TETRIS: Template Transformer Networks for Image Segmentation With Shape Priors

Matthew Chung Hai Lee, Kersten Petersen, Nick Pawlowski, Ben Glocker, and Michiel Schaap

***Abstract*— In this paper, we introduce and compare different approaches for incorporating shape prior information into neural network-based image segmentation. Specifically, we introduce the concept of template transformer networks, where a shape template is deformed to match the underlying structure of interest through an end-to-end trained spatial transformer network. This has the advantage of explicitly enforcing shape priors, and this is free of discretization artifacts by providing a soft partial volume segmentation. We also introduce a simple yet effective way of incorporating priors in the state-of-the-art pixel-wise binary classification methods such as fully convolutional networks and U-net. Here, the template shape is given as an additional input channel, incorporating this information significantly reduces false positives. We report results on synthetic data and sub-voxel segmentation of coronary lumen structures in cardiac computed tomography showing the benefit of incorporating priors in neural network-based image segmentation.**

***Index Terms*— Image segmentation, shape priors, neural networks, template deformation, image registration.**

I. Introduction

SEGMENTATION of anatomical structures can be greatly improved by incorporating priors on shape, assuming population wide regularities are observed, or that expert knowledge is available. Shape priors help to reduce the search space of potential solutions for machine learning algorithms, improving the accuracy and plausibility of solutions [1]. Priors are particularly useful when data is ambiguous, corrupt, exhibits low signal-to-noise or if training data is scarce.

Some of the first attempts to explicitly enforce shape priors in segmentation pipelines made use of deformable templates [2], combining image registration with a shape template to perform segmentation. Subsequently this method was combined with anatomical atlases to perform segmentations of different organs [3]–[5]. However these methods require either an image-to-image or image-to-segmentation likelihood function to drive atlas matching or alignment of the deformation model. Statistical methods such as active shape models [6] have been explored extensively with the difficulty of constructing shape models in the first place which are then often limited in their expressiveness due to the underlying manifold learning method (linear or non-linear principal component analysis).

State-of-the-art neural network based segmentation models [7]–[10] typically optimize pixel-wise loss functions such as mean squared error or cross entropy, and more recently differentiable Dice [11]. These objective functions do not take explicit priors into consideration during training. Nevertheless, smoothness priors can be enforced during test time by using conditional random fields or similar post-processing techniques. More recent work has shown improved results by directly incorporating shape constraints into their learning algorithm rather than applying them as post-processing such as in [12]–[14], where priors are learnt to regularize neural network embeddings during training. While this can lead to networks that favor plausible segmentations, there is no guarantee that the outputs adhere to desired shape constraints, such as a single connected component or a closed surface.

A. Contributions

In this paper we introduce a new neural network model based on template deformations which utilizes spatial transformer networks [15]. Our model leverages the representational power of neural networks while *explicitly* enforcing shape constraints in segmentations by restricting the model to perform segmentation through deformations of a given shape prior. We call this Template Transformer Networks for Image Segmentation (TETRIS). As with template deformations, our method produces anatomically plausible results by regularizing the deformation field. This also avoids discretization artifacts as we do not restrict the network to make pixel-wise classifications. By using a neural network that is trained to align the shape prior to the structure of interest visible in the input image there is no need for a hand crafted (intensity-based) image-to-segmentation registration measure as with other template deformation models. To the best of our knowledge, this is the first full 3D neural network-based image segmentation through

Manuscript received January 16, 2019; revised March 12, 2019; accepted March 12, 2019. Date of publication March 22, 2019; date of current version October 25, 2019. *(Ben Glocker and Michiel Schaap have shared senior authorship.) (Corresponding author: Matthew Chung Hai Lee.)*

M. C. H. Lee, B. Glocker, and M. Schaap are with HeartFlow, Redwood, CA 94063 USA, and also with the Biomedical Image Analysis Group, Imperial College London, London SW7 2AZ, U.K. (e-mail: matthew.lee13@imperial.ac.uk).

K. Petersen is with HeartFlow, Redwood, CA 94063 USA.

N. Pawlowski is with the Biomedical Image Analysis Group, Imperial College London, London SW7 2AZ, U.K.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2905990

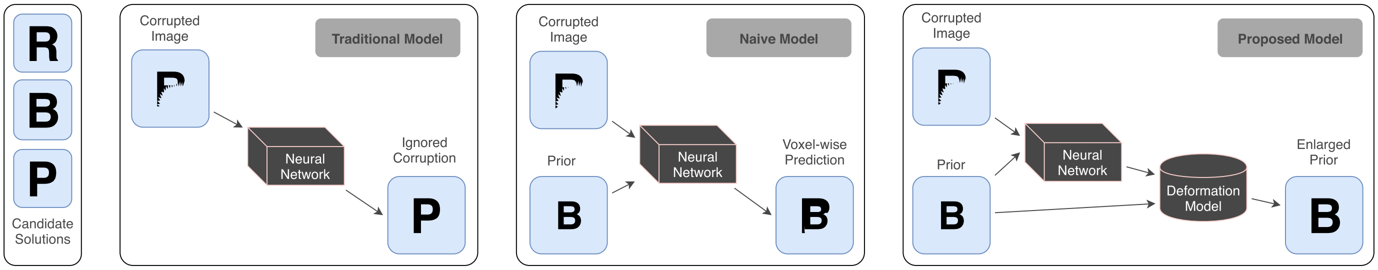


Fig. 1. Schematic to illustrate the differences between traditional, pixel-wise segmentation models, the naive way of incorporating priors through additional input, and our TeTrIS model which produces a set of parameters for a transformation. By restricting the output space of the network to only deformations of the prior, we obtain guarantees on topology.

registration method combining deep convolutional neural nets with spatial transformers.

Another contribution of this paper is the demonstration that state-of-the-art segmentation algorithms can be easily extended to incorporate *implicit* shape priors by providing a shape template as additional input during training. To the best of our knowledge, this simple yet effective enhancement has not been considered in the past. Our benchmarking shows that this can lead to a significant increase in segmentation accuracy, a high level graphical overview of these methods are given in Fig. 1.

We present promising results on coronary artery segmentation from cardiac computed tomography which further strengthens the case for the use of priors in medical image segmentation with deep neural nets. Experimentally we show that all methods which utilize prior information are able to consistently improve the cross entropy score of segmentation, and that our method is able to retain a singly connected component segmentation. Our quantitative results show the varying strengths and weakness of the two introduced methods. We also present qualitative results on synthetic examples to demonstrate the effects of out-of-sample data and how this affects neural network segmentations.

B. Related Work

Our proposed model lies in the intersection of machine learning based image registration and segmentation, and the incorporation of shape priors into neural networks. The following discusses the most related work but due to space limitations and the large amount of work in these fields, this cannot provide a comprehensive overview.

1) Atlas and Registration Based Segmentation: Atlas based segmentation algorithms [16] are among the most popular methods and rely on two key components, an intensity-based image-to-image matching term $\mathcal{L}_\eta$ and a set of training examples, i.e. the atlases, with corresponding labels. During testing, images can be compared to examples in the atlas dataset using $\mathcal{L}_\eta$ and the label mask of the most similar atlases are selected as candidate segmentations. This concept can be extended to employ patch based techniques or advanced label fusion procedures and is robust when label boundaries occur in homogeneous regions. However, this method often offers a coarse segmentation, which may lack precision and can be refined using linear and nonlinear registration. This refinement can be done in either image or segmentation space [17]. The authors of [18] proposed a combination of an image-to-image and segmentation-to-segmentation likelihood function, using a Lagrange multiplier to weight the contribution of each term. If an image-to-segmentation likelihood function is used then this approach is better referred to as template registration [19].

2) Statistical Shape Models: Active shape models as introduced in [20] explicitly model shape based on training examples. By discretizing the k dimensional shapes using n control points, they create point distribution models using an ellipsoid prior. Principle modes of variation can then be found using principle component Analysis (PCA) [21]. New shapes can be represented as a linear weighting of these components. Additionally, by restricting the model to use t principle components where $t < kn$ and restricting the range of values each linear weighting can have, a valid shape space is produced. Active appearance models [22] build on this technique and jointly model appearance together with shape. These models have been use widely in the medical imaging community to perform segmentation [6]. However, such models are heavily biased by the distribution of the training set used to build them.

3) Network Based Image Registration: Traditional registration algorithms take two images, a moving $\mathcal{M}$ and fixed $\mathcal{F}$ and perform registration by iteratively updating some parameterized transformation $\mathcal{T}_\theta$ which maps image grid locations to each other, such that some loss function $\mathcal{L}_\eta(\mathcal{M} \circ \mathcal{T}_\theta, \mathcal{F})$ is minimized, where bespoke parameters θ are found for a given pair of images during test time. Optimization of the algorithm can be considered as optimizing η, some parameterization of the loss function which results in the ‘best’ registrations (such as the values of Lagrange multipliers) or by optimizing the choice of $\mathcal{T}$ (the transformation family expressible).

The key difference between neural network based image registration and traditional, iterative registration algorithms is that the loss function is only computed during training for neural networks. The parameters of the neural network implicitly encode what transformation, conditioned on the input, is needed to register the image with minimal cost instead of repeatedly calculating a loss to iteratively update the parameters θ.

Recent works on neural network based image registration fall into two major categories, the first treats network based registration as a regression problem on a given ground truth deformation field such as in [23]–[27]. These methods, unlike ours, can be used as fast approximations to other registration models. The second group of methods learn the deformation field implicitly while optimizing a downstream task. For example, [28] combine momentum-parametrization for LDDMM shooting [29] and neural networks to learn an end-to-end model for registration. Reinforcement learning approaches have also been used to perform image registration [30]. They treat the registration problem as an iterative update of four translation and two rotation parameters so do not handle free form deformations.

Related network-based registration methods [31] and [32] use 2D spatial transformer networks to embed the deformation model into a neural network pipeline in order to learn the registration model end-to-end. The latter of which uses a FlowNet architecture [24]. This work was built on in [33] which performs full unsupervised 3D registration. However, unlike our model, all of these methods perform image-to-image registration rather than template deformation for a downstream segmentation task. The authors of [34], [35] begin to investigate template deformations but do not investigate this to its full 3D potential.

4) Shape Priors in Neural Networks: Finally we discuss methods which incorporate shape priors into neural networks. Though conditional random fields are considered smoothness priors, they do assist in providing shape consistency in segmentations. CRFs are incorporated into the training process in [12] by casting CRFs as recurrent neural networks. This allows the segmentation and refinement model to be trained end-to-end. Adversarial training was used in [36] as a means of learning such regularization without the need of an explicit model whilst still being able to train end-to-end. A discriminator network was used to distinguish segmentations from a network and ground truth segmentations, this training process encourages the network segmentations to look more plausible. An interleaving process was proposed in [37] where iterative training of a neural network and CRF refinement was performed inspired by the grab cut [38] method, though this model was not trained end-to-end. More recent work has shown improved results by directly incorporating principle component based shape constraints into their learning algorithm. Building on the work of active shape models [20], the authors of [13] use a PCA layer embedded in the neural network to restrict its output space to be weightings of the principle components, this was extended to a probabilistic model in [39]. Another approach proposed in [14] exploits the fact that autoencoders are able to capture a low dimensional representation of the shapes of segmentation maps. This encoding is then used at training time to constrain the outputs of a segmentation network to be close to this low dimensional manifold via adversarial training. The latter two methods utilize anatomical consistency across subjects.

5) Spatial Transformer Networks: Our work builds heavily on spatial transformer networks (STNs) [15] which we describe below. STNs are a neural network model that, conditioned on some input I returns θ for some parameterized transformation model $\mathcal{T}_\theta(G)$. That is $\theta = f_\psi(I)$ where f_ψ is a neural network, itself parameterized by ψ. Once we have θ, we are able to differentiably re-sample our image I to V using $\mathcal{T}_\theta$, as with image registration, here the image itself is re-sampled. The STN model then passes the re-sampled image V to another neural network, $g_\xi(V)$ which performs some down stream task. In [15], they utilize this powerful model to train g_ξ, which performs a classification task and simultaneous train f_ψ a deformation model that makes the down stream task easier via a combination of rescaling, region of interest extraction and rotation of the input images.

During training, the loss is calculated on the down stream task only, for classification this could be the cross entropy loss between the predicted class produced by $g_\xi(V)$ and the true class. Since the neural network g_ξ is a differentiable function and sampling from I to V is also differentiable we are able to train both tasks end to end. Inherently a spatial transformer is performing deformations that assist the down stream task, as opposed to having a loss calculated directly on the task of deforming. This can be considered an implicit registration step, where the registration is autonomously discovered by the network for optimal downstream performance. We do not have to decide what kind of deformation will be good for the task, though we do need to specify the family of deformations $\mathcal{T}$. During test time, unlike iterative registration models, no loss value needs to be calculated. We simply need to perform a forward pass through the network to get both the deformation and the class prediction

II. Template Transformer Networks

Traditional template deformation models require the definition of an image-to-segmentation matching function as an approximation or surrogate to the actual segmentation objective. Iterative optimization is then used to incrementally update the transformation parameters in order to maximize agreement between a template and the image to be segmented. In contrast, our method makes use of network based registration, which only requires the computation of a corresponding loss function (equivalent to the matching function) during *training* time. This important difference means we no longer need to approximate our actual segmentation function via an intensity-based surrogate and can directly optimize for the task at hand.

We introduce a novel template deformation model that exploits the power of neural network-based registration. Our end-to-end model takes a shape prior in the form of a partial volume image (PVI) and an image as input to a neural network which learns to deform the input prior so as to produce an accurate segmentation of the input image. This is done by implicitly estimating a deformation field so as to maximize template alignment corresponding to optimal segmentation accuracy. We provide a detailed description of the main steps below. An overview of our method is shown in Fig. 2. In the following subsections (II-B, II-C and II-D) we describe in detail how we perform deformations, how we regularize our deformation field and how we handle large volume sizes.

A. Obtaining Shape Templates

Shape priors can be utilized in neural networks in various forms such as in the form of level sets, PVIs, binary masks or as shape parameters (e.g., mesh control points).

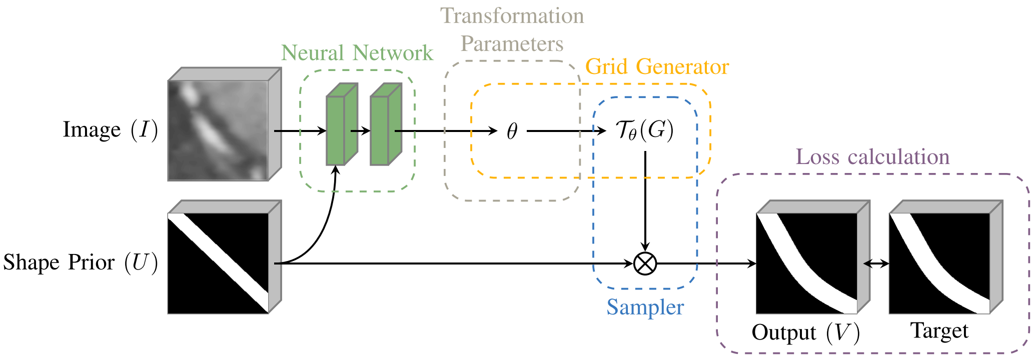


Fig. 2. TETRIS takes as input an image and a shape prior in the form of a partial volume image and produces a set of parameters for a transformation, this transformation is then applied to the prior and the loss is calculated on the deformed prior and the target segmentation during training.

In this work we focus on the use of a deformation model, conditioned on a shape prior to deform a PVI into another PVI. Our shape prior itself in this particular case is also a PVI but we emphasize this is not a necessity and richer priors such as statistical appearance models can also be used. As template transfer networks predict a transformation instead of a point-wise segmentation map, they lend themselves naturally to the ability of using other geometric representations for the priors such as mesh-based models. Shape priors can be generally obtained via manual, semi-automatic and automatic methods and the exact mechanism is application specific. We will later discuss one particular approach for obtaining shape priors for the application of coronary artery segmentation.

B. Deformation Model

To deform a template, consider some input image I, shape prior U, ground truth segmentation T all of size $H \times W \times D$. We have a sampling scheme (or deformation model) $\mathcal{T}_\theta(G)$ where G is considered a standard co-ordinate grid, and loss function $\mathcal{L}_\eta$. $\mathcal{T}_\theta(G)$ is a function on (x_i^t, y_i^t, z_i^t), grid coordinates in our target space, that maps to (x_i^s, y_i^s, z_i^s) co-ordinates in our original source space, where we index voxel locations by $i \in [1, \ldots, H'W'D']$ for notational simplicity. Given this we can define V, a re-sampling of a prior U, based on $\mathcal{T}_\theta$ as

$$V_i = \sum_n^H \sum_m^W \sum_l^D U_{mnl} \times k(x_i^s - m; \Phi_x) \times k(y_i^s - n; \Phi_y) \times k(z_i^s - l; \Phi_z) \quad \forall i \in [1, \ldots, H'W'D'] \qquad (1)$$

where k is any sampling kernel with parameters Φ. For image interpolation we use a trilinear kernel to prevent re-sampled pixel values from being extrapolated to outside of the original intensity domain, that is

$$V_i = \sum_n^H \sum_m^W \sum_l^D U_{mnl} \times \max(0, 1 - |x_i^s - m|) \times \max(0, 1 - |y_i^s - n|) \times \max(0, 1 - |z_i^s - l|) \qquad (2)$$

We choose $\mathcal{T}_\theta$ to be a free form deformation, i.e. θ is a three dimensional vector field. For notational simplicity we define the sampling grid function as

$$\begin{bmatrix} x_i^s \\ y_i^s \\ z_i^s \end{bmatrix} = \mathcal{T}_\theta(G_i) = \begin{bmatrix} x_i^t \\ y_i^t \\ z_i^t \end{bmatrix} - \begin{bmatrix} \theta_i^x \\ \theta_i^y \\ \theta_i^z \end{bmatrix} \qquad (3)$$

If $\mathcal{T}_\theta$ is a free form deformation which is not in the same resolution as the target image we are required to re-sample the deformation field. Potentially using a different set of sampling kernels k with it's own parameters Φ. We choose to use B-Spline interpolation to ensure smooth fields [40], utilizing the Catmull-Rom solution to the interpolation problem [41].

Our method takes inspiration from STNs by using a neural network $f_\psi(I, U)$, which is conditioned on both the input image I and the shape prior U, to produce parameters θ of the B-Spline deformation model $\mathcal{T}_\theta$. We can then perform a deformation of the prior U, calculate a segmentation loss and update the parameters ψ of our network.

By combining template deformation with neural networks, we mitigate the key problem with traditional template deformation models, that being the need to hand craft a good image to segmentation alignment function. The source of this problem, as with any registration technique, lies in the fact that a loss calculation must be made during test time to update the deformation field parameters θ. By utilizing STNs to produce θ during test time and instead updating a neural network f_ψ during training, we can train a registration model with the true segmentation loss function (based on alignment between prior and reference segmentation) avoiding the need for surrogate functions at test time.

The template deformation model is network agnostic so any neural network can be used. We choose a simple feed forward network architecture with convolutions and max pooling to produce a deformation field which we use in the STN to deform the prior. Full details of which are provided in Figs. 3 and 12.

Our method is able to take any shape prior and deform it with sub-pixel accuracy, unlike other neural network based segmentation algorithms which typically treat segmentation as pixel-wise classification. Since our model smoothly deforms a prior, we are able to produce partial volume segmentations, reducing discretization artifacts in final segmentation maps. We provide experiments on both partial volume data as well as voxel-wise classification results.

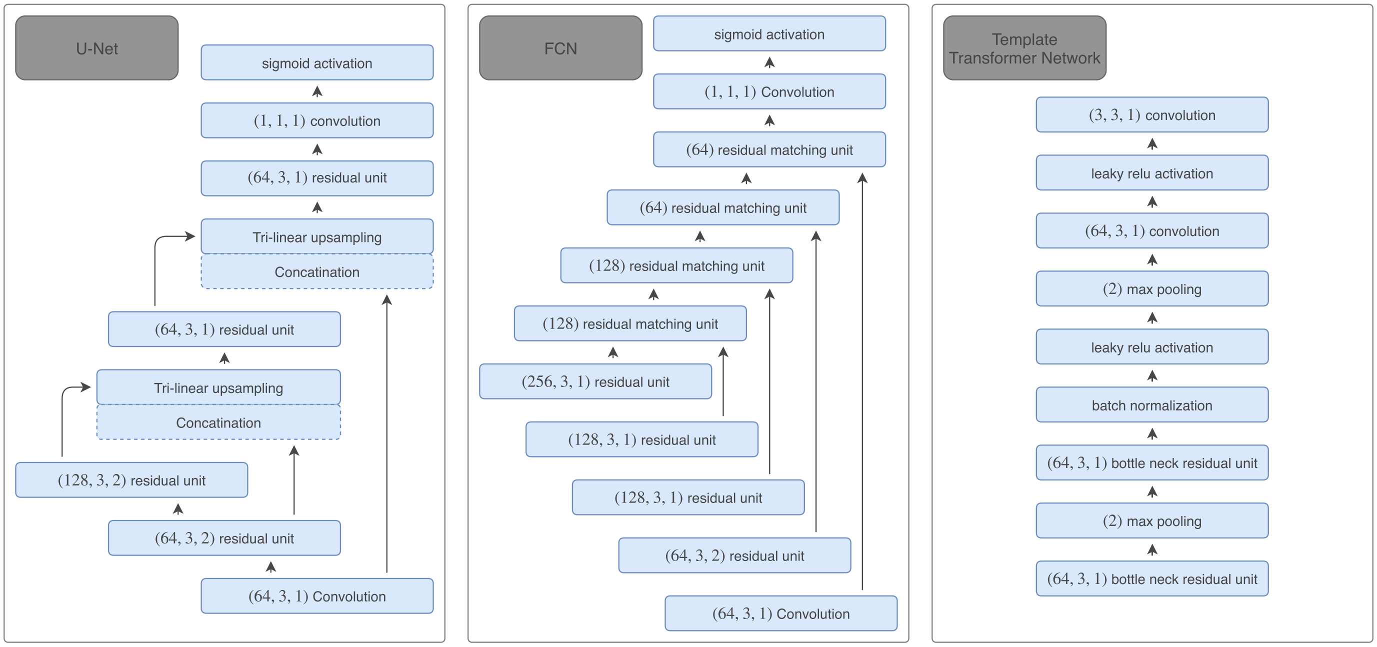


Fig. 3. Graphical representation of the three different models explored in this paper. from left to right, the U-Net, FCN and the black box model used to produce deformation parameters for TeTrIS. Building blocks are described in Fig. 12.

C. Field Regularization

Due to the ill-posed nature of registration problems, it is common to constrain deformation fields by adding a regularization term to the optimization problem that favors some desired property, such as locally smooth deformations, or an $l2$ penalty on the vector field itself to favor minimum displacement solutions. We investigate two regularization terms

$$\mathcal{L}_{l2} := \frac{1}{V}\int_0^X\int_0^Y\int_0^Z T(x,y,z)^2 dxdydz \tag{4}$$

and

$$\begin{aligned}\mathcal{L}_{\text{smooth}} := \frac{1}{V}\int_0^X\int_0^Y\int_0^Z &\left(\frac{\partial^2 T}{\partial x^2}\right)^2 + \left(\frac{\partial^2 T}{\partial y^2}\right)^2 \\ &+ \left(\frac{\partial^2 T}{\partial z^2}\right)^2 + 2\left(\frac{\partial^2 T}{\partial zx}\right)^2 + 2\left(\frac{\partial^2 T}{\partial xy}\right)^2 \\ &+ 2\left(\frac{\partial^2 T}{\partial yz}\right)^2 dxdydz \end{aligned} \tag{5}$$

$\mathcal{L}_{l2}$ penalizes the $l2$-norm of the field and $\mathcal{L}_{\text{smooth}}$ penalizes the sum of squared second order derivatives.

D. Field Aggregation

To deal with the size of the data and the memory restrictions of modern graphics processing units we do inference on a patch basis, we collect control points across patches and aggregate them before re-sampling using B-Spline interpolation. This combined with only using valid padding prevents ill-posed boundary conditions across the image. This also allows us to perform inference on variable size volumes with consistent control point spacing without modification to the neural network.

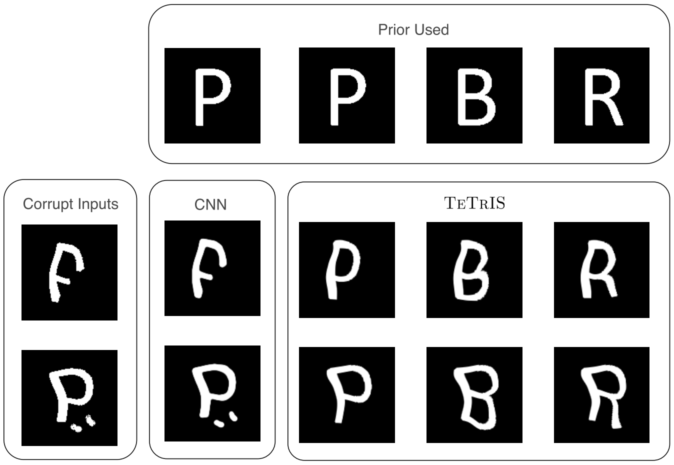


Fig. 4. Examples of where a deformation model can extrapolate well outside of the distribution of the training data compared to a standard convolutional neural network.

III. Illustrative Example

As a proof of concept, we present qualitative results on the effects of corruption in the data as these are not easily quantifiable. To investigate how incorporating a prior into a neural network can help when corruption is present, we create a toy dataset of 1500 randomly deformed P's, B's and R's for training and two hand crafted test images which we provide qualitative results for. We then train a deformation model to deform the prior (the letter that was originally deformed) to match the deformed letter. Additionally, we train a normal convolution neural network to predict the deformed letter on a pixel-wise level. Both the TeTrIS model and the convolutional neural network are conditioned on the prior and the target. As is expected, when the image signal is strong,

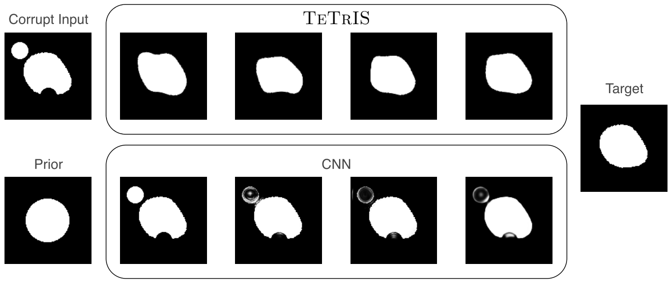


Fig. 5. Qualitative results from varying the amount of corrupted training examples in the dataset. From left to right, the models are trained with 0%, 5%, 10% and 15% of the training set consisting of corrupted examples.

TABLE I
QUANTITATIVE SYNTHETIC DATA RESULTS

Corruption in training set	Dice score		Connected Components		Hausdorff Distance	
	CNN	TETRIS	CNN	TETRIS	CNN	TETRIS
0%	0.9409	0.9441	1.80	1.02	25.56	7.60
5%	0.9720	0.9256	4.80	1.00	22.10	7.47
10%	0.9775	0.9268	3.98	1.00	23.25	6.84
15%	0.9710	0.9537	2.11	1.00	20.41	5.80

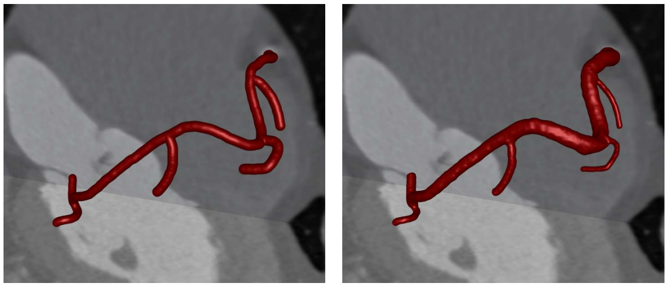
the network learns to rely heavily on the image signal and ignores the prior, this can be seen in Fig. 4. We trained both models on only uncorrupted deformations to see how each model can handle an out-of-sample test case. We see that the vanilla CNN learns to completely ignore prior information, so when inferring on corrupt data, it is not able to extrapolate, unlike the TETRIS model. By restricting our model's output space to be within the range of deformations of the prior we are able, even in the presence of corruption to produce plausible results, consistent with our prior.

IV. SYNTHETIC EXPERIMENTS

We argue that for a CNN to handle such corruption it would need to be present in the training set, Fig. 5 shows the effects of having an increased amount of corrupted data in the training set. We construct a secondary dataset where corruption is more easily generated which consists of 1000 randomly deformed discs, where corruption is in the form of smaller discs being cut from the main central disc and smaller peripheral discs being placed around the main central disc and the set is split in half for training and testing. We trained the models with 0%, 5%, 10% and 15% of the training set consisting of corrupted examples. As more corruption is present in the training set, the better the standard CNN model is able to handle them during test time as expected. Though the artifacts that occur are not topologically as plausible as those produced by our TETRIS model, which is reflected in the high dice scores but also high Hausdorff distances.

V. CORONARY ARTERY SEGMENTATION EXPERIMENTS

In this paper we focus on the application of vessel segmentation where ambiguities arise from the functional distinction between veins and arteries, which may have similar image features. This has lead some methods to approach the problem as a multi stage process, first centerlines are extracted [42], then the vessels are segmented [43]. Shape priors can be enforced once good candidate centerlines have been extracted by treating the segmentation task as a wall distance estimation task. By utilizing curved planar reformation [44] the segmentation problem can be cast as a wall distance regression from the centerline and topology can be guaranteed.



(a) (b)

Fig. 6. An example of the shape prior on the left and manual segmentation on the right. The shape prior is a tubed human annotated centerline with a fixed one millimeter radius. (a) Tubed prior. (b) Target segmentation.

We train our network on a set of 274 annotated cardiac CT volumes with 0.5 millimeter isotropic spacing and reserving 138 volumes for validation and an additional 136 for testing. The ground truth labels obtained through manual expert segmentation are in the form of partial volumes.

A. Generating Priors for Coronary Arteries

To generate a shape prior for coronary artery segmentation, we first extract out a centerline using a semi-automatic method which consists of a Random Forest voxel-wise classification, a Dijkstras shortest path based tree extraction and finally a human review and correction step to correct outliers. The centerline is converted to a 3D volumetric representation by creating a tube around it with a fixed radius of 1 mm in a partial volume image, since the centerline exists in arbitrary space rather than voxel space. More examples of coronary centerline extraction method in cardiac CTA can be found in [42]. Fig. 6 is a volume rendering illustrating the difference between the prior and the ground truth segmentation in an example vessel. More generally, priors can be extracted from sources such as automated algorithms, weak labels, human expert knowledge or population based statistics and will inherently be application specific.

B. Model Details and Baselines

We use two baseline models to compare the three models we present i) the residual fully convolutional network (FCN) and ii) a residual U-net architecture utilizing the implementations from [45] using residual blocks from [46]. Details of the architectures can be found in Fig. 3, where the building blocks are described in Fig. 12.

We also present results on naively incorporating shape priors into these state-of-the-art models. We do this by feeding the networks two channels of input, the image to be segmented and the prior that we have of the image at that location. This alternative method is a very simple extension of existing state-of-the-art pixel-wise approaches, computationally cheap and easy to implement. The shape prior, in this case, acts as a kind of initialization for the network's output.

C. Training Details

For all models we use the same patch extraction parameters, during training we dynamically extract 32 patches from each volume and randomly shuffle them into a buffer of 512 patches. Patches are extracted if they are near the centerline, biasing the sampling around the vessel. We use a batch size of 8 for all models and train them using the Adam optimizer [47] while exponentially decaying the learning rate. The learning rate at step i is as defined as

$$l_i = l_0 \cdot r^{\frac{i}{s}} \tag{6}$$

where our initial learning rate $l_0 = 1 \cdot 10^{-5}$, decay rate $r = 0.99$, decay step $s = 500$ and where regularization is used, we weight it by $5 \cdot 10^{-6}$.

We pretrain our baseline models using a weighted cross entropy function as defined in Equation 7, where p is our target distribution, q is our candidate distribution and w is a weighting factor. By setting $w > 1$, we bias the loss term to penalize false negatives. This is beneficial as voxels containing the vessel interior are sparse in any given patch. Penalizing false negatives more prevents the network from predicting all voxels as background voxels during the initial stages of optimization, a trivial local optimum. Note this does not need to be done with our TETRIS model as the network is already biased towards the identity transform thanks to our regularization term which favors a smooth deformation field. For our experiments we set $w = 2$ and pretrain our non-TETRIS models for 1000 iterations.

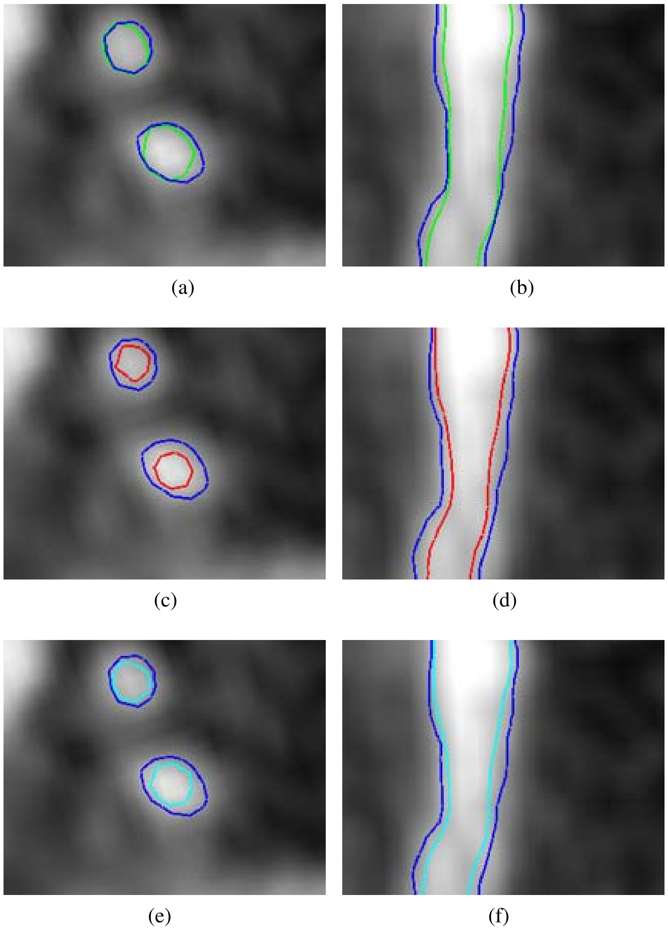
$$-w \cdot p \log(q) - (1 - p) \log(1 - q) \tag{7}$$

We fine tune the models using normal, un-weighted, cross entropy for a further 5000 iterations.

D. Results

Results are presented on a test set of 136 cases, for the task of partial volume estimation we use the cross entropy as a measure of accuracy, we can see from Table II that incorporating shape priors into state of the to state-of-the-art neural network segmentation models significantly improves results. For comparison we include results on using the identity function on the prior, i.e. naively taking the shape prior as the segmentation.

We provide box plots of the results in Fig. 9 for a more fine grained break down of the results, where we have plotted the cross entropy on a log scale. Though our model with $l2$ field regularizations performs the best, there is no significant difference between the methods, exhibiting the expressiveness of a deformation model despite constraining the output space to be a deformation of the prior.


(a) (b)

(c) (d)

(e) (f)

Fig. 7. Close up of qualitative results shown as contours for the different methods where the blue, green, red and cyan contours are of the target segmentation, TETRIS, FCN (with prior) and U-Net (with prior) respectively. On the left and right are orthogonal views of the left anterior descending artery, near the first diagonal bifurcation where TETRIS outperforms other methods. (a) TETRIS-$l2$. (b) TETRIS-$l2$. (c) FCN (w/ prior). (d) FCN (w/ prior). (e) U-Net (w/ prior). (f) U-Net (w/ prior).

TABLE II
QUANTITATIVE SEGMENTATION RESULTS ON TEST CASES

	Cross Entropy	Connected Components	Dice Score	Hausdorff Distance	Trainable Parameters
U-Net	0.01219	26.4	0.336	73.41	11.94M
FCN	0.01190	27.0	0.406	59.94	13.74M
U-Net (w/ prior)	0.00186	1.0	0.854	2.86	11.95M
FCN (w/ prior)	0.00163	1.1	0.790	2.87	13.75M
TETRIS-no-reg	0.00162	1.0	0.779	3.20	1.38M
TETRIS-$l2$	0.00160	1.0	0.787	3.36	1.38M
TETRIS-smooth	0.00163	1.0	0.768	3.55	1.38M

Both our U-Net (with prior) model and our proposed TETRIS model are able to consistently produce singly connected components without post processing by incorporating prior information, further demonstrating the potential of our approachs. However the FCN (with prior) model often does not capture these higher order requirements, even with prior information, we believe this is due to the inherent multi-scale nature of the U-Net architecture. TETRIS shines when using metrics that take into account partial volumes, however our CNNs with priors added as input channels work consistently well when the goal is pixel-wise binary segmentation. We also

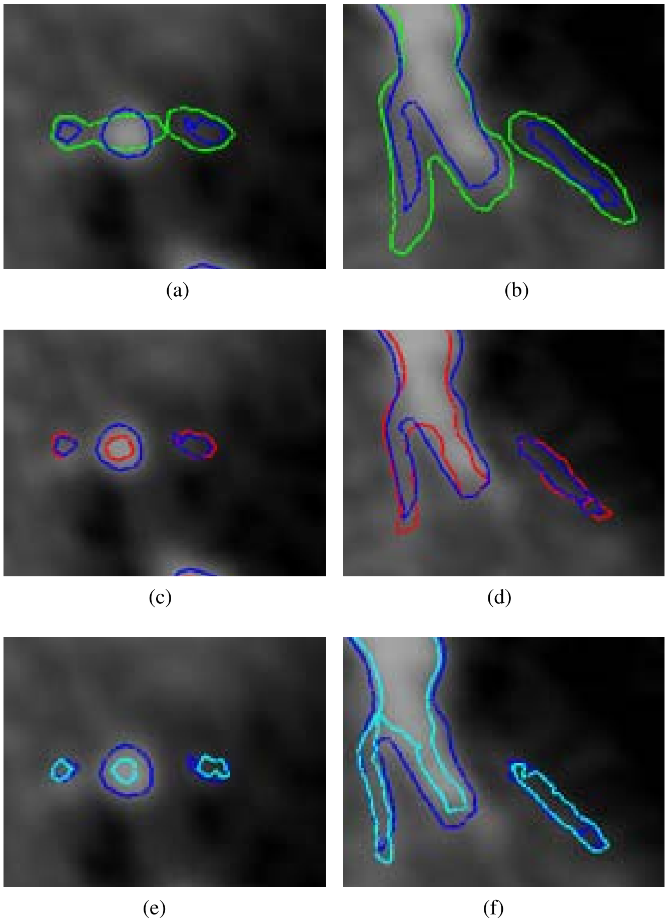


Fig. 8. Close up of qualitative results shown as contours for the different methods where the blue, green, red and cyan contours are of the target segmentation, TETRIS, FCN (with prior) and U-Net (with prior) respectively. On the left and right are orthogonal views of a trifurcation where TETRIS over segments and the FCN and U-Net model under segment. (a) TETRIS-*l*2. (b) TETRIS-*l*2. (c) FCN (w/ prior). (d) FCN (w/ prior). (e) U-Net (w/ prior). (f) U-Net (w/ prior).

note a drastic reduction of trainable parameters by a factor of ten for TETRIS compared to U-Net and FCN, indicating a better balance between performance and model complexity.

To obtain the number of connected components, we threshold the partial volume segmentations at 0.5 and perform a 26-connected component analysis. Ideally, all segmentations should have only one connected component. We note that TETRIS without regularization may result in discontinuous segmentations, but did not find this to be the case in practice. We notice no major difference between penalizing the field with an *l*2 penalty or by the sum of second order derivatives. Fig. 10 shows an example case where neural networks are not able to recover the vessel in the image without prior information, where as all three our models are able to fall back on the prior when the image signal may be weak.

We further investigated the use of more complex models for TETRIS but found that convergence became slow and often resulted in similar validation scores, hence our choice for a simple TETRIS model. We notice that the models also have different strengths and weaknesses, as mentioned previously, the U-Net (with prior) model is better than the FCN (with prior) model at capturing global consistency of shape but in regions where contrast is low, our TETRIS model produces smoother more accurate segmentations as seen in Fig. 7.

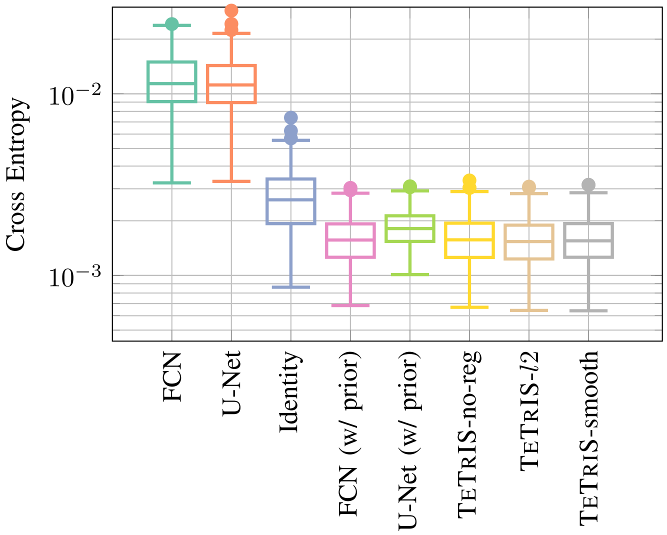


Fig. 9. Cross entropy for partial volume estimation of test cases for the different methods investigated, clearly demonstrating the benefits of incorporating prior information and the ability of a deformation model to perform just as well, if not better than an standard neural network which naively incorporates prior information.

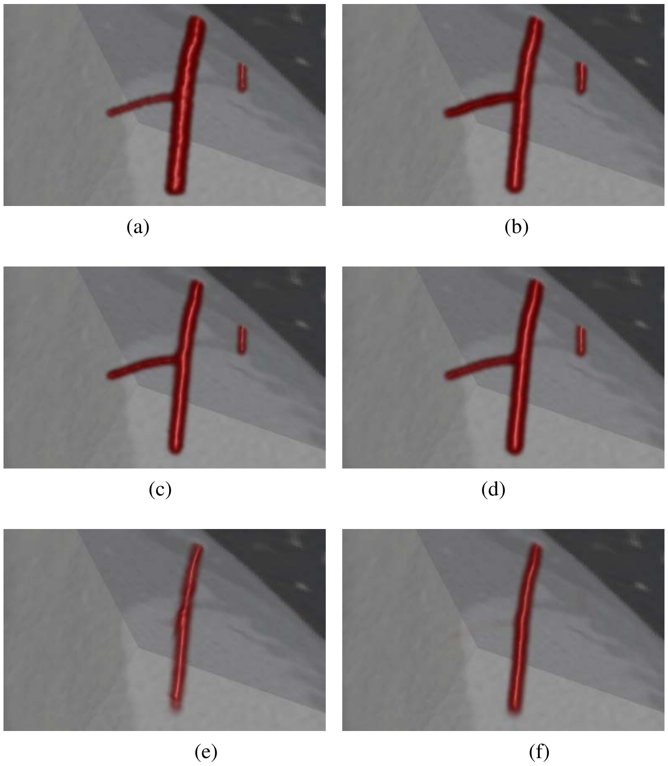


Fig. 10. Close up of qualitative results shown as volume rendering for the different methods with and without shape priors compared to the manual target segmentation. (a) Target segmentation. (b) TETRIS-*l*2. (c) FCN (w/ prior). (d) U-Net (w/ prior). (e) FCN. (f) U-Net.

Using a deformation model does have caveats, in Fig. 8 we see a trifurcation region where TETRIS over-segments and both the U-Net/FCN with prior under-segment. The resolution

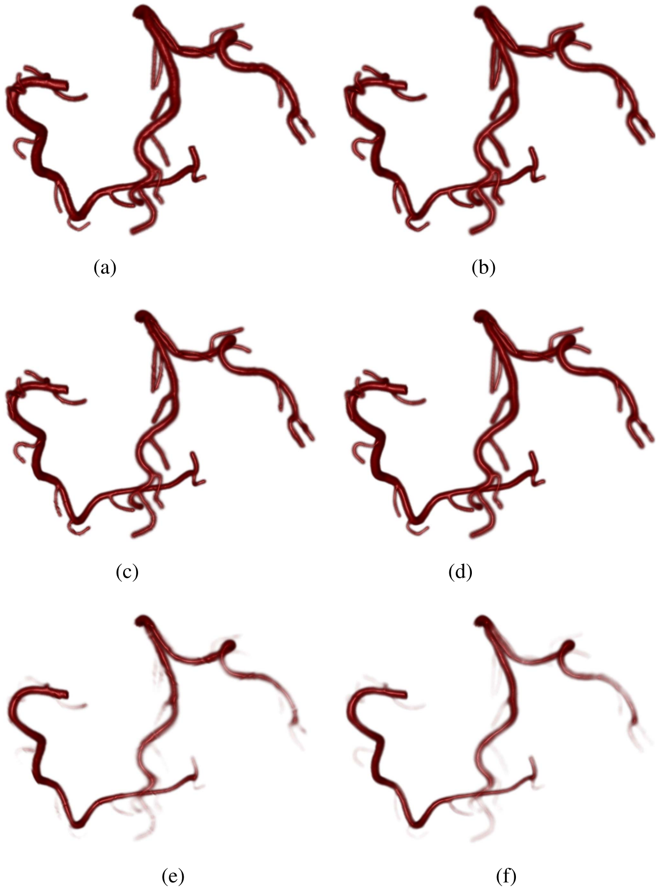


Fig. 11. Qualitative results shown as volume rendering for the different methods with and without shape priors compared to the manual target segmentation. The transfer function from partial volume probability to opacity is set as the identity function from [0,1] → [0,1]. (a) Target segmentation. (b) TETRIS-/2. (c) FCN (w/ prior). (d) U-Net (w/ prior). (e) FCN. (f) U-Net.

of the field and the penalty applied to large deformations prevents our model from doing well in such regions.

In summary, we should highlight the advantages of the template transformer based networks over point-wise segmentation models such as U-net and FCN, as it might not be apparent from the segmentation scores. Although, U-Net performs best on Dice, TETRIS performs better on cross-entropy assessing the agreement for the soft, partial volume predictions. Additionally, the template transformer networks can provide guarantees on the resulting shape while both U-Net and FCN do not. This can be important in applications where the segmentations are used for downstream tasks such as shape analysis or blood flow calculation. One other important benefit of the TETRIS model (although not explored in this work) is the ability to incorporate a variety of shape priors such as mesh-based representations or probabilistic shape and appearance models (e.g. a mean and variance image).

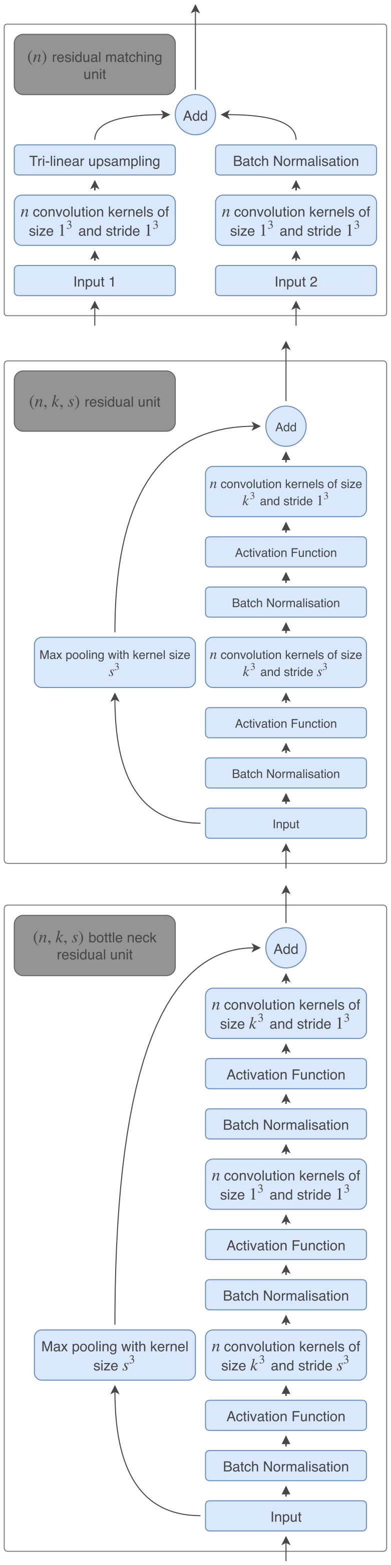

(n) residual matching unit
Add
Tri-linear upsampling
Batch Normalisation
n convolution kernels of size 1^3 and stride 1^3
n convolution kernels of size 1^3 and stride 1^3
Input 1
Input 2
(n, k, s) residual unit
Add
n convolution kernels of size k^3 and stride 1^3
Activation Function
Batch Normalisation
Max pooling with kernel size s^3
n convolution kernels of size k^3 and stride s^3
Activation Function
Batch Normalisation
Input
(n, k, s) bottle neck residual unit
Add
n convolution kernels of size k^3 and stride 1^3
Activation Function
Batch Normalisation
n convolution kernels of size 1^3 and stride 1^3
Activation Function
Batch Normalisation
Max pooling with kernel size s^3
n convolution kernels of size k^3 and stride s^3
Activation Function
Batch Normalisation
Input

Fig. 12. Graphical representation of the building blocks used to construct all models.

VI. Discussion

We introduced Template Transformer Networks for image segmentation which are able to deform shape priors into segmentations. This work builds on template deformations by no longer requiring the need for hand-crafted image to segmentation cost functions and makes use of Spatial Transformer Networks for differentiable end-to-end learning. Our method is competitive with state of the art segmentation algorithms while being able to guarantee topological constraints.

Our work is a proof of concept which relied on a simple architecture that can be easily extended. Though our model is restricted in the sense that it can only perform deformations of a prior, we argue this can be an advantage where shape guarantees are important. Arguably, our model strikes a better balance between performance and model complexity due to a significantly fewer number of trainable parameters.

We consider the prior extraction beyond the scope of this work, but we believe that it is a critical part of not only our method, but all methods which require a prior. In problems where no sensible priors can used template based methods are likely not suitable. Though not explored in this work, our approach lends itself to the incorporation of much richer priors, such as probabilistic shape priors and other geometric representations such as meshes or point distribution models.

Our method replaces an iterative method with a one-shot method, we believe a natural extension to investigate would be to incorporate template deformations with recurrent or auto-regressive neural networks for more flexible and potentially larger deformations, mitigating the effects of the chosen resolution for the control point grid. Though in this work we chose to use B-Splines, our method is agnostic to the choice of parameterization of the deformation field. The exploration of other and potentially more flexible parameterizations is also of great interest. Additionally, we would also like to explore the use of deformation fields which are not on fixed grids so as to allow for finer deformation fields as and when is needed without the burden of excess computation.

Appendix

We provide full examples of vessel segmentations in Fig. 11 to give the reader larger context into the accuracy of the model, where we have trimmed the aorta for clearer visualizations. Without prior information it is clear that the sensitivity of the networks drop substantially.

Acknowledgement

The authors would like to thank Sarah Parisot, Katarína Tóthová, Ozan Oktay and Sofy Weisenberg for insightful feedback on the manuscript as well as Vishwanath Sangale, David Lesage, James Batten and Karl Hahn for fruitful discussions. They would also like to thank the reviewers for their constructive feedback.

References

[1] M. S. Nosrati and G. Hamarneh. (2016). "Incorporating prior knowledge in medical image segmentation: A survey." [Online]. Available: https://arxiv.org/abs/1607.01092

[2] A. L. Yuille, D. S. Cohen, and P. W. Hallinan, "Feature extraction from faces using deformable templates," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1989, pp. 104–109.

[3] G. Subsol, J.-P. Thirion, and N. Ayache, "A scheme for automatically building three-dimensional morphometric anatomical atlases: Application to a skull atlas," *Med. Image Anal.*, vol. 2, no. 1, pp. 37–60, 1998.

[4] T. Kapur, P. Beardsley, S. Gibson, W. Grimson, and W. Wells, "Model-based segmentation of clinical knee MRI," in *Proc. IEEE Int. Workshop Model-Based 3D Image Anal.*, Jan. 1998, pp. 97–106.

[5] D. Perperidis *et al.*, "Building a 4D atlas of the cardiac anatomy and motion using mr imaging," in *Proc. IEEE Int. Symp. Biomed. Imag., Nano Macro*, Apr. 2004, pp. 412–415.

[6] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Med. Image Anal.*, vol. 13, no. 4, pp. 543–563, 2009.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, May 2015, pp. 234–241.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2016). "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1606.00915

[10] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[11] F. Milletari, N. Navab, and S.-A. Ahmadi. (2016). "V-Net: Fully convolutional neural networks for volumetric medical image segmentation." [Online]. Available: https://arxiv.org/abs/1606.04797

[12] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1529–1537.

[13] F. Milletari, A. Rothberg, J. Jia, and M. Sofka, "Integrating statistical prior knowledge into convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.* Cham, Switzerland: Springer, 2017, pp. 161–168.

[14] O. Oktay *et al.*, "Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 384–395, Feb. 2018.

[15] J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, "Deep transfer metric learning," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5576–5588, Dec. 2016.

[16] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.

[17] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens, "An information theoretic approach for non-rigid image registration using voxel class probabilities," *Med. Image Anal.*, vol. 10, no. 3, pp. 413–431, 2006.

[18] W. Bai *et al.*, "A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: Application to cardiac MR images," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1302–1315, Jul. 2013.

[19] K. A. Saddi, C. Chefd'hotel, M. Rousson, and F. Cheriet, "Region-based segmentation via non-rigid template matching," in *Proc. IEEE 11th Int. Conf. Comput. Vis. (ICCV)*, Oct. 2007, pp. 1–7.

[20] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.

[21] B. Flury, *Multivariate Statistics: A Practical Approach*. London, U.K.: Chapman & Hall, 1988.

[22] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[23] H. Uzunova, M. Wilms, H. Handels, and J. Ehrhardt, "Training cnns for image registration from few samples with model-based data augmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham, Switzerland: Springer, 2017, pp. 223–231.

[24] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2758–2766.

[25] H. Sokooti, B. de Vos, F. Berendsen, B. P. F. Lelieveldt, I. Išgum, and M. Staring, "Nonrigid image registration using multi-scale 3D convolutional neural networks," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham, Switzerland: Springer, 2017, pp. 232–239.

[26] X. Cao *et al.*, "Deformable image registration based on similarity-steered CNN regression," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham, Switzerland: Springer, 2017, pp. 300–308.

[27] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "SVF-Net: Learning deformable image registration using shape matching," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham, Switzerland: Springer, 2017, pp. 266–274.

[28] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration—A deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, Sep. 2017.

[29] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *Int. J. Comput. Vis.*, vol. 61, no. 2, pp. 139–157, 2005.

[30] K. Ma, J. Wang, V. Singh, B. Tamersoy, Y.-J. Chang, A. Wimmer, and T. Chen, "Multimodal image registration with deep context reinforcement learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.* Cham, Switzerland: Springer, 2017, pp. 240–248.

[31] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum. (2017). "End-to-end unsupervised deformable image registration with a convolutional neural network." [Online]. Available: https://arxiv.org/abs/1704.06065

[32] S. Shan, X. Guo, W. Yan, E. I.-C. Chang, Y. Fan, and Y. Xu. (2017). "Unsupervised end-to-end learning for deformable medical image registration." [Online]. Available: https://arxiv.org/abs/1711.08608

[33] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9252–9260.

[34] H. Zhang and X. He, "Deep free-form deformation network for object-mask registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2017, pp. 4251–4259.

[35] C. Qin *et al.* (2018). "Joint learning of motion estimation and segmentation for cardiac mr image sequences." [Online]. Available: https://arxiv.org/abs/1806.04066

[36] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. (2016). "Semantic segmentation using adversarial networks." [Online]. Available: https://arxiv.org/abs/1611.08408

[37] M. Rajchl *et al.*, "DeepCut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 674–683, Jun. 2017.

[38] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[39] K. Tóthová *et al.* (2018). "Uncertainty quantification in CNN-based surface prediction using shape priors." [Online]. Available: https://arxiv.org/abs/1807.11272

[40] D. Rueckert, P. Aljabar, R. A. Heckemann, J. V. Hajnal, and A. Hammers, "Diffeomorphic registration using B-splines," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, vol. 9. Berlin, Germany: Springer, 2006, pp. 702–709.

[41] E. Catmull and R. Rom, "A class of local interpolating splines," in *Computer Aided Geometric Design*. Amsterdam, The Netherlands: Elsevier, 1974, pp. 317–326.

[42] M. Schaap *et al.*, "Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms," *Med. Image Anal.*, vol. 13, no. 5, pp. 701–714, 2009.

[43] H. A. Kirişli *et al.*, "Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography," *Med. Image Anal.*, vol. 17, no. 8, pp. 859–876, 2013.

[44] A. Kanitsar, D. Fleischmann, R. Wegenkittl, P. Felkel, and M. E. Gröller, "CPR: Curved planar reformation," in *Proc. Conf. Vis.*, 2002, pp. 37–44.

[45] N. Pawlowski *et al.* (2017). "DLTK: State of the art reference implementations for deep learning on medical images." [Online]. Available: https://arxiv.org/abs/1711.06853

[46] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385

[47] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980