	169x138	138x169x138	-
	input patch size	192x160	128x128x128	-
	batch size	89	2	-
	num pool per axis	5, 5	5, 5, 5	-
Heart	median patient shape	320x232	115x320x232	58x160x116
	input patch size	320x256	80x192x128	64x160x128
	batch size	33	2	2
	num pool per axis	6, 6	4, 5, 5	4, 5, 5
Liver	median patient shape	512x512	482x512x512	121x128x128
	input patch size	512x512	128x128x128	128x128x128
	batch size	10	2	2
	num pool per axis	6, 6	5, 5, 5	5, 5, 5
Hippocampus	median patient shape	50x35	36x50x35	-
	input patch size	56x40	40x56x40	-
	batch size	366	9	-
	num pool per axis	3, 3	3, 3, 3	-
Prostate	median patient shape	320x319	20x320x319	-
	input patch size	320x320	20x192x192	-
	batch size	26	4	-
	num pool per axis	6, 6	2, 5, 5	-
Lung	median patient shape	512x512	252x512x512	126x256x256
	input patch size	512x512	112x128x128	112x128x128
	batch size	10	2	2
	num pool per axis	6, 6	4, 5, 5	4, 5, 5
Pancreas	median patient shape	512x512	96x512x512	96x256x256
	input patch size	512x512	96x160x128	96x160x128
	batch size	10	2	2
	num pool per axis	6, 6	4, 5, 5	4, 5, 5

Table 1. Network topologies as automatically generated for the seven phase 1 tasks of the Medical Segmentation Decathlon challenge. 3D U-Net lowres refers to the first stage of the U-Net Cascade. The configuration of the second stage of the U-Net Cascade is identical to the 3D U-Net.

Part 2,

Methods

Preprocessing

- **Cropping**
  - All data is cropped to the region of nonzero values. It can reduce size and computational burden.

- **Resampling**
  - CNNs do not natively understand voxel spacings.
  - To enable networks to properly learn spatial semantics, all patients are resampled to the median voxel spacing of their respective dataset, where third order spline interpolation is used for image data and nearest neighbor interpolation for the corresponding segmentation mask.
  - If the median shape of the resampled data has more than 4 times the voxels that can be processed as input patch by the 3D U-Net then it qualifies for the U-Net Cascade and this dataset is additionally resampled to a lower resolution.
  - If the dataset is anisotropic, the higher resolution axes are first downsampled until they match the low resolution axis/axes and only then all axes are downsampled simultaneously.

- **Normalization**
  - CT - [0.5 -0.95]percentiles of intensity z-score normalization
  - Others - just z-score normalization
  - Z-score normalization = (x – μ) / σ

Methods

Training Procedure

- **Five-fold cross validation**
- **Loss function is combination of dice and cross entropy** $\mathcal{L}_{total} = \mathcal{L}_{dice} + \mathcal{L}_{CE}$
- **Dice loss -** where u is the **softmax** output of the network and v is a **one hot encoding** of the ground truth segmentation map. Both u and v have shape I × K with i ∈ I being the number of pixels in the training patch/batch and k ∈ K being the classes

$$\mathcal{L}_{\mathrm{dc}} = -\frac{2}{|K|} \sum_{k \in K} \frac{\sum_{i \in I} u_i^k v_i^k}{\sum_{i \in I} u_i^k + \sum_{i \in I} v_i^k}$$

- **Optimizer** – Adam(lr=0.003, when Whenever validation and training losses's EMA dosen't improve 0.005 within 30 epochs, lr is reduced by factor 5.)
- **Epoch** – 250
- **Early stop** – Whenever if EMA dosen't improve 0.005 within 60 epochs
- **Data Agumentation -** according to github.com/MIC-DKFZ/batchgenerators and after first stage of U-Net Cascade : random morphological operators (erode, dilate, open, close) and randomly remove connected components of these segmentations.
- **Patch Sampling -** To increase the stability of network training we enforce that more than a third of the samples in a batch contain at least one randomly chosen foreground class.

Part 2,

Methods

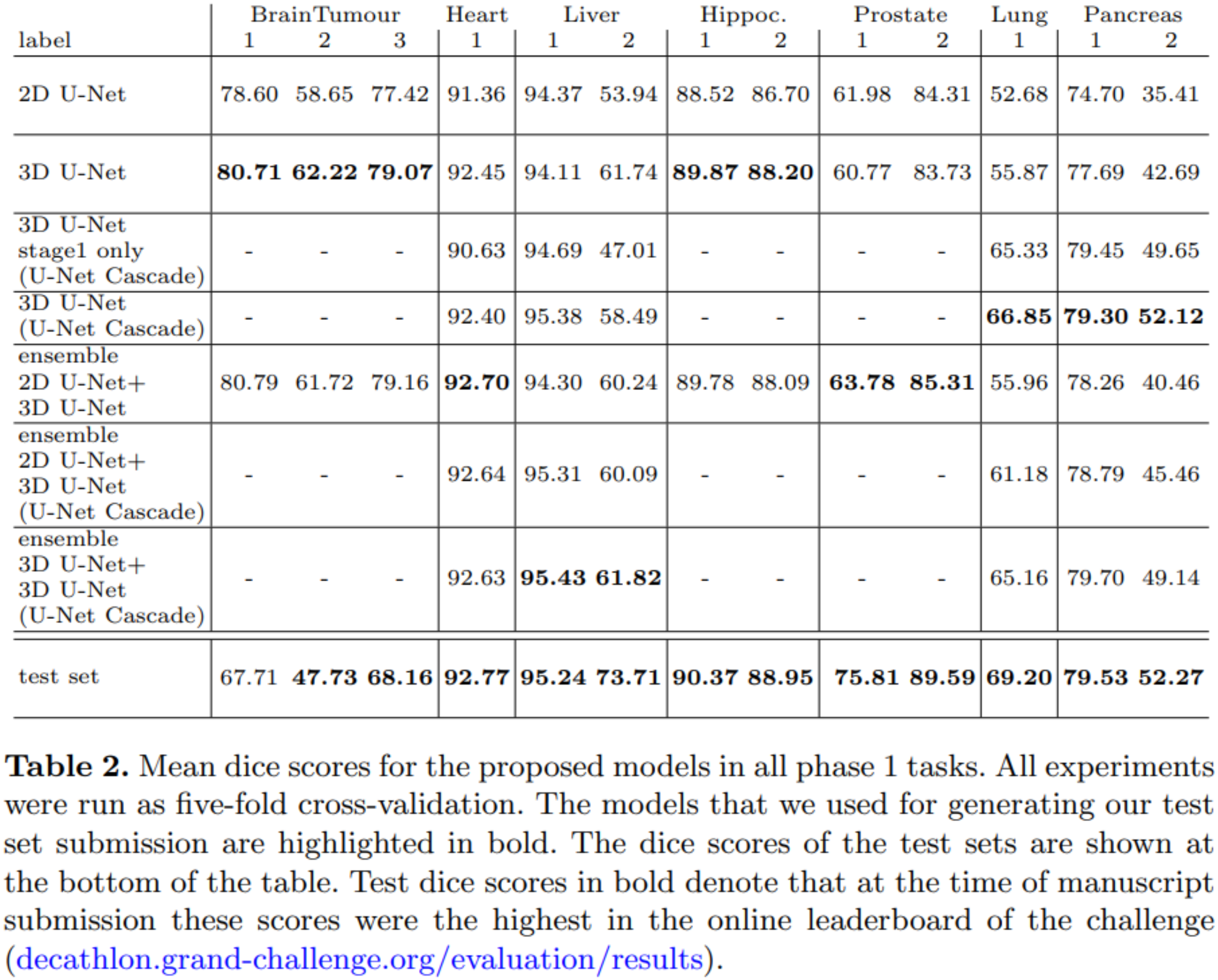
Inference, Post-processing, Ensembling and Submission

- **Inference**
  - When aggregating predictions across patches The closer to the center, the more weight was given.
  - Patches are chosen to overlap by patch size / 2 and we further make use of test time data augmentation by mirroring all patches along all valid axes.
  - Combining the tiled prediction and test time data augmentation result in segmentations where the decision for each voxel is obtained by aggregating up to 64 predictions (in the center of a patient using 3D U-Net).
- **Post-processing**
  - A connected component analysis of all ground truth segmentation labels is performed on the training data.
  - the largest connected component for this class are automatically removed on predicted images of the corresponding dataset.
- **Ensembling and Submission**
  - All possible combinations of two out of three of our models are ensembled for each dataset.
  - the model (or ensemble) that achieves the highest mean foreground dice score on the training set cross-validation is automatically chosen.

Part 3,

Experiments and Results

List of dataset and preprocessing



	BrainTumour			Heart	Liver		Hippoc.		Prostate		Lung	Pancreas	
label	1	2	3	1	1	2	1	2	1	2	1	1	2
2D U-Net	78.60	58.65	77.42	91.36	94.37	53.94	88.52	86.70	61.98	84.31	52.68	74.70	35.41
3D U-Net	**80.71**	**62.22**	**79.07**	92.45	94.11	61.74	**89.87**	**88.20**	60.77	83.73	55.87	77.69	42.69
3D U-Net stage1 only (U-Net Cascade)	-	-	-	90.63	94.69	47.01	-	-	-	-	65.33	79.45	49.65
3D U-Net (U-Net Cascade)	-	-	-	92.40	95.38	58.49	-	-	-	-	**66.85**	**79.30**	**52.12**
ensemble 2D U-Net+ 3D U-Net	80.79	61.72	79.16	**92.70**	94.30	60.24	89.78	88.09	**63.78**	**85.31**	55.96	78.26	40.46
ensemble 2D U-Net+ 3D U-Net (U-Net Cascade)	-	-	-	92.64	95.31	60.09	-	-	-	-	61.18	78.79	45.46
ensemble 3D U-Net+ 3D U-Net (U-Net Cascade)	-	-	-	92.63	**95.43**	**61.82**	-	-	-	-	65.16	79.70	49.14
test set	67.71	**47.73**	**68.16**	**92.77**	**95.24**	**73.71**	**90.37**	**88.95**	**75.81**	**89.59**	**69.20**	**79.53**	**52.27**

Table 2. Mean dice scores for the proposed models in all phase 1 tasks. All experiments were run as five-fold cross-validation. The models that we used for generating our test set submission are highlighted in bold. The dice scores of the test sets are shown at the bottom of the table. Test dice scores in bold denote that at the time of manuscript submission these scores were the highest in the online leaderboard of the challenge (decathlon.grand-challenge.org/evaluation/results).

Part 4,

Disscusion

nnU-Net

- nnU-Net is framework for the medical domain that directly builds around the original U-Net architecture and dynamically adapts itself to the specifics of any given dataset.
- non-architectural modifications can be **much more powerful** than some of the recently presented architectural modifications, the essence of this framework is a thorough design of adaptive pre-processing, training scheme and inference.
- nnU-Net performs competitively on the held-out test sets of 7 highly distinct medical datasets, achieving the highest mean dice scores for all classes of all tasks (except class 1 in the BrainTumour dataset)
- In this paper, design choices such as the use of Leaky ReLUs instead of regular ReLUs and the parameters of our data augmentation were not properly validated.
- **Future work** will focus on systematically evaluating all design choices via ablation studies.