Shape Feature Loss for Kidney Segmentation in 3D Ultrasound Images

Haithem Boussaid[1]

James Jago[2]

Laurence Rouet[1]

[1] Philips Research
Suresnes, FR

[2] Philips Ultrasound
Bothell, USA

Abstract

Kidney segmentation from 3D ultrasound images remains a challenging task due to low signal-to-noise ratio and low-contrasted object boundaries. Most of recently proposed segmentation CNNs rely on loss functions where each voxel is treated independently, which does not convey the overall high-dimensional structure of a 3D organ. Such approaches fail to produce regularly shaped segmentation masks especially in complex cases. In this work, we design a loss function to compare segmentation masks in a feature space designed to describe explicit global shape attributes. We use a Spatial Transformer Network to derive the 3D pose of a mask and we project the resulting aligned mask on a linear sub-space describing the variations across objects. The resulting shape-feature vector is a concatenation of weighted shape rigid pose parameters and non-rigid deformation parameters with respect to a mean shape. We use the L1 function to compare the prediction and the ground-truth shape-feature vectors. We validate our method on a large 3D ultrasound kidney segmentation dataset. Using the same U-net Architecture, our loss function outperforms dice and cross entropy standard loss functions used in the nnU-net state-of-the-art approach.

1 Introduction

Current clinical practice for kidney volume assessment is based on the combination of 2D diameter measurements, from 2D Ultrasound (US), combined with an ellipsoid model [1] and is known to have an error of estimation around 25% . Automatic kidney volume quantification from 3D US is foreseen as a mean to improve patient surveillance in the context of renal diseases, such as chronic kidney disease (CKD). The acquired images often suffer from limited image quality, shadows, making the segmentation process challenging for a non-expert sonographer. Moreover, depending on both the considered acquisition protocol and anatomies, kidneys exhibit high variability with respect to their 3D pose, geometry and appearance.

To tackle this problem, classical methods in medical image shape segmentation use prior knowledge about the object shape to be segmented [12, 13, 14]. This expertise is acquired from previously seen objects and expressed through a model. The model describes a class of local [16] or global [15] statistical relationships between shapes to be extracted. By incorporating the desirable properties of the envisioned solution, these methods tend to give regularized, plausible shapes but are not necessarily faithful to the considered image due to limitations of the used hand-crafted image features and/or shadow learning and inference techniques.

With the advent of deep learning based segmentation methods, this long standing idea of combining a feature extraction model with a prior knowledge regularization model has faded out. Recent works henceforth rely on the deep neural networks to be able to implicitly learn high order relations between the input image and the desired high-dimensional structured segmentation mask, eventhough they use loss functions designed for voxel-wise classification. CNNs risk to over-fit to only variability seen in the training dataset and interpretability of neural network learned expertise is still an open question. In our experiments with challenging ultrasound segmentation tasks, we witnessed several failure cases with current CNNs.

Since the introduction of the popular U-net paper [3], there was not a major improvement to dethrone the proposed encoder-decoder with skip connections architecture. Dilated convolutions [21], attention mechanism [19], residual connections [17], dense connections [18], and squeeze and excitation [20] were since then incorporated into U-net. [2] empirically demonstrate that these architectures bring little systematic improvement agnostic to the considered dataset. Other recent research papers focus on designing 'better' loss functions. Cross entropy and Dice are 'Gold Standard' [2, 11]. Sensitivity-specificity loss [25], IoU loss [26], Tversky loss [27] Penalty loss [28] and Hausdorff Distance loss [29] are related to the Dice loss [11]. Weighted cross entropy [3], TopK loss [22] Focal loss [23] and Distance map penalized cross entropy loss [24] are derived from the Cross entropy loss. Overall, these losses are still limited to voxel-to-voxel comparison and do not fully take into account the underlying semantic information and dependencies in the output space.

