Investigating CoordConv for Fully and Weakly Supervised Medical Image Segmentation

Rosana El Jurdi
LITIS Lab
Université de Rouen Normandie
Rouen, France
rosana.el-jurdi@univ-rouen.fr
Lebanese University
Beirut, Lebanon

Thomas Dargent
LITIS Lab
Université de Rouen Normandie
Rouen, France
thomas.dargent@etu.univ-rouen.fr

Caroline Petitjean
LITIS Lab
Université de Rouen Normandie
Rouen, France
caroline.petitjean@univ-rouen.fr

Paul Honeine
LITIS Lab
Université de Rouen Normandie
Rouen, France
paul.honeine@univ-rouen.fr

Fahed Abdallah
Lebanese University
Beirut, Lebanon
fahed.abdallah76@gmail.com

***Abstract*—Convolutional neural networks (CNN) have established state-of-the-art performance in computer vision tasks such as object detection and segmentation. One of the major remaining challenges concerns their ability to capture consistent spatial attributes, especially in medical image segmentation. A way to address this issue is through integrating localization prior into system architecture. The CoordConv layers are extensions of convolutional neural network wherein convolution is conditioned on spatial coordinates. This paper investigates CoordConv as a proficient substitute to convolutional layers for organ segmentation in both fully and weakly supervised settings. Experiments are conducted on two public datasets, SegTHOR, which focuses on the segmentation of thoracic organs at risk in computed tomography (CT) images, and ACDC, which addresses ventricular endocardium segmentation of the heart in MR images. We show that if CoordConv does not significantly increase the accuracy with respect to standard convolution, it may interestingly increase model convergence at almost no additional computational cost.**

***Index Terms*—Image segmentation, Fully Convolutional Networks, CoordConv, Location Prior, Weakly Supervised Learning, MRI, CT**

I. Introduction

Convolutional neural networks (CNN), a class of deep learning models, have long emerged as powerful tools with outstanding performance given a variety of applications such as object detection and semantic segmentation. Despite their breakthrough, CNN performance is still prone to degradation due to the lack of spatial features that would be especially helpful for image segmentation, where pixelwise decision must be taken [1]. Recent advances in the domain have focused on integrating location prior onto CNN training in order to overcome this inadequacy.

To guide medical segmentation, priori information such as shape and the topology of organs have often been investigated as means of maintaining anatomical plausibility. In this context, plenty of work has been developed prior to the advent of deep learning for segmentation given variational approaches. Multi-atlas based approaches and machine learning based techniques were also ways to integrate prior information through relying on labeled data [2]. However, there is no straightforward way to transfer these previous works to deep learning networks, since the latter have some specific constraints including differentiability and optimization of the loss function. Moreover, there is still limited research on which information to model, how to model it, and how to integrate it into deep neural networks, and more specifically into the CNN. Since the priors compensate for the need of a massive training set and are not based explicitly on the ground truth labels, weakly supervised image segmentation, which makes use of only coarse-grained annotations, can greatly benefit from this type of constrained loss [3], [4].

The CoordConv layers, recently introduced in [5], are extensions of convolutions that allow convolution filter to take into account the spatial coordinates of the pixels. The goal of CoordConv is to learn a mapping between coordinates in the Cartesian space and coordinates in the one-hot pixel space. CoordConv has shown its potential for object localization [5], [6], and has rightfully raised interest for image segmentation [7], [8]; however, the CoordConv solution's added value has not been yet assessed in image segmentation.

This paper investigates CoordConv as a proficient substitute to convolutional layers in FCN segmentation models dedicated to organ contouring in anatomical images. We explore the effect of CoordConv on model performance and rate of convergence when integrated into different layers of the network, replacing standard convolution in both convolutional and deconvolutional layers.

The authors would like to acknowledge the National Council for Scientific Research of Lebanon (CNRS-L) and the *Agence Française de la Francophonie* (AUF) for granting a doctoral fellowship to Rosana El Jurdi, as well as the ANR (Project APi, grant ANR-18-CE23-0014).

978-1-7281-8750-1/20/$31.00 ©2020 IEEE

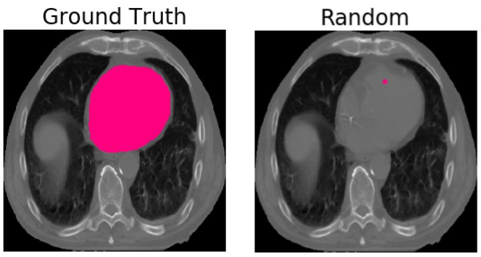

Fig. 1. (left) Ground truth of heart segmentation overlaid on a CT image from SegTHOR, and (right) an example of weak label: randomly located circle of diameter 4 pixels.

