Robust Resolution-Enhanced Prostate Segmentation in Magnetic Resonance and Ultrasound Images through Convolutional Neural Networks

Oscar J. Pellicer-Valero [1,*], Victor Gonzalez-Perez [2], Juan Luis Casanova Ramón-Borja [3], Isabel Martín García [2], María Barrios Benito [2], Paula Pelechano Gómez [2], José Rubio-Briones [3], María José Rupérez [4] and José D. Martín-Guerrero [1]

1 Intelligent Data Analysis Laboratory, Department of Electronic Engineering, ETSE (Engineering School), Universitat de València (UV), Av. Universitat, sn, 46100 Bujassot, Valencia, Spain; jose.d.martin@uv.es
2 Department of Medical Physics, Fundación Instituto Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009 Valencia, Spain; vgonzalez@fivo.org (V.G.-P.); mismaga99@gmail.com (I.M.G.); mar7esc@gmail.com (M.B.B.); ppelechano@hotmail.com (P.P.G.)
3 Department of Urology, Fundación Instituto Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009 Valencia, Spain; jcasanova@fivo.org (J.L.C.R.-B.); jrubio@fivo.org (J.R.-B.)
4 Centro de Investigación en Ingeniería Mecánica (CIIM), Universitat Politècnica de València (UPV), Camino de Vera, sn, 46022 Valencia, Spain; mjrupere@upvnet.upv.es
* Correspondence: Oscar.Pellicer@uv.es; Tel.: +34-9635-44022

Featured Application: The proposed model has the potential of having a significant impact on current prostate procedures, undercutting, and even eliminating, the need of manual segmentations through improvements in terms of robustness, generalizability and output resolution.

Abstract: Prostate segmentations are required for an ever-increasing number of medical applications, such as image-based lesion detection, fusion-guided biopsy and focal therapies. However, obtaining accurate segmentations is laborious, requires expertise and, even then, the inter-observer variability remains high. In this paper, a robust, accurate and generalizable model for Magnetic Resonance (MR) and three-dimensional (3D) Ultrasound (US) prostate image segmentation is proposed. It uses a densenet-resnet-based Convolutional Neural Network (CNN) combined with techniques such as deep supervision, checkpoint ensembling and Neural Resolution Enhancement. The MR prostate segmentation model was trained with five challenging and heterogeneous MR prostate datasets (and two US datasets), with segmentations from many different experts with varying segmentation criteria. The model achieves a consistently strong performance in all datasets independently (mean Dice Similarity Coefficient -*DSC*- above 0.91 for all datasets except for one), outperforming the inter-expert variability significantly in MR (mean *DSC* of 0.9099 vs. 0.8794). When evaluated on the publicly available Promise12 challenge dataset, it attains a similar performance to the best entries. In summary, the model has the potential of having a significant impact on current prostate procedures, undercutting, and even eliminating, the need of manual segmentations through improvements in terms of robustness, generalizability and output resolution.

Keywords: MR prostate imaging; US prostate imaging; convolutional neural network; prostate segmentation; neural resolution enhancement

1. Introduction

In the field of medical imaging, segmentations are extremely useful for a plethora of tasks, including image-based diagnosis, lesion detection, image fusion, surgical planning or computer-aided surgery. For the prostate, in particular, fusion-guided biopsy and focal therapies are quickly gaining popularity due to the improved sensitivity and specificity for

lesion detection [1], and the low complication profile [2], respectively, although they are still not fully accepted in clinical guidelines.

Nevertheless, accurate prostate segmentations are still hard and laborious to obtain, since they have to be manually annotated by expert radiologists and, even then, the inter- and intra-observer variability may be significant due to factors such as the lack of clear boundaries between neighboring tissues or the huge size and texture variation of this gland among patients. In our experiments, 14 images were segmented by a second expert, and the inter-expert agreement was found to be of 0.8794 in terms of Sørensen-Dice Similarity Coefficient (*DSC*) and 1.5619 mm in terms of the Average Boundary Distance (ABD). Very similar results were obtained in [3], with experts achieving a *DSC* and an ABD of 0.83 and 1.5 mm, respectively. Because of this, automatic segmentation algorithms for the prostate are increasingly sought-after.

Before the rise of Deep Learning (DL) around 2012 [4], many different techniques for automatic prostate segmentation coexisted. For instance, in [5], MR images of the prostate were segmented by using voxel threshold-based classification followed by 3D statistical shape modeling. An alternative approach suggested by [6] attempted to match the probability distributions of the photometric variables inside the object of interest with an appearance model, and then evolved the shape of the object until both distributions matched best. Another technique that has been widely used in the literature for medical image segmentation is atlas matching, which consists in non-rigidly registering a set of labeled atlas images to the image of interest, and then somehow combining all resulting segmentations into a single one [7].

Despite the strengths of these methods, the true revolution in this field came with the advent of CNNs, which are a kind of DL algorithm formed by a stack of convolutional filters and non-linear activation functions, wherein the filter parameters are learned by stochastic gradient descent. New CNN architectures have been steadily raising state-of-the-art performance in computer vision tasks, such as image classification [8], image segmentation [9] or object detection [10]. Similarly, this trend has carried over to medical imaging, and prostate segmentation in particular.

One of the first approaches to CNN-based segmentation consisted in sliding a classification CNN over a whole image to provide pixel-wise classifications, which then were combined into a single segmentation mask [11]. Shortly after, fully convolutional neural networks for semantic segmentation were proposed [12]; they allowed for much faster training and inference, as the whole image was processed at once, and also made a better use of spatial information by utilizing activation maps from different layers. Later, the U-net architecture [9] introduced the encoder-decoder design with skip connections that is still predominantly used. In a U-net, the image is first processed through an encoder CNN, which is similar to a classification CNN. The output of the encoder is then connected to the input of the decoder CNN, which is an inverted version of the encoder where the pooling operators have been exchanged by up-scaling convolutions. Additionally, skip connections transfer information from the encoder to the decoder at several stages other than the output. This idea, in combination with residual connections [13] and a cost function based on *DSC*, was quickly extended from two-dimensional (2D) convolutions to 3D convolutions by the V-net [14], in order to better deal with 3D medical images. Similar to the transition from pixel-wise classification to fully convolutional CNNs, 3D-CNNs are better able to use the context of the whole image and provide faster speeds in comparison with a per-slice 2D segmentation.

In the field of prostate segmentation in MR imaging, many different CNN architectures building on top of the V-net or the U-net have been proposed. For instance, ref. [15] proposed the addition of deep supervision, ref. [16] used a more recent densenet-resenet architecture [13,17,18] introduced a boundary-aware cost function. Prostate segmentation in 3D Ultrasound (US) imaging, although much less prevalent, has experienced a similar development, with a recent paper employing the attention mechanism to exploit the

information from several layers [19]. Some of these architectural choices, and several others, will be further elaborated in Section 2.

Despite the high performances reported by many of the aforementioned papers, it could be argued that they all incur in a common pitfall: they are designed to perform well on one single prostate dataset. Therefore, it is unknown how robust the model would be when applied to any other dataset. This kind of robustness is paramount if the model is to be applied in a real-life scenario, where the images may come from many different scanners, and may be analyzed by many different experts. Furthermore, robustness is also desirable in the sense that the produced segmentations should be accepted by different experts, despite their possibly varying criteria for segmentation; in other words, the produced segmentations should ideally behave like an average prediction from several experts.

In this paper, a robust algorithm based on CNNs for MR and US prostate image segmentation is proposed. It leverages both common and not-so-common design choices, such as a hybrid densenet-resnet architecture (Section 2.3), deep supervision (Section 2.4), 3D data augmentation (Section 2.5), a cyclic learning rate (Section 2.7), checkpoint ensembling (Section 2.8) and a simple yet effective post-processing technique to increase the resolution of the segmentations known as Neural Resolution Enhancement (Section 2.9). This technique, besides improving the segmentation performance, allows the CNN to produce segmentations with resolutions beyond that of the original image. Furthermore, the model is trained on five different datasets simultaneously (Section 2.1), achieving an excellent performance on all of them. Finally, the weights obtained from this model are used to train (through transfer learning) an US segmentation model on two different datasets, achieving also an excellent performance on them both. Results are presented both quantitatively (Section 3.1), by presenting the metrics (Section 2.10) achieved on every dataset, and qualitatively (Section 3.2), by showing images of the predicted and Ground Truth (GT) segmentations on several patients. The paper is closed by a discussion, (Section 4), about the clinical impact of the proposed model, and a brief conclusion (Section 5).