Recently, a number of works attempted to incorporate prior shape knowledge into CNNs for object segmentation. [4] used a convolutional auto-encoder to learn a non-linear compact representation of the underlying anatomy. The encoder part is then plugged with a standard segmentation network as a shape regularisation loss. It encourages the CNN to predict segmentations that lie on the extracted low dimensional data manifold. It is unclear though if the learned encoder managed to learn global geometric variability like rotation and translation. A complex disentanglement technique is needed to interpret and validate the learned variability. The Adversarial loss was also considered for segmentation to progressively build a prior on the space of feasible segmentations [5]. The discriminator learns an implicit feature space to distinguish between fake and real segmentations. However, GAN-based techniques still suffer from hallucinating artifacts and are difficult to train. The perceptual loss [6] uses features from intermediate layers of a generic pre-trained network. It was proposed for image-to-image translation tasks and super resolution. This loss might need adaptation when considered for segmentation tasks since segmentation mask statistics are different from natural images statistics. [7] introduced a PCA-based loss and showed promising results on 2D X-ray cervical vertebra images. However, no pose variability was treated which limits the application of the method to only already aligned datasets. [32] used the encoder part of an encoder-decoder framework to predict the shape and pose parameters. In our experiments, we found it difficult to regress directly these heterogeneous parameters with a mean squared error loss.

In this work, we keep the successful U-Net based CNN architecture and augment it with a shape feature loss, noted L_{sf}, which uses a simple explicit and interpretable shape feature space as a compact representation of kidneys. L_{sf} enforces network predictions to follow the learnt statistical shape distributions. In particular, we use a Spatial Transformer Network (STN) to reveal the 3D pose of a predicted segmentation mask and transform it to a normalized pose. Then we project this normalized shape to obtain a feature vector describing shape deformation with respect to a mean shape. We use the L1 function to compare the prediction and the ground-truth shape-feature vectors. Furthermore, in order to guarantee optimal contribution of each loss component, we use a multi-task learning technique [8] to automatically calibrate the weights assigned to each term of the loss function. This technique adjusts the weights automatically during training according to the network homoscedastic uncertainty [8]. Our method is validated on a large 3D ultrasound kidney segmentation dataset. Using the same U-net Architecture, our loss function outperforms Dice and cross entropy standard loss functions.

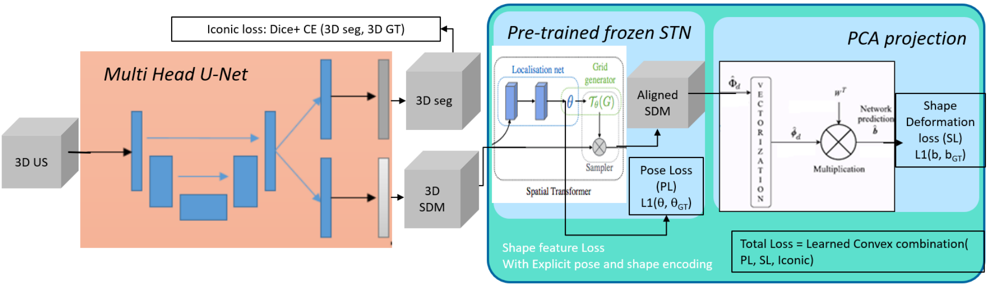

Figure 1: Our network architecture introduces the shape feature loss, L_{sf}. The pose loss (PL) part constrains the segmentation pose using a pre-trained frozen STN. The Shape Deformation loss (SL) part constrains the normalized pose segmentation anatomically using a PCA model. A learned combination calibrates the contribution of each loss component.

2 Preliminaries

2.1 Ground Truth Generation

We convert each available training ground-truth segmentation mask to a signed distance map (SDM) -noted $\mathbf{y}$- using Danielsson algorithm [31]. An SDM represents shape in a nonparametric manner. The shape is defined implicitly as the set of its SDM. We then obtain aligned SDMs -noted $\mathbf{y_N}$- with a reference SDM by using a classic iterative rigid registration algorithm. This allows us to also obtain the ground-truth 3D pose parameters θ_{GT} of each $\mathbf{y}$

2.1.1 SDM based Statistical Shape Model

With the obtained aligned training SDMs, we build a statistical shape model (SSM) of the kidney. We apply Principal Component Analysis (PCA) on normalized pose ground-truth Signed distance maps . This allows representing each SDM as

$$\mathbf{y_N} = \bar{\mathbf{y}}_\mathbf{N} + W\mathbf{b} \tag{1}$$

where $\mathbf{\bar{y}_N}$ is the mean normalized pose SDM, W is the eigenvectors matrix and $\mathbf{b_{GT}}$ is a vector of shape parameters. The resulting SSM model explains 19 modes of variation, that represent 99% of the shape variability.

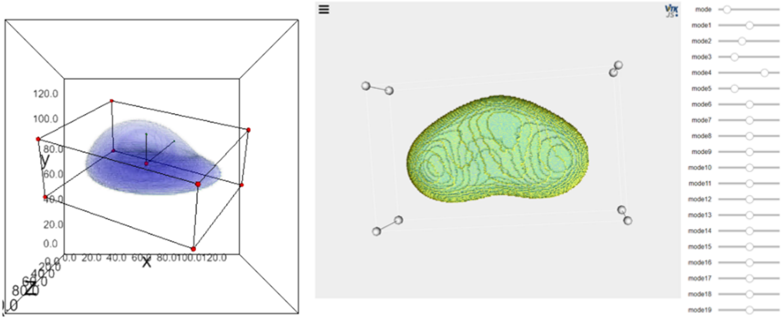

Figure 2: Compact representation of a kidney shape. A feasible kidney shape is described by its pose parameters θ_{GT} (left) and its global deformation parameters with respect to a mean shape $\mathbf{b_{GT}}$ (right). The shape is represented by a 7+19 dimensional feature vector.

2.2 SDM pose regression

We pre-train a Supervised Spatial Transformer Network (STN) that takes an SDM as input, predicts its 3D pose and transform it to a normalized pose in a differentiable manner. We augment our dataset with synthetic data by applying random rigid transformation and composing the corresponding rigid pose. We use a 3D resnet architecture for the localisation network part F_{loc}, which is less prone to over-fitting. Instead of using only a mean squared error (MSD) loss on pose parameters, we also used a MSD loss on normalized pose SDMs in addition to the aforementioned loss. This allowed us to obtain more stable training and better overall performance. Once the training finished, the STN is henceforth frozen to be used inside our loss function.

3 Backbone Model

The No New U-Net (nnU-Net) [2] paper compiles best practices in the use of U-net for 3D medical image segmentation. It automatically finds optimal hyper-parameters suited to the considered dataset. It showed state-of-the-art performance in 3D medical image segmentation on several different datasets and won several challenges such as the Medical Segmentation Decathlon. It showed best performance when using a combination of Dice and cross entropy loss. We consider the nnU-Net implementation as our standardized baseline and we build upon it to incorporate prior shape knowledge regularization loss.

Since our loss involves shape properties, we augment the nnU-Net to predict also an SDM as shown in Figure 1 to represent a shape as proposed in [10]. A recent comparison study [9] showed that best segmentation performance is achieved when using a multi head U-net with two predictions, a segmentation and an SDM regression. This is inherent to multi task learning paradigm, the network is equivalent to two networks sharing weights and regularizing each other.

The loss assigned to this network is the sum of an ‘iconic’ loss computed on the segmentation branch, and a ‘shape regularization’ loss computed on the regression part. The iconic segmentation loss is the sum of a Dice and Cross entropy loss as used in nnU-Net. The shape regularization loss is our proposed Shape Feature Loss.

4 Shape Feature loss for 3D anatomical structure segmentation

Our proposed loss function acts on the regression branch and is an L1 function between features extracted from the predicted $\hat{\mathbf{y}}$ and the ground-truth $\mathbf{y}$ signed distance map

$$L_{sf}(\hat{\mathbf{y}},\mathbf{y}) = \|\phi(\hat{\mathbf{y}}) - \phi(\mathbf{y})\|_1 \tag{2}$$

Our feature extraction function ϕ is the concatenation of 3D pose feature function ϕ_{pl} and a shape deformation feature function ϕ_{sd} as shown in Figure 1 and noted in $\phi = \left[\phi_{pl}, \phi_{sd}\right]$.

4.1 3D pose feature extraction

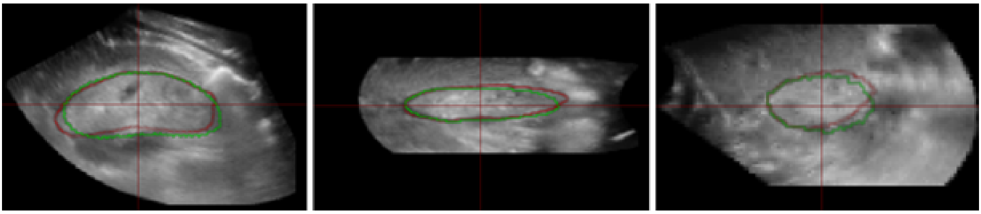

Figure 3: Besides giving good segmentation results (prediction in red, ground-truth in green) our method produces aligned images with a reference pose, which provides a standardized view for further analysis.

We use ou pre-trained STN to align and extract rigid parameters from an SDM. These parameters represent 3D rotations $\theta_x, \theta_y, \theta_z$, 3D translations T_x, T_y, T_z and Scale S and will be compared to their ground-truth counterparts.

$$\phi_{pl}(\mathbf{y}) = F_{loc}(\mathbf{y}) = [T_x, T_y, T_z, \theta_x, \theta_y, \theta_z, S] \tag{3}$$

4.2 Shape deformation feature extraction

Our shape deformation feature function $\phi_{sd}(\mathbf{y})$ acts on normalized pose SDM $\mathbf{y_N} = STN(\mathbf{y}$ produced by the STN and extracts the shape vector as follows:

$$\phi_{sd}(\mathbf{y}) = W^T(STN(\mathbf{y}) - \bar{\mathbf{y_N}}) \tag{4}$$

This is derived from Equation 1, where $\bar{\mathbf{y}}_{\mathbf{N}}$ is the mean normalized pose SDM and W^T is the transpose of the eigenvectors matrix.

5 Dynamic loss-term calibration

In order to fully take advantage of our loss function, we calibrate the multiple terms of our loss. Our loss function can be written as, with $\phi = [\phi_i]$:

$$\hat{L} = \sum_{k=1}^{k=n} \hat{L}_i, \text{ where } \hat{L}_i = w_i L_i + c_i \text{ and } L_i = \|\phi_i(\hat{\mathbf{y}}) - \phi_i(\mathbf{y})\|_1 \tag{5}$$

The number of loss-terms n is equal to 26, since we have 7 pose parameters and 19 modes of variations. Inspired by multi-task learning, we use the self paced learning technique based on homoscedastic uncertainty introduced in [8]. The basic idea is to monitor the learning progress signal and learn a policy to adjust the relative weights to the loss terms. Hence we learn dynamic weights (w_i, c_i) for each loss component. Specifically, we learn the network homoscedastic uncertainty and use it to weigh different tasks. This uncertainty does not change with input data and is loss-term-specific. In practice, this results in changing each loss term with the following loss:

$$\hat{L}_i = \frac{1}{\sigma_i^2} L_i + \log(\sigma_i^2) \tag{6}$$

where σ_i is the learnable homoscedastic uncertainty associated with each loss component. The second term avoids the uncertainties to be set too high. We refer to [8] for derivation of this formulation of the multi-task loss which is based on the maximization of the Gaussian likelihood with homoscedastic uncertainty.

Our final loss is sum of an iconic segmentation loss computed on the segmentation branch of the network and

6 Experimental Validation

6.1 Dataset and experimental setting

We use a dataset of left and right adult kidney 3D ultrasounds from 667 patients. These images were collected from several hospital sites and include various kidney conditions such as normals, CKD and transplants. Reference segmentations were performed by manual annotation under the supervision of medical experts. These exams were performed in the course of the normal care pathway of the patient and were studied retrospectively after anonymization. The 3D images were preprocessed to have the same voxel size and cropped afterward to have same image size of $192 \times 160 \times 80$ containing the kidney structure. We used 80% of the data for training, 20% for validation by random selection. We let the nnU-Net pre-processing routines determine all hyper-parameters. We use the same setting when adding our Shape Feature loss function L_{sf} to the network.

6.2 Results

We compare the state-of-the-art nnU-Net implementation with our approach. For our experiments, we retain exactly the same nnU-Net implementation and its hyperparameters and we add to the network our L_{sf} loss branch. Since our STN applies rotation and translation to SDMs, we add zero padding to our input images, so that the aligned SDMs remain valid after STN transformations. We use Dice similarity coefficient as a validation metric. Quantitatively, Our method outperforms the nnU-Net as shown in Table 1. We obtain a p value < 0.002. Qualitatively, we can see in Figure 4 that our results are more regularized and anatomically plausible. Furthermore, Figure 5 shows a case where our method detects the 3D kidney pose better than the nnU-Net method. We believe that our method is clinically significant to practitioners as it provides anatomically more accurate segmentations, requiring less time for manual correction and standardizes pose as shown in Figure 3 enabling less subjective visual medical assessment.

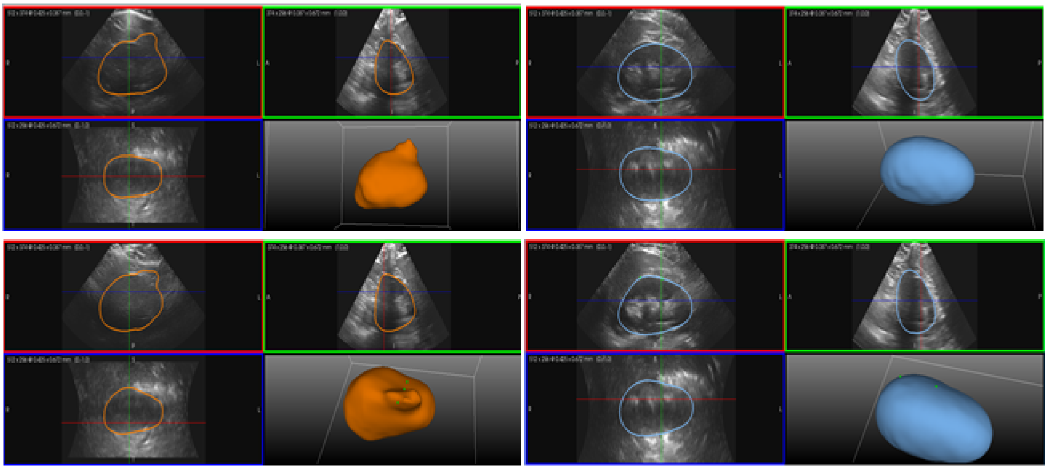

Figure 4: In Orange, two results (top and bottom) of segmentation with nnU-Net approach. The results show artifacts that are not anatomically plausible. In Blue, results of the segmentation using our method. The shape constraints embeded in our Shape Feature loss enable to produce regularized segmentations

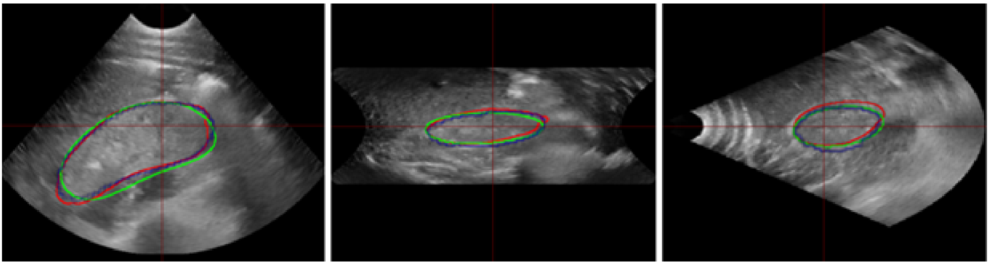

Figure 5: A case where our method better detects the 3D pose (more visible on the third image) of the kidney than the nnU-Net approach. Our results are shown in blue, nnU-Net in red and ground-truth in green.

6.3 Ablation study

In order to show the importance of each component of our method, we proceed to an ablation study. As shown in Table 2 we first discarded the dynamic loss-term calibration. We found

Table 1: Mean and standard deviation of Dice score, total training time and memory allocated comparison between our method and the nnU-Net on our dataset

Architecture	Training Time	Memory allocated	Dice score kidney
nnU-Net	124h	8.15Gb	90.64 (1.61)
Ours	167h	16.35Gb	**92.07 (1.91)**

that the decrease in performance is marginal but the network converged more slowly than with the dynamic calibration term. We also discarded the FSL term and replaced it with an L1 directly on SDMs -meaning there is non feature extraction form the predicted SDMs. The decrease is performance is significant but the performance is still better that the nnU-Net. We also replaced L1 by L2, we found similar results. We can't discard the STN part as projecting unaligned SDMs on PCA space would give incorrect shape parameters. Hence we built an aligned version of our dataset by registering all images to a reference image. In Table 3 we show an ablation study on the aligned kidney dataset. In this setting we don't use an STN and our method becomes similar to [7]. This is a simpler task. We can see that each component of our method brings improvement in performance.

Table 2: Ablation study on the kidney dataset

Method	Dice score kidney
nnU-Net	90.64 (1.61)
nnU-Net+L1 SDM regression	91.05 (1.21)
nnU-Net+L2 SDM regression	91.02 (1.34)
nnU-Net+FSL	92.01 (2.01)
nnU-Net+FSL+loss calibration	92.07 (1.91)

6.4 Computing infrastructure

In all our experiments, we use an Nvidia Titan RTX graphics card which has 24 GB of memory, an Intel 32-core Xenon processor, 64 Gb of Ram workstation. This machine runs on Ubuntu 18.04 LTS. Our code is built on top on the publicly available Pytorch (version 1.6) original implementation of nnU-net.

Table 3: Ablation study on the aligned version of the kidney dataset

Method	Dice score kidney
nnU-Net	94.44 (1.21)
nnU-Net+L1 SDM regression	95.24 (1.31)
nnU-Net+SSM loss	96.34 (1.61)
nnU-Net+SSM loss+loss calibration	96.74 (1.41)

7 Conclusion

In this paper, we introduced the Shape Feature Loss. Plugged into the U-Net architecture, this simple loss function is suited for 3D anatomical structure segmentation. It incorporates prior knowledge to regularize output needed for challenging tasks such as 3D ultrasound kidney segmentation. Thanks to the used STN and a PCA model, the rigid and non rigid parameters are decoupled and explicitly encoded from a U-Net prediction. Applied to the currently considered kidney conditions, this proposed approach improves the segmentation results compared to the state-of-the-art nnU-Net baseline. However, in the case of highly pathological cases, for example large infiltrating tumors, a manual correction of the output will be necessary to adapt to non-learned shape irregularities.

Indeed, the use of a PCA model in an SDM space might require a relatively big training segmentation dataset to reveal meaningful deformations. Moreover PCA is limited to model a linear subspace which is not well suited to all anatomical deformations. Hence, as a future work, we will consider designing a shape deformation loss function acting on 3D mesh representation using Graph Convolutional Neural Networks [30]. This would allow to encode local shape deformations. As a further perspective, we could take advantage of the pre-trained STN layers and use them to formulate an additional perceptual loss.

References

[1] Bakker, J. and Olree, M. and Kaatee, R. and Lange, E. and Moons, K. and Beutler, J. and Beek, F., Renal Volume Measurements: Accuracy and Repeatability of US Compared with That of MR Imaging1, In: Radiology (1999)

[2] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., Maier-Hein, K. H. . nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. In: Nature Methods (2021)

[3] O. Ronneberger, P. Fischer, and T. Brox,: “U-Net: Convolu-tional networks for biomedical image segmentation,” in: MICCAI (2015)

[4] Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., … Rueckert, D. : Anatomically Constrained Neural Networks (ACNN): Application to Cardiac Image Enhancement and Segmentation. In: IEEE Transactions on Medical Imaging.(2017)

[5] Luc, P., Couprie, C., Chintala, S., Verbeek, J. : Semantic Segmentation using Adversarial Networks. In: NIPS Workshop on Adversarial Training. (2016)

[6] Johnson, J., Alahi, A., Fei-Fei, L. (n.d.). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: ECCV (2016)