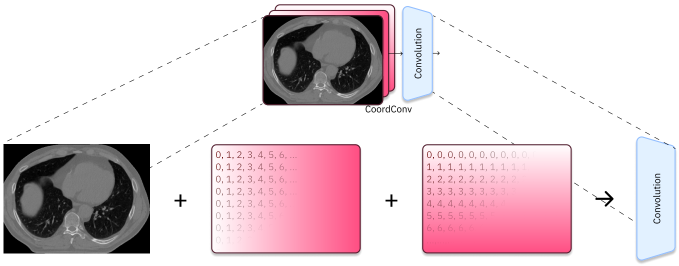

Fig. 2. CoordConv main principle: x-layer and y-layer are concatenated to the input image (here, a CT image from the SegTHOR dataset).

II. Related works

In the literature, there are several ways to inject spatial, geometrical or anatomical high-level information into segmentation networks. One way is at the level of the loss function. This consists in designing specific prior-based loss functions to enforce high-level label dependencies. Prior losses are usually designed based on features extracted from ground truth during training or estimated priori. Such prior losses may integrate adjacency relations between organs [9], organ volume size [3] or Betti values expressing the number of connected components [10], [11] to name a few. They can also stem from transformations [12] or (often non-linear) representations of the ground truth, such as distance map [13], VAE encoding [14], or parametric functions [15]. Despite the versatility of such loss based methods, however, designing novel loss functions often face considerable challenges including differentiability of the losses and their optimization strategies.

Another way to inject anatomical prior is via network architecture and layer design [16]–[21]. For example, in [17] collaborative architectures are implemented to iteratively refine the posterior probability given the previous probabilities of a large number of context locations, thus providing information about neighboring organs. In [20], authors integrated location and shape prior onto the learning process through introducing Bounding filters at the level of the skip-connections in a U-Net base model. In [21] the spatial location of patches extracted from the image into a CNN model structure is injected posterior to the convolutional layers. Another example is in [16] where the authors modified the decoder layers of a U-Net-like structure in order to incorporate prior via super resolution ground-truth maps.

Different from works that enforce some constraints through the loss, thereby requiring loss differentiability or costly optimization tool, the CoordConv layer idea of Liu et al. in [5] consists in modifying the input to the network. The goal is to establish mappings between the Cartesian space and the pixel space, by enabling the filters to know where pixels are located. The implementation of CoordConv is done by concatenating two additional X and Y channels to the input channel as shown in Fig. 2. In doing so, CoordConv ensures the best of both convolutional and spatial features.

III. Methodology: Segmentation Model U-Net and CoordConv

Our model builds upon the segmentation model proposed by [22], which is a typical residual U-Net, and the CoordConv layers. Similar to U-Net, the network is composed of encoder/decoder layers of convolutional and deconvolutional blocks with skip connections. However, instead of consecutively feeding the output of each convolution within the layer to the convolution that proceeds it, the outputs from different convolutions per layer are rather combined and convolved for more fine-grained feature extraction. For fully supervised experiments, cross entropy was adopted as loss functions. In the following, we describe in more details the weakly supervised setting.

Weakly supervised setting: For weakly supervised segmentation, the labels are not the full ground truth, but are constituted by random seeds (see an example in Fig. 1). In this case, the loss is enriched with a size constraint as described in [3].

Let us denote by S the softmax output (or probability map) of the network. One can write the loss function as a combination of two terms:

$$L = \mathcal{H}(S) + \lambda \mathcal{C}(V_s), \tag{1}$$

where $\mathcal{H}(S)$ is the cross-entropy between prediction and true pixel labels and $\mathcal{C}(V_s)$ is a size-constraint regularization as proposed by [3], where $V_s = \sum_p S_p$ can be interpreted as the area of the segmented region. The size constraint $\mathcal{C}(V_s)$ in the loss function (1) consists in enforcing the size of the segmented area to be in a specific (empirically defined) range, denoted by $[a, b]$:

$$\mathcal{C}(V_s) = \begin{cases} (V_s - a)^2, & \text{if } V_s < a \\ (V_s - b)^2, & \text{if } V_s > b \\ 0 & \text{else.} \end{cases} \tag{2}$$

Following [3], the value of the parameter λ was set to 0.01 in our experiments. We refer the reader to [3] for more details.

Regarding CoordConv, we implement it by adding two extra X and Y coordinate channels to the input channel, as shown in Fig. 2.