2. Materials and Methods

2.1. Description of the Datasets

One of the main strengths of this study is the use of five different prostate T2-weighted MR datasets. As shown in Table 1, there is a significant variability in scanner manufacturers, resolutions and magnetic field strengths, among other factors. Datasets "Girona" [20], "Promise12" [21] and "Prostate-3T" [22] are all freely available for download on the Internet, while "IVO" comes from the Valencian Institute of Oncology, and "Private" comes from a private institution which has decided to remain anonymous. Furthermore, Promise12 is an ongoing prostate segmentation challenge, wherein 50 MR prostate images are provided along with their segmentation masks (dataset "Promise12"), and 30 additional images are provided without segmentations as a test set (dataset "Promise12_test"). The participants must submit their predictions to the challenge server, where they are evaluated. Hence, "Promise12_test" will only be used for testing.

In addition to that, the prostate segmentations follow varying criteria depending on the expert who segmented them. In "IVO" dataset, three different radiologists with two, five and seven years of experience in prostate cancer imaging took turns to perform the segmentations. In "Private" dataset, a single medical physicist with two years of experience in MR prostate imaging segmented all the images. In "Promise12", each of the four rows in Table 1 corresponds to a different medical center and, by extension, were also segmented by at least one different expert each, although an expert from the Promise12 challenge [21] corrected some of them. For the other datasets, no further information about the segmentations is known.

Table 1. Details of the MR datasets.

Dataset	N	Scanner Manufacturer (%)	Endorectal Coil	Pixel Spacing (mm)	Slice Spacing (mm)	Field Strength (T)
Girona	34	G. Elec. (59%)	No	0.27–0.55	1.00	1.5
		Siemens (41%)	No	0.68–0.79	1.00	3.0
Promise12	48	Siemens (25%)	Yes	0.63	3.6	1.5
		G. Elec. (25%)	Yes	0.25	2.20–3.00	3.0
		Siemens (25%)	No	0.33–0.63	3.00–3.60	1.5 & 3.0
		Siemens (25%)	No	0.50–0.75	3.60–4.00	3.0
Promise12_test	30	Unknown	Yes & No	0.27–0.63	2.2–3.6	1.5 & 3.0
Prostate-3T	12	Siemens	No	0.60–0.62	3.60–4.00	3.0
IVO	280	G. Elec. (96%)	No	0.35–0.74	0.60–7.00	1.5
		Philips (3%)	No	0.28–0.49	3.00	1.5 & 3.0
		Siemens (1%)	No	0.62–0.69	3.00–3.50	1.5
Private	90	Philips (81%)	No	0.30–0.62	2.91–5.00	1.5 & 3.0
		Siemens (11%)	No	0.52–0.69	3.00–3.60	1.5 & 3.0
		G. Elec. (8%)	No	0.37–0.43	3.40–6.00	1.5 & 3.0

Regarding exclusion criteria, before separating the images into any subsets, all segmentations were examined and those with obvious errors were directly excluded. Therefore, no corrections were made, so as to better preserve the particular criteria from each expert (except for the "Private" dataset, in which all segmentations were revised). The number of samples (N) in Table 1 is computed after this filtering. As a special mention, 18 images from "Prostate-3T", which were also present in "Promise12" or "Promise12_test" (although with different GT segmentations), were also discarded; and other 30 images from "Prostate-3T" (half of the original dataset), which systematically left many slices in the base and apex unsegmented, had to be discarded as well. Figure 1 shows the center slice of a sample from each of the datasets.

For the 3D-US segmentation model, two different datasets were employed: "IVO" and "Private", both coming from the same institutions as their homonymous MR datasets. For "IVO" ($N = 160$ images), five different urologists with six to thirty years of experience segmented the images, while for "Private" ($N = 82$ images), it was two urologists with more than ten years of experience; no exclusion criteria were applied. Images from both datasets were captured using Hitachi scanners at spacings of 0.20 mm to 0.41 mm in any axis. Figure 1 shows the center slice of a sample from each axis. Unfortunately, no further segmented datasets were found on the Internet for this image modality.
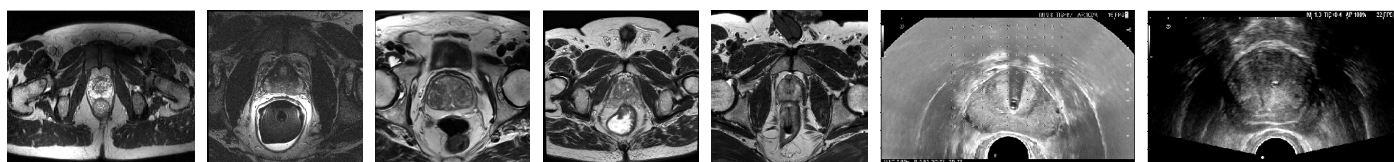
Figure 1. Center slice of a sample from MR and US datasets (from left to right): "Girona", "Promise12", "Prostate-3T", "IVO (MR)", "Private (MR)", "IVO (US)" and "Private (US)".

2.2. Image Pre-Processing

Before using the images to train the CNN, they all had to be pre-processed to alleviate their heterogeneity. First, their intensity was normalized by applying Equation (1) to every image I, such that 98% of the voxels in I_{new} fall within the range $[0, 1]$.

$$I_{new} = \frac{I - percentile(I, 1)}{percentile(I, 99) - percentile(I, 1)} \tag{1}$$

Then, the center crop of each image (and its respective segmentation mask) was taken, using a size of $112 \times 112 \times 32$ and a spacing of (1, 1, 3) mm for the MR images, and a size of $160 \times 112 \times 80$ and a spacing of (0.75, 0.75, 0.75) mm for the US images. B-Spline interpolation of third order was employed for all image interpolation tasks, while Gaussian label interpolation was used for the masks.

2.3. Hybrid Densenet-Resnet Architecture

The proposed CNN architecture (Figure 2) is based on the V-Net and, more precisely, on the architecture proposed by [16], which combines a densenet [17] encoder with a resnet [13] decoder. All design decisions were guided by validation results.
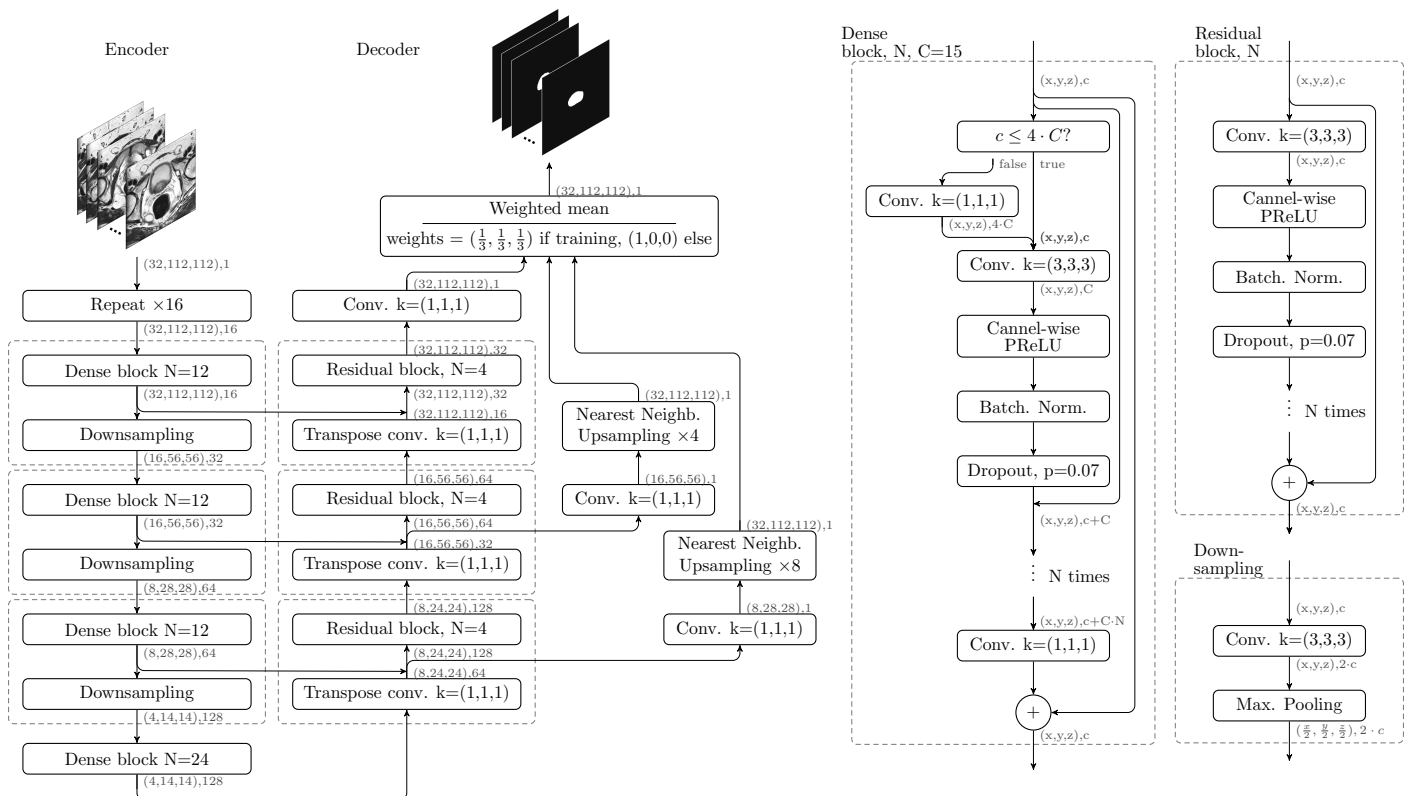
Figure 2. Architecture of the CNN. The encoder is composed of four Dense Blocks connected by Downsampling blocks. The decoder uses three Residual Blocks connected by transpose convolutions. Several skip connections transfer information from the encoder to the decoder. Furthermore, to the right of the decoder, intermediate outputs are used to perform Deep Supervision.

The full architecture is sufficiently described in Figure 2. Therefore, only a few interesting design choices will be discussed here. Firstly, the proposed Dense block includes a residual connection, which empirically helped the CNN converge faster. Secondly, every Dense block contains between 12 and 24 "standard" convolutions (kernel of size (3, 3, 3)), as well as several "bottleneck" convolutions (kernel of size (1, 1, 1)), for a total of 72 "standard" convolutions, which is a huge number compared to similar architectures such as V-net (with only 12 convolutions in its resnet-based encoder) or BOWDA-net [23] (with 28 convolutions in its densenet-based encoder). This makes the encoder better capable of learning more complex representations of the input data. Comparatively, the decoder can have a simpler resnet architecture, since the heavy lifting (which is feature extraction) has already been done by the encoder. Thirdly, channel-wise PReLU was employed as activation function [24], as it provides a slightly better performance at a negligible additional computational cost. A channel-wise PReLU function is similar to a ReLU (Rectified Linear Unit) [25] function, but with a learnable slope α for the negative inputs (instead of being just zero); α is shared among all activations in a channel, but is

different for every channel. Fourthly, transpose convolutions were used in the decoder, since they were found to provide a better performance when compared to upsampling followed by a convolution.

Due to the huge memory requirements intrinsic to the densenet architecture, very small batch sizes had to be employed (4 for the MR dataset, and 2 for the US dataset), as well as a technique known as Gradient Checkpointing [26], which allows to reduce the Graphics Processing Unit (GPU) memory requirements at the cost of increased computation times. It works by keeping a fraction of the CNN activations in memory at any given time (instead of all of them), and recomputing the rest when they are needed.

2.4. Deep Supervision

To further improve the performance of the CNN, a simple implementation of Deep Supervision [27] is used. Unlike regular CNNs, which predict the segmentation mask from the last layer only, deeply supervised CNNs attempt to predict it from several intermediate layers as well. In Figure 2 this is implemented by the branches to the right of the decoder, which take the activation maps at two points along the decoder, reduce the number of channels to one by means of a "bottleneck" convolution, and then upsample them to the CNN output resolution using Nearest Neighbors interpolation. During training, the final output of the CNN is averaged with these intermediate predictions while, during inference, only the final output is considered. A similar implementation for this technique is also successfully used by [15]. Figure 3 shows the GT mask of a prostate MR image, as well as the final and intermediate predictions, which are used for Deep Supervision. As it can be seen, intermediate predictions resemble a downscaled version of the final mask.

Figure 3. From left to right: first intermediate prediction, second intermediate prediction, third (and final) prediction, and original MR prostate image with GT label.

As demonstrated in [27], Deep Supervision serves a twofold purpose: on one hand, it forces all the layers throughout the network to learn features which are directly useful for the task of image segmentation; on the other hand, the gradients are better able to flow towards the deeper layers, which accelerates training, and helps prevent problems related to gradient vanishing.

2.5. Online Data Augmentation

Online data augmentation was used to artificially increase the amount and variability of the training images, thus improving the generalization capabilities of the model and, ultimately, its performance. Before feeding an image to the CNN during training, the following transformations were sequentially applied to it:

1. 3D rotation along a random axis with random magnitude in the range $[0, \pi/20]$ radians.
2. 3D shift of random magnitude in the range $[0, 15]$ mm along every axis.
3. 3D homogeneous scaling of random magnitude in the range $[1/1.15, 1.15]$ times.
4. Flipping along x-axis with probability $1/2$.
5. Adding Normally distributed noise with a random magnitude in the range $[0, 0.05]$ relative to the normalized image.

When required, a random number would be sampled from the uniform distribution and then scaled and shifted to the appropriate range.

2.6. Model Training

For both the MR and US segmentation models, images were split into three subsets: training (70% of the images), validation (15%) and test (15%). These proportions were computed dataset-wise, such that the relative representation of each datset on every subset was the same. The MR training set was used to update the weights of the CNN through stochastic gradient descent (Adam optimizer with default parameters), while the MR validation set was used to choose the best set of hyper-parameters (such as learning rate schedule, CNN depth, CNN width, input resolution or even internal CNN architecture). Once the MR segmentation model was considered final, the CNN was retrained one last time using both MR training and validation subsets, and the results were evaluated in the MR test subset.

For the US segmentation model no hyper-parameters were changed, except the input size and spacing (to better fit the prostate in the image, as discussed in Section 2.2) and the batch size (due to GPU memory limitations, as discussed in Section 2.3). Furthermore, transfer learning was employed [28]: the weights from the MR segmentation model were used as an initialization to the US segmentation model, thus leveraging the feature extraction capabilities of the pre-trained model. The US model was directly trained using both US training and validation subsets (as no validation subset was actually required), and the results were evaluated in the US test subset.

2.7. Cyclic Learning Rate

A cyclic learning rate [29] was chosen for training, as it presents several advantages with respect to a fixed schedule. Firstly, an optimal fixed schedule must be learned from the data, which is cumbersome and requires extensive trial and error; secondly, this cyclic schedule will allow us to use a technique known as Checkpoint Ensembling, which will be explained in Section 2.8. Thirdly, a cyclic learning rate supposedly helps the optimizer escape saddle point plateaus, which is desirable. The chosen cyclic schedule is a decaying triangular wave (schedule known as "triangular2" in [29]) of period 48 epochs (at 180 batches per epoch), with a minimum learning rate of $5.5 \cdot 10^{-5}$, a maximum of $7.5 \cdot 10^{-4}$, and a decay such that the maximum value of the wave is halved every period. The CNN was trained for six periods (a total of 288 epochs).

2.8. Checkpoint Ensembling

Checkpoint Ensembling [30] is a strategy that allows to capture the effects of traditional ensembling methods within a single training process. It works by collecting checkpoints of the best k weights (those that lead to the best validation scores of the CNN during its training process). Then, during inference, for each input to the CNN, k predictions are obtained and combined into a single one (by averaging, for instance). In theory, this method makes a compound prediction from weights which may have settled into different local minima, thus simulating the compound segmentation proposal from several experts.

As for our model, using a cyclic learning rate opens up the possibility of using weight checkpoints that coincide with the minima of the learning rate schedule, as it is at these points where the gradient stabilizes most, and local minima are supposedly reached. Therefore, in our particular case, six checkpoints will be used for Checkpoint Ensembling. As a bonus, this technique incurs in no additional costs, other than inference costs, which are obviously increased by a factor of six. Traditional ensembling was also tested, although finally discarded, as it did not provide any performance improvements and incurred in much higher training costs.

2.9. Neural Resolution Enhancement

The last technique that will be discussed is Neural Resolution Enhancement [31], which leverages the properties of any already trained image segmentation CNN to intelligently increase the resolution of the output mask at no cost, even beyond the resolution of the original image.

To understand how this method operates, let us describe how a threefold increase in the resolution along the z-axis would be performed (refer to Figure 4). First, the resolution of the input image X is triplicated along the z-axis (by using bicubic interpolation, for instance), therefore becoming $\hat{X}$. Then, three new images (X_0, X_1, X_2) are built by taking z-slices from $\hat{X}$ in such a way that they have the same size (dimensions) and spacing (voxel size) as X, but are offset by different sub-voxel amounts along z-axis (in fact, note that $X_0 \equiv X$). Then, X_0, X_1 and X_2 are fed through the CNN, and three segmentation masks are obtained (Y_0, Y_1, Y_2). Finally, all three predictions are combined by stacking them in the correct order, hence obtaining $\hat{Y}$, which is a predicted mask with three times the resolution of X along the z-axis. This same procedure could be applied to any number of dimensions simultaneously, although the inference cost would scale abruptly, as all the possible sub-voxel displacement combinations would have to be computed.
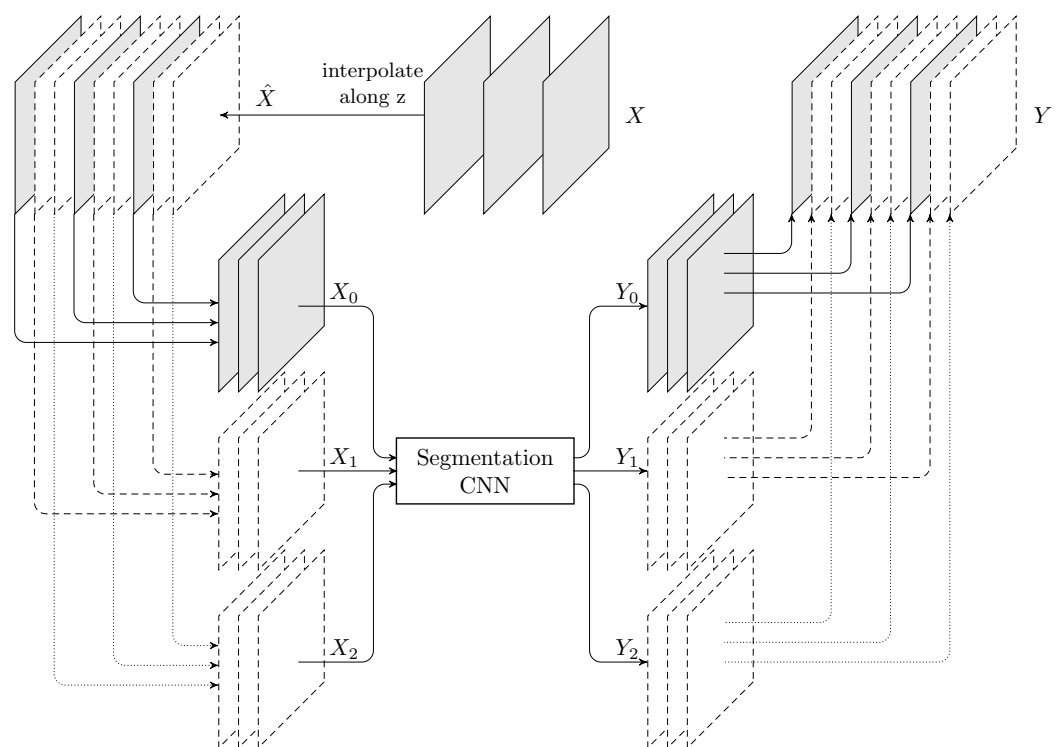
Figure 4. Visual representation of the Neural Resolution Enhancement method to triplicate the resolution along the z-axis.

This method, albeit simple, is extremely powerful, as it allows to predict (rather than to interpolate) segmentation masks beyond the resolution of the original image. The problem of interpolation is therefore shifted from the mask domain to the image domain, where the conveyed information is still complete and not yet binarized. Furthermore, it can be applied to any already trained segmentation CNN, as a simple post-processing step. Figure 5 shows an example application.

In the context of our problem, z-axis resolution is triplicated to reduce the impact of the final mask interpolation. This is: once the CNN outputs a segmentation mask, it must be transformed back to the space of the original input image (same resolution, spacing, physical orientation, position, etc.). Although this is necessarily a lossy process, by leveraging this technique, the predicted mask can have a higher resolution, which significantly mitigates the issue.
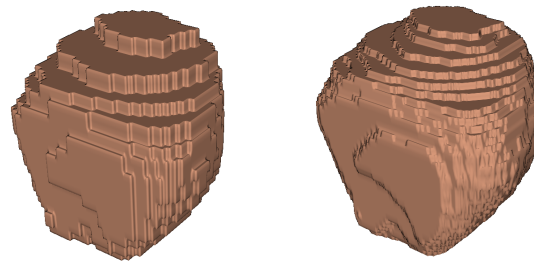
Figure 5. Mask predicted by a prostate segmentation CNN and upscaled along the z-axis three times using: nearest-neighbor interpolation (**left**), and Neural Resolution Enhancement (**right**).

2.10. Evaluation Metrics and Loss

As it is customary in semantic segmentation problems, *DSC* was employed as the main evaluation metric, which guided most design decisions. *DSC* is defined in Equation (2), where N denotes the total number of voxels in an image, $\hat{y}_i \in [0, 1]$ represents the prediction of the CNN at voxel i, $y_i \in \{0, 1\}$ is the GT label at voxel i, and $\epsilon = 1$ is a small arbitrary value that prevents division by zero.

$$DSC(y, \hat{y}) = \frac{2 \cdot \sum_i^N \hat{y}_i \cdot y_i + \epsilon}{\sum_i^N \hat{y}_i + \sum_i^N y_i + \epsilon} \qquad (2)$$

As a loss function, *DSC* is much better able to deal with unbalanced segmentation masks in comparison with binary cross-entropy. However, several studies acknowledge its deficiencies along the boundaries of the mask [23], or when the target is very small [32] (as in lesion segmentation). Ref. [23], for instance, utilizes a composite loss which penalizes wrong segmentations proportionally to the distance to the boundary of the GT. Despite multiple attempts at incorporating a similar loss to our model, we finally decided against it, since it did not provide any performance advantages during validation. Therefore, the finally used loss function $\mathcal{L}$ is directly derived from *DSC*, as illustrated in Equation (3).

$$\mathcal{L} = 1 - DSC \qquad (3)$$

In addition to *DSC*, two distance-based metrics were also employed: Average Boundary Distance (ABD) and 95th percentile Hausdorff Distance (HD95). These metrics were computed as described in the Promise12 challenge [21] and represent the average and the 95th percentile largest distance (in mm) between the surface of the predicted mask and the GT mask, respectively.

When comparing these metrics among groups, the Wilcoxon signed-rank test was employed, which is the non-parametric equivalent of the paired t-test. The Wilcoxon test was needed due to the distribution of the metrics in the test set not being normal (*p*-value ≤ 0.001 using D'Agostino and Pearson's normality test for *DSC*, ABD and HD95 results).

3. Results

3.1. Quantitative Results

The quantitative test results (in terms of *DSC*, HD95 and ABD metrics) for both MR and US segmentation models (globally, and by dataset) are shown in Table 2. Both models achieve a mean and median *DSC* above the 0.91 threshold for all datasets, meaning that they are very strong performers and, more interestingly, that they are robust to the heterogeneity of the various datasets. As an exception, the mean *DSC* on the "Girona" dataset falls to around 0.90 due to a single relatively weak prediction (*DSC* of 0.8467) dragging down the mean of this extremely small set, as evidenced by the otherwise inexplicably high median value. Also, the mean *DSC* for the MR segmentation model on the "Private" dataset is exceptionally high, probably due to it being the only dataset where GT segmentations were revised.

These observations are further supported by the HD95 metric. For all MR and US datasets, it sits mostly just below 4 mm in average. Since the slices in a typical prostate MR image are about 3 mm apart, achieving an HD95 below 3 mm is extremely unlikely due to the different criteria regarding how far the base and the apex should extend. Thus, an average HD95 below 4 mm is a very good result. Finally, the ABD metric lies mainly below 1 mm in average for all MR datasets, and below 1.2 mm for the US datasets.

Table 2. Quantitative results for all datasets and models.

Dataset		N	*DSC*			HD95 (mm)			ABD (mm)		
			Mean	**Median**	**Min.**	**Mean**	**Median**	**Max.**	**Mean**	**Median**	**Max.**
MR	Girona	5	0.8980	0.9113	0.8467	3.7240	3.7873	4.2054	1.3305	1.3187	2.0323
	Promise12	7	0.9148	0.9118	0.8919	4.2876	3.6000	7.2000	1.0135	0.9680	1.2757
	Prostate-3T	2	0.9222	0.9222	0.9099	3.6000	3.6000	3.6000	0.9190	0.9190	1.0719
	IVO	42	0.9136	0.9182	0.8094	3.9947	4.0002	6.9347	0.9569	0.9013	2.0356
	Private	13	0.9251	0.9228	0.8993	3.3363	3.1736	4.5122	0.9190	0.8525	1.2815
	All	69	0.9150	0.9179	0.8094	3.8693	3.9995	7.2000	0.9815	0.9311	2.0356
US	IVO	24	0.9215	0.9256	0.8456	3.4295	3.1210	9.9997	1.1825	1.0539	2.8573
	Private	12	0.9131	0.9133	0.8960	3.6317	3.7025	6.1216	1.1872	1.2008	1.7809
	All	36	0.9187	0.9235	0.8456	3.4969	3.2863	9.9997	1.1840	1.1102	2.8573

For comparison purposes, a second segmentation (GT2) was created for the first three datasets by one of the IVO experts. Table 3 shows the mean *DSC* of the predictions of the model against each of the GTs (GT and GT2), as well as the mean *DSC* of the GTs against themselves (the inter-expert agreement). As it can be seen, the *DSC* of the model against both GT and GT2 surpasses by a large margin the inter-expert agreement (except for one case), suggesting that the model is more robust and reliable than any given expert by itself. Two Wilcoxon tests confirm that these differences in *DSC* are statistically significant (at a significance threshold of 0.05).

Table 3. Evaluation of the predictions against GT and GT2, as well as GT against GT2 (inter-expert performance).

Dataset (MR)	N	Mean *DSC*		
		Predicted & GT	**Predicted & GT2**	**GT & GT2**
Girona	5	0.8980	0.9057	0.8657
Promise12	7	0.9148	0.9032	0.8825
Prostate-3T	2	0.9222	0.8995	0.9026
All above	14	0.9099	0.9035	0.8794
		p-value against last column		
		0.0035	0.0258	-

Since most authors focus on performing well in one single dataset, it is difficult to compare these results against other published models. As an exception [16] used a private dataset (with diffusion-weighted MR images and ADC-maps in addition to the T2-weighted MR images) in conjunction with "Promise12", achieving an impressive mean *DSC* of 0.9511 on their own dataset, but only a mean *DSC* of 0.8901 on "Promise12_test".

Regarding 3D-US image segmentation, few publications were found (most use 2D-US), and none employed more than one dataset. As for recent 3D-US papers, ref. [19] achieved a mean *DSC* of 0.90 by leveraging an attention mechanism. Ref. [33] obtained a *DSC* 0.919 by using a contour-refinement post-processing step, however, the results are not reported on a proper test set, but rather, using leave-one-out cross-validation. More recently, ref. [34]

achieved an excellent 0.941 mean *DSC* by applying a 2D U-Net on radially sampled slices of the 3D-US and then reconstructing the full 3D volume. As an example on the problem of 2D-US segmentation, ref. [35] achieved a mean *DSC* of 93.9 by using an ensemble of five CNNs. This last result, however, is not directly comparable, as in 2D-US segmentation the *DSC* is evaluated on a per-slice basis, instead of the prostate as a whole.

Promise12 is an ongoing prostate segmentation challenge, wherein 50 MR prostate images are provided along with their segmentation masks (dataset "Promise12"), and 30 additional images are provided without segmentations as a test set (dataset "Promise12_test"). Table 4 shows the performance of the model on "Promise12_test" along with the five best entries to the Promise12 challenge. For this specific dataset, the predicted segmentation masks are uploaded and evaluated in the servers of the challenge, and the results are publicly posted online thereafter [36].

As it should be expected, the mean and median for our model are similar to the results obtained for the other test sets (Table 2). Also, comparing it to the other entries (Table 4), our model achieves very similar results for all metrics; yet, its Challenge Score falls behind. That said, this is also to be expected since, unlike the other contestants, no fine-tuning was performed to improve the results for this dataset in particular. BOWDA-net [23], for example, uses an adversarial domain adaptation strategy to transform the images from a second training dataset to the domain of the "Promise12" dataset, therefore improving the performance only on "Promise12_test". Lastly, our model used just 41 out of the 50 images provided in the "Promise12" dataset for training, as two were discarded and seven were used for testing. When comparing our model against each of the others with a Wilcoxon test, only the first contender (MSD-Net) was found to be significantly (p-value ≤ 0.01) better in all metrics, while the fourth contender (nnU-Net) was better in terms of *DSC* (p-value = 0.037) and ABD (p-value = 0.030), but not HD95 (p-value = 0.439). The nnU-Net [37] is a very recent and interesting method that tries to automate the process of adapting a CNN architecture to a new dataset by making use of a sensible set of heuristics. Regarding the MSD-Net, unfortunately, its specifics are yet to be published as of the writing of this paper.

Table 4. MR model performance on "Promise12_test" along with the five best entries as of December 2020.

Challenge Score	Name	DSC			HD95 (mm)			ABD (mm)		
		Mean	Median	Min.	Mean	Median	Max.	Mean	Median	Max.
91.9072	MSD-Net	0.9299	0.9323	0.8890	3.5512	3.3454	7.3344	1.1160	1.0968	1.6777
90.7993	Edge Att.	0.9118	0.9136	0.8672	4.3095	3.9362	7.6217	1.4264	1.4004	2.3584
90.3441	HD_Net	0.9135	0.9129	0.8398	3.9331	3.7134	5.9674	1.3614	1.3090	2.2662
89.6507	nnU-Net	0.9194	0.9272	0.8406	3.9509	3.7276	6.8301	1.2431	1.1771	2.1693
89.5858	Bowda-Net	0.9141	0.9222	0.8367	4.2654	3.8969	7.7235	1.3451	1.2763	2.2920
88.5397	Ours	0.9137	0.9168	0.8741	4.1176	3.8449	7.8605	1.3197	1.2864	1.8129

Ultimately, beating this challenge was never the focus of this paper. No other single model (to the author's knowledge) is able to perform as consistently as ours in so many different datasets simultaneously. This is of utmost importance if such a model is to be used in a real-life scenario, where the MR images may come from many different scanners, and may be analyzed by many different experts.

3.2. Qualitative Results

To asses these results qualitatively, in Figure 6a–n, the center 100 mm $\times$ 100 mm crop (85 mm $\times$ 85 mm in the case of US images) of three slices from the worst and best performing images (in terms of *DSC*) from each dataset have been represented, along with the GT (in red), the GT2 (in blue, when available) and the predicted segmentations (in green). Figures were generated using Python library plot_lib [38].
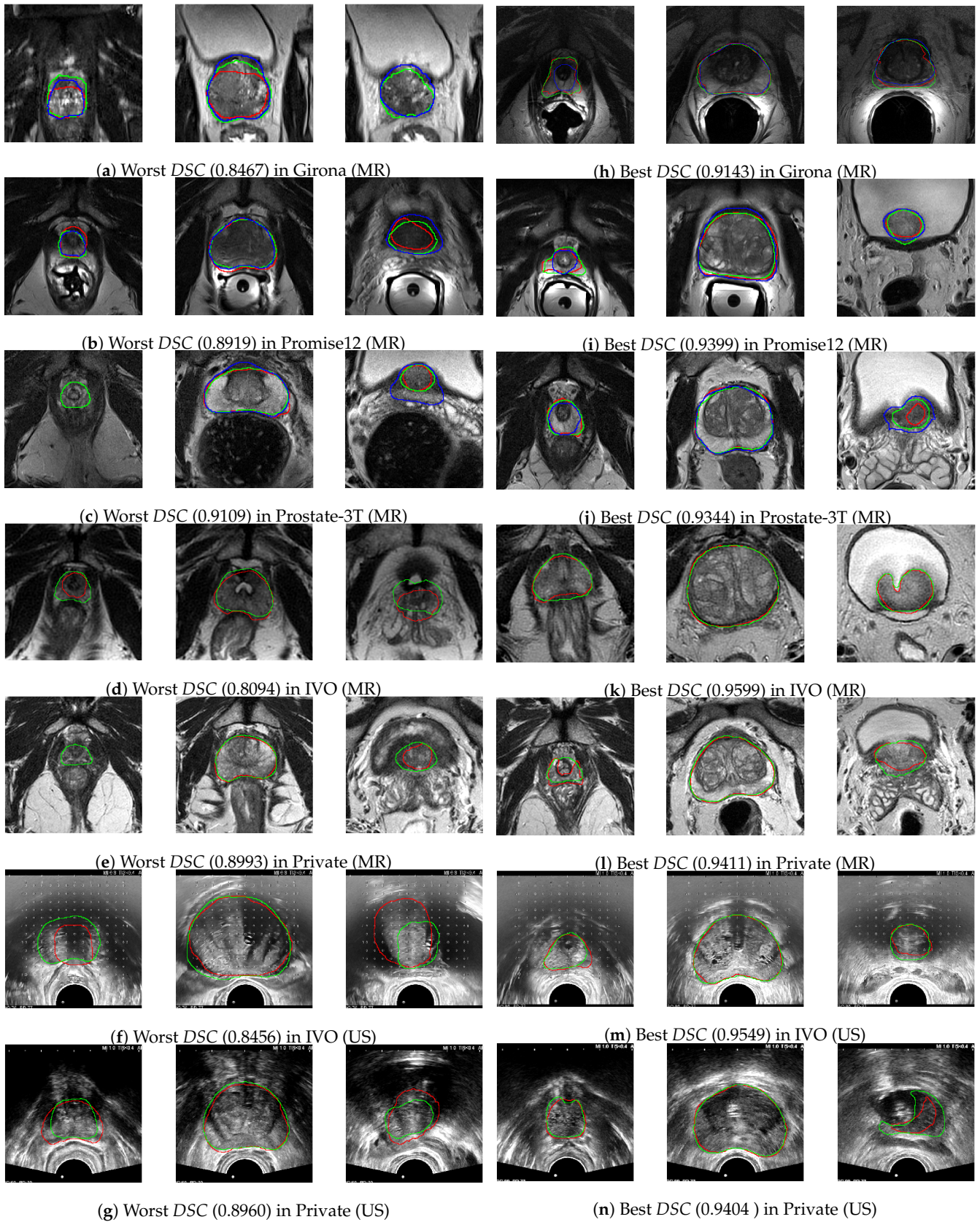
Figure 6. Worst (**left**) and best (**right**) segmentations in terms of *DSC* for each dataset (green: model, red: GT, blue: GT2).

Regarding the worst cases, despite being the poorest performers, the differences are relatively small and often the model proposal is arguably superior to the GT. Furthermore, the central slices are almost identical in all instances, and it is only towards base and apex

where the differences emerge. One of such discrepancies is the point at which the apex and the base begin, which oftentimes depends on the segmentation criteria, as it can be seen, for instace, in Figure 6a, where the CNN indicates the presence of prostate in the rightmost slice (at the base), while the GT label does not (although GT2 does). Finally, at several ambiguous instances (such as in the middle slice of Figure 6a and the rightmost slice of Figures 6b,j), the predicted mask (in green) behaves as an average between both experts. As discussed, this is a very desirable property for the model to have, and this is what allows it to outperform single experts on their own (as demonstrated in Table 3).

As for the best cases, it can be seen that they are mostly represented by larger prostates, as they are comparatively easier to segment, and also the *DSC* metric is biased towards them. As a curiosity, the rightmost slice in Figure 6n shows how the model has learned to avoid segmenting the catheter balloon that is used in prostate biopsies, the procedure during which the US images were acquired.

In terms of HD95, the worst MR case, which corresponds to an HD95 of 7.2 mm, is shown in Figure 7. As it can be seen, two slices from the apex are missed by the algorithm, hence amounting to a minimum of 2×3 mm of error, plus some extra mm. The worst performing MR case in terms of ABD coincides with the worst performing prostate in terms of *DSC*, which can be found in Figure 6d.
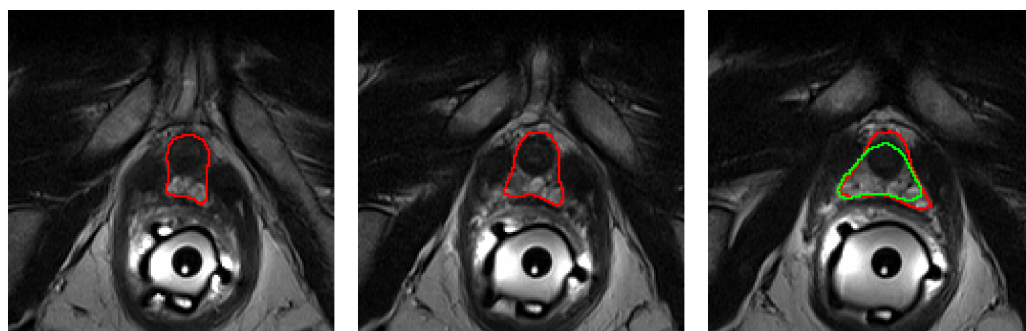
Figure 7. Worst HD95 (7.2 mm) of all MR test datasets. Two slices from the apex (**left** and **center**) are missed by the algorithm, hence amounting to a minimum of 2×3 mm of error, plus some extra mm from the segmentation errors commited in the third slice (**right**).

3.3. Ablation Studies

Table 5 contains the results of the ablation studies, which were performed by changing one single aspect of the baseline MR model at a time. Wilcoxon tests were performed against the baseline to check for significance (p-value < 0.05). Also, a single experiment was performed on the US model by retraining it without the use of transfer learning.

Firstly, the two post-processing techniques discussed in this paper (Checkpoint Ensembling and Neural Resolution Enhancement) are analyzed. Both show high statistical significance (p-values < 0.01) in terms of *DSC* and ABD. In fact, out of all the experiments conducted in this Section, only these two were found to make a statistically significant difference, probably since the worsening of the metrics, even if minor, is sustained for all images. These post-processing methods affect in no way the training process or the model, as they are applied at a later stage; therefore they are a simple and free bonus in performance, only at the cost of increased inference time.

Secondly, a battery of tests involving architectural changes (which require retraining) is presented. Even if none of these experiments showed statistical significance, several conclusions can still be extracted cautiously.

The first two experiments are an attempt to lower the complexity of the baseline model by either reducing the number of resolution levels of the network, or the amount of layers (this is: "standard" convolutions) per level. In both cases, even if the differences with respect to the baseline were small, a decrease in performance can be observed for the majority of the metrics, which justifies the use of the more complex baseline architecture if possible.

In the next two experiments, models based exclusively on the resnet architecture (with residual connections applied every four consecutive convolutions) were employed. Despite having as much as four times the amount of parameters as compared to the baseline, these models were the worst performing out of all analyzed in this Section, hence showing the power of the densenet architecture.

The next test consisted in replacing the PReLU activations with ReLU activations. Despite this having a very small influence in performance, the metrics are overall better in the baseline, and PReLU is therefore preferred given its negligible impact on model complexity.

For the following test, Deep Supervision was deactivated. In general, most of the metrics show a small improvement with this architectural modification. In our internal validation tests, this technique seemed to provide a small boost in performance and, as such, it was added to the final model. Furthermore, it stabilized the initial steps of the training procedure. However, in light of these results, its usefulness remains now in question.

Table 5. Ablation studies.

	Experiment Description	DSC			HD95 (mm)			ABD (mm)		
		Mean	Median	Min.	Mean	Median	Max.	Mean	Median	Max.
MR	Baseline (4M params.)	0.9150	0.9179	0.8094	3.8693	3.9995	7.2000	0.9815	0.9311	2.0356
	No neural Resolution Enhancement	0.9127	0.9153	0.8064	3.8991	3.9994	7.2000	1.0344	1.0090	2.1380
	No Checkpoint Ensembling	0.9128	0.9178	0.8009	4.0071	3.9997	8.0018	1.0269	0.9674	2.2703
	One less resolution reduction level	0.9145	0.9163	0.8404	3.9413	3.8665	11.1732	0.9952	0.9468	2.6608
	Half the amount of layers per level	0.9144	0.9161	0.7556	3.9394	3.8177	9.2372	0.9912	0.9316	2.3902
	Resnet with 20 layers (7M params.)	0.9063	0.9112	0.4647	4.2393	3.9995	15.1950	1.1174	0.9842	6.4507
	Resnet with 72 layers (16M params.)	0.9033	0.9150	0.1525	4.1926	3.9997	17.4923	1.1348	0.9207	9.9787
	ReLU instead of PReLU activations	0.9146	0.9171	0.8135	3.9631	3.7499	13.8931	1.0052	0.9561	3.1201
	No Deeep Supervision	0.9152	0.9143	0.8257	3.8033	3.7499	7.0495	0.9732	0.9802	1.9869
US	Baseline	0.9187	0.9235	0.8456	3.4969	3.2863	9.9997	1.1840	1.1102	2.8573
	No transfer learning (fine-tuning)	0.9207	0.9236	0.8576	3.4166	3.2372	8.3040	1.1536	1.1096	2.2470

For the last test, the US segmentation model was retrained using random weight initialization, instead of using the weights from the MR model for transfer learning. Although the fine-tuned model converged faster, the results suggest that training from scratch might be preferable in situations like this one, where the amount of training data is sufficient. In any event, the generality of the architecture and pre-/post-processing methodology still holds, even if the weights of the MR prostate segmentation model are not particularly useful for the problem of US prostate segmentation.

4. Discussion

In this paper, a robust and generalizable model for accurate prostate segmentation has been proposed.

To achieve robustness, the model was trained with five very challenging and heterogeneous MR prostate datasets (and two US prostate datasets) with GTs originating from many different experts with varying segmentation criteria. Additionally, several key design choices, such as the use of Checkpoint Ensembling and a relatively heavy data augmentation regime, were explicitly made.

In clinical practice, the MR and US images may originate from many different scanners, with widely different characteristics (field intensities, scanner manufacturers, use of endorectal coil, etc.), to which any segmentation algorithm should be robust by design. As we have seen in Table 2, the proposed model has a similarly good performance for all images, no matter the dataset nor its specific characteristics.

Furthermore, such an algorithm may be used by different experts with varying criteria for segmenting the prostate. Even if it is impossible to please every criterion simultaneously, the proposed model is shown to behave as an average prediction among the different experts, as seen for instance in the rightmost slice of Figure 6j. This is corroborated by Table 3, where it shows a significantly higher overlap with any given expert (it tries to please all criteria), than the experts between themselves.

Concerning generalizability, the proposed architecture can be very easily applied to different tasks by means of transfer learning. In this paper, the MR segmentation model is simply retrained, with no hyperparameter tuning or image pre- or post-processing changes (other than the change of input resolution), on the problem of US prostate segmentation, achieving equally good performances despite the smaller dataset sizes, as seen in Table 2.

The main clinical applications of the proposed model lie in the context of fusion-guided biopsy and focal therapies on the prostate, which require an accurate segmentation of both MR and US prostate images. These segmentations are employed to perform registration between both modalities, which is needed to transfer the prostate lesions detected by the radiologists in the preoperative MR to the intraoperative US image in order to guide the procedure.

The proposed model can undercut, and even eliminate, the need of manual segmentations, which require expertise, are very time-consuming, and are prone to high inter- and intra-expert variability. Hence, more accurate segmentations may lead to better inter-modal prostate image registration and better prognosis in the aforementioned procedures; while almost instant results can be of particular interest for the segmentation of intraoperative US images, where the urologist currently has to spend around ten minutes manually performing this task next to the sedated patient.

Finally, a technique known as Neural Resolution Enhancement was employed as a post-processing step to reduce the impact of the lossy CNN output interpolation. This method, which leverages any already trained segmentation CNN, can also be used to improve the resolution of the output mask even beyond that of the original input image, as discussed in Figure 5.

This technique could be especially appealing for simulating the biomechanical behavior of the prostate, which is required by many registration algorithms and surgical simulators. To function properly, such simulations demand very high resolution meshes of the prostate geometry, which are inherently impossible to obtain due to the reduced resolution of the original MR and US images. However by using Neural Resolution Enhancement, a much higher resolution mask is obtained, which is not the result of mere interpolation, but rather a prediction of the missing geometry by combining the contextual information contained in the original image with the knowledge that the CNN has acquired about the general shape of this gland.

5. Conclusions

In conclusion, this paper proposes a prostate segmentation model with the potential of having a significant impact on the efficacy and efficiency of current guided prostate procedures, through improvements in terms of performance, robustness, generalizability and output resolution.

In our future work, the proposed model will be applied to different organs and tasks, such as lesion segmentation. Furthermore, different backbone architectures, such as those based on Neural Architecture Search, will be tested.

Author Contributions: Conceptualization, J.D.M.-G.; Data curation, O.J.P.-V. and V.G.-P.; Methodology, O.J.P.-V.; Project administration, J.D.M.-G.; Resources, V.G.-P., I.M.G. and M.B.B.; Software, O.J.P.-V.; Supervision, J.L.C.R.-B., J.R.-B., M.J.R. and J.D.M.-G.; Validation, I.M.G., M.B.B. and P.P.G.; Visualization, O.J.P.-V.; Writing—original draft, O.J.P.-V.; Writing—review & editing, V.G.-P., M.J.R. and J.D.M.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by a doctoral grant of the Spanish Ministry of Innovation and Science, with reference FPU17/01993.

Institutional Review Board Statement: The study was approved by the Ethical Committee of the Valencia Institute of Oncology (CEIm-FIVO) with protocol code PROSTATEDL (2019-12) and date 17 July 2019.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Dataset Girona is available at Zenodo (https://doi.org/10.5281/zenodo.162231), Promise12 at Grand Challenge (https://promise12.grand-challenge.org/Download) and Prostate-3T at the Cancer Imaging Archive (https://wiki.cancerimagingarchive.net/display/Public/Prostate-3T). Dataset IVO, from the Valencian Institute of Oncology, is not publicly available, since the ethical committee (CEIm-FIVO) only approved its use for the current study. Dataset Private comes from a private institution which retains all rights of usage for the images, and hence it is not publicly available either.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Marra, G.; Ploussard, G.; Futterer, J.; Valerio, M.; Ploussard, G.; De Visschere, P.J.; Tsaur, I.; Tilki, D.; Ost, P.; Gandaglia, G.; et al. Controversies in MR targeted biopsy: Alone or combined, cognitive versus software-based fusion, transrectal versus transperineal approach? *World J. Urol.* **2019**. [CrossRef] [PubMed]
2. Ahdoot, M.; Lebastchi, A.H.; Turkbey, B.; Wood, B.; Pinto, P.A. Contemporary treatments in prostate cancer focal therapy. *Curr. Opin. Oncol.* **2019**. [CrossRef] [PubMed]
3. Shahedi, M.; Halicek, M.; Li, Q.; Liu, L.; Zhang, Z.; Verma, S.; Schuster, D.M.; Fei, B. A semiautomatic approach for prostate segmentation in MR images using local texture classification and statistical shape modeling. In *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling. International Society for Optics and Photonics*; Fei, B., Linte, C.A., Eds.; SPIE: Bellingham, WA, USA, 2019; Volume 10951, p. 91. [CrossRef]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2012**, *60*, 84–90. [CrossRef]
5. Allen, P.D.; Graham, J.; Williamson, D.C.; Hutchinson, C.E. Differential segmentation of the prostate in MR images using combined 3D shape modelling and voxel classification. In Proceedings of the 2006 33rd IEEE International Symposium on Biomedical Imaging, Arlington, VA, USA, 6–9 April 2006; Volume 2006, pp. 410–413. [CrossRef]
6. Freedman, D.; Radke, R.J.; Zhang, T.; Jeong, Y.; Lovelock, D.M.; Chen, G.T. Model-based segmentation of medical imagery by matching distributions. *IEEE Trans. Med. Imaging* **2005**, *24*, 281–292. [CrossRef] [PubMed]
7. Klein, S.; van der Heide, U.A.; Lips, I.M.; van Vulpen, M.; Staring, M.; Pluim, J.P.W. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Med. Phys.* **2008**, *35*, 1407–1417. [CrossRef]
8. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946.
9. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lect. Notes Comput. Sci.* **2015**, *9351*, 234–241. [CrossRef]
10. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
11. Cireşan, D.C.; Giusti, A.; Gambardella, L.M.; Schmidhuber, J. Deep neural networks segment neuronal membranes in electron microscopy images. *Adv. Neural Inf. Process. Syst.* **2012**, *4*, 2843–2851.
12. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *39*, 640–651. [CrossRef]
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 770–778. [CrossRef]
14. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016. [CrossRef]
15. Zhu, Q.; Du, B.; Turkbey, B.; Choyke, P.L.; Yan, P. Deeply-supervised CNN for prostate segmentation. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017. [CrossRef]

16. To, M.N.N.; Vu, D.Q.; Turkbey, B.; Choyke, P.L.; Kwak, J.T. Deep dense multi-path neural network for prostate segmentation in magnetic resonance imaging. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 1687–1696. [CrossRef]
17. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
18. Zhu, Y.; Wei, R.; Gao, G.; Ding, L.; Zhang, X.; Wang, X.; Zhang, J. Fully automatic segmentation on prostate MR images based on cascaded fully convolution network. *J. Magn. Reson. Imaging* **2019**, *49*, 1149–1156. [CrossRef] [PubMed]
19. Wang, Y.; Dou, H.; Hu, X.; Zhu, L.; Zhu, L.; Yang, X.; Xu, M.; Qin, J.; Heng, P.A.; Wang, T.; et al. Deep Attentive Features for Prostate Segmentation in 3D Transrectal Ultrasound. *IEEE Trans. Med. Imaging* **2019**, *38*, 2768–2778. [CrossRef] [PubMed]
20. Lemaître, G.; Martí, R.; Freixenet, J.; Vilanova, J.C.; Walker, P.M.; Meriaudeau, F. Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review. *Comput. Biol. Med.* **2015**, *60*, 8–31. [CrossRef] [PubMed]
21. Litjens, G.; Toth, R.; van de Ven, W.; Hoeks, C.; Kerkstra, S.; van Ginneken, B.; Vincent, G.; Guillard, G.; Birbeck, N.; Zhang, J.; et al. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Med. Image Anal.* **2014**, *18*, 359–373. [CrossRef] [PubMed]
22. Litjens, G.; Futterer, J.; Huisman, H. Data From Prostate-3T. 2015. Available online: https://cloud.google.com/healthcare/docs/resources/public-datasets/tcia-attribution/prostate-3t (accessed on 28 December 2020). [CrossRef]
23. Zhu, Q.; Du, B.; Yan, P. Boundary-weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 753–763. [CrossRef]
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034. [CrossRef]
25. Nair, V.; Hinton, G.E. Rectified linear units improve Restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; pp. 807–814.
26. Chen, T.; Xu, B.; Zhang, C.; Guestrin, C. Training Deep Nets with Sublinear Memory Cost. *arXiv* **2016**, arXiv:1604.06174.
27. Lee, C.Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-Supervised Nets. *J. Mach. Learn. Res.* **2014**, *38*, 562–570.
28. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**. [CrossRef]
29. Smith, L.N. Cyclical Learning Rates for Training Neural Networks. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 464–472. [CrossRef]
30. Chen, H.; Lundberg, S.; Lee, S.I. Checkpoint Ensembles: Ensemble Methods from a Single Training Process. *arXiv* **2017**, arXiv:1710.03282.
31. Pellicer-Valero, O.J.; Martín-Guerrero, J.D.; Rupérez, M. Cost-free resolution enhancement in Convolutional Neural Networks for medical image segmentation. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2–4 October 2020; pp. 145–150.
32. Abraham, N.; Khan, N.M. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 683–687. [CrossRef]
33. Lei, Y.; Tian, S.; He, X.; Wang, T.; Wang, B.; Patel, P.; Jani, A.B.; Mao, H.; Curran, W.J.; Liu, T.; et al. Ultrasound prostate segmentation based on multidirectional deeply supervised V-Net. *Med. Phys.* **2019**, *46*, 3194–3206. [CrossRef]
34. Orlando, N.; Gillies, D.J.; Gyacskov, I.; Romagnoli, C.; D'Souza, D.; Fenster, A. Automatic prostate segmentation using deep learning on clinically diverse 3D transrectal ultrasound images. *Med. Phys.* **2020**, *47*, 2413–2426. [CrossRef] [PubMed]
35. Karimi, D.; Zeng, Q.; Mathur, P.; Avinash, A.; Mahdavi, S.; Spadinger, I.; Abolmaesumi, P.; Salcudean, S.E. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. *Med. Image Anal.* **2019**, *57*, 186–196. [CrossRef] [PubMed]
36. Litjens, G.; Toth, R.; van de Ven, W.; Hoeks, C.; Kerkstra, S.; van Ginneken, B.; Vincent, G.; Guillard, G.; Birbeck, N.; Zhang, J.; et al. PROMISE12 Results. 2020. Available online: https://promise12.grand-challenge.org/ (accessed on 28 December 2020).
37. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2020**. [CrossRef] [PubMed]
38. Pellicer-Valero, O.J. OscarPellicer/plot_lib. 2020. Available online: https://zenodo.org/record/4395272 (accessed on 28 December 2020). [CrossRef]