[7] Al Arif, S. M. M. R., Knapp, K., Slabaugh, G. : SPNet: Shape prediction using a fully convolutional neural network in MICCAI (2018)

[8] Cipolla, R., Gal, Y., Kendall, A. : Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In: CVPR (2018)

[9] Ma, J., Wei, Z., Zhang, Y., Wang, Y., Lv, R., Zhu, C., … Chen, J. (2020).: How Distance Transform Maps Boost Segmentation CNNs: An Empirical Study CNNs with Distance Transform Maps. In: Proceedings of Machine Learning Research (2020)

[10] Fernando Navarro, Suprosanna Shit, Ivan Ezhov, Johannes Paetzold, Andrei Gafita, Jan C. Peeken, Stephanie E. Combs, and Bjoern H. Menze.: Shape-aware complementary-task learning for multi-organ segmentation. In: Machine Learning in Medical Imaging (2019)

[11] M. Jun : Segmentation Loss Odyssey. In: arXiv preprint arXiv:(2005)

[12] T. F. Cootes and C. J. Taylor. : Combining point distribution models with shape models based on finite element analysis. In: Image and Vision Computing (1995)

[13] W. Bai, W. Shi, D. P. O'Regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert. : A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images. In: IEEE TMI (2013)

[14] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. E. Grimson, and A. Willsky,: A shape-based approach to the segmentation of medical imagery using level sets, In: IEEE Transactions on Medical Imaging (2003)

[15] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, : Active shape models-their training and application In: Computer Vision and Image Understanding (1995)

[16] Boussaid, H., Kokkinos, I. : Fast and exact: ADMM-based discriminative shape segmentation with loopy part models. CVPR (2014)

[17] He, K., Zhang, Z., Ren, S. Sun, J. : Deep residual learning for image recognition. In : CVPR (2016)

[18] Huang, G., Liu, Z., van der Maaten, L. Weinberger, K. Q.: Densely connected convolutional networks. In :CVPR (2017).

[19] Oktay, O. et al. : Attention U-net: learning where to look for the pancreas. Arxiv Preprint Arxiv (2018).

[20] Hu, J., Shen, L. Sun, G. : Squeeze-and-excitation networks. In : CVPR (2018)

[21] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. a Yuille, A.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs In: IEEE Trans. Pattern Anal. Mach. Intell. (2017).

[22] Wu, Z., Shen, C., Hengel, A.v.d.: Bridging category-level and instance-level semantic image segmentation. arXiv preprint arXiv (2016)

[23] Lin, T., Goyal, P., Girshick, R., He, K., Dollr, P.: Focal loss for dense object detection. In: ICCV (2017)

[24] Caliva, F., Iriondo, C., Martinez, A.M., Majumdar, S., Pedoia, V.: Distance map loss penalty term for semantic segmentation. In: MIDL (2019)

[25] Brosch, T., Yoo, Y., Tang, L.Y.W., Li, D.K.B., Traboulsee, A., Tam, R.: Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In: MICCAI (2015)

[26] Rahman, M.A., Wang, Y.: Optimizing intersection-over-union in deep neural networks for image segmentation. In: International symposium on visual computing. (2016)

[27] Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging. (2019)

[28] Su, Y., Jihoon, K., Young-Hak, K.: Major vessel segmentation on x-ray coronary angiography using deep networks with a novel penalty loss function. In: MIDL (2019)

[29] Karimi, D., Salcudean, S.E.: Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. arXiv preprint arXiv (2019)

[30] Liang, J., Homayounfar, N., Ma, W.-C., Xiong, Y., Hu, R., Urtasun, R. : PolyTransform: Deep Polygon Transformer for Instance Segmentation. Proceedings of the In: CVPR (2020)

[31] Danielsson, Per-Erik. : Euclidean Distance Mapping. In: Computer Graphics and Image Processing (1980).

[32] Tilborghs S., Dresselaers T., Claus P., Bogaert J., Maes F. Shape Constrained CNN for Cardiac MR Segmentation with Simultaneous Prediction of Shape and Pose Parameters. In: STACOM (2020)