TABLE I
AVERAGE DICE INDEX (± STANDARD DEVIATION) AND AVERAGE HAUSDORF DISTANCE FOR FULLY AND WEAKLY SUPERVISED SEGMENTATION ON SEGTHOR. COORDCONV-1ST (RESP. -ENC AND -ALL) MEANS THE FIRST (RESP. THE ENCODER AND ALL) CONVOLUTIONAL LAYERS OF THE NETWORK HAVE BEEN REPLACED BY COORDCONV.

	SegTHOR Dataset			
	Fully superv.		Weakly superv.	
	Dice	Hausdorff	Dice	Hausdorff
U-Net (no CoordConv)	0.86 ±0.24	3.50 ±1.19	0.82 ± 0.20	4.91 ±1.49
U-Net+CoordConv-1st	0.89 ±0.18	3.44 ±0.87	0.82 ± 0.21	4.28 ±1.35
U-Net+CoordConv-EnC	0.89 ±0.18	3.36 ±0.87	0.83 ± 0.20	4.79 ±1.07
U-Net+CoordConv-All	0.89 ± 0.17	3.54 ±0.86	0.83 ±0.20	4.59 ±1.25

TABLE II
AVERAGE DICE INDEX (± STANDARD DEVIATION) AND AVERAGE HAUSDORF DISTANCE FOR FULLY AND WEAKLY SUPERVISED SEGMENTATION ON ACDC. COORDCONV-1ST (RESP. -ENC AND -ALL) MEANS THE FIRST (RESP. THE ENCODER AND ALL) CONVOLUTIONAL LAYERS OF THE NETWORK HAVE BEEN REPLACED BY COORDCONV.

	ACDC Dataset			
	Fully superv.		Weakly superv.	
	Dice	Hausdorff	Dice	Hausdorff
U-Net (no CoordConv)	0.85 ± 0.27	2.14 ± 0.96	0.73 ± 0.26	3.51 ±3.00
U-Net+CoordConv-1st	0.85 ± 0.26	2.28 ±0.92	-	-
U-Net+CoordConv-EnC	0.86 ± 0.22	2.26 ± 1.11	-	-
U-Net+CoordConv-All	0.88 ± 0.22	2.06 ± 0.83	0.72 ± 0.24	3.16 ±1.63

IV. EXPERIMENTS

A. Datasets and protocol

Experiments are conducted on two public datasets: the SegTHOR dataset[1], where the goal is to segment thoracic organs at risk in computed tomography (CT) images [23], and the ACDC dataset[2], which addresses segmentation of the cardiac ventricular endocardium in MR (magnetic resonance) images [24].

Regarding SegTHOR, we will only focus on the segmentation of the heart. The proposed model is trained on 219 slices (6 patients) from SegTHOR which were augmented to 1096 through rotation, random mirroring and flipping. Validation is conducted on a set of 155 slices (4 patients). With regards to the ACDC dataset, training and validation are performed on 1675 and 229 image slices corresponding to 75 and 25 patients respectively. Images from both datasets were resized to 256×256 pixels and normalized to value between 0 and 1.

Models were evaluated using the Dice index and Hausdorff distance. For the fully residual U-Net, we use the implementation from [25] and modify it accordingly with CoordConv. Thus, we have conducted training using the Adam optimizer with a batch size of 4 over 200 epochs. Adopting the same framework as [3], the learning rate was set to 5×10^{-4} and halved each 20 epochs if the validation performance did not improve.

Regarding the weakly supervised segmentation, the weak labels are random seeds, which are circles of diameter 4 pixels and less, generated from the manual ground truth. Bounds for training under the size constraint loss were also extracted from the ground-truth segmentation maps and were set to (a = 97.9, b = 1722) and (a = 210, b = 7800) for ACDC and SegTHOR respectively.

B. Results and analysis

In our experiments, we replace the standard convolution by the CoordConv component with respect to both the encoding and the decoding path of the network. We investigate three different settings in order to assess the CoordConv-based networks: in Experiment 1, we replace only the input, which is a regular mono-channel image, with a concatenation of the image and the X and Y coordinate channels. We call this model **CoordConv-1st**. Experiment 2, which is denoted by **CoordConv-EnC**, involves replacing all convolutional layers within the network with CoordConv layers. Experiment 3 consists in concatenating Cartesian coordinates at the level of both convolutional and deconvolutional layers and is labeled by **CoordConv-All**.

Interestingly, results from TABLE I for SegTHOR and TABLE II for ACDC show that adding CoordConv does not significantly improve the segmentation accuracy with respect to the baseline ("U-Net (no CoordConv)" row in the tables), as confirmed by the p-value (>0.05) from a paired Student t-test between each model's (row denoted by "U-Net+CoordConv-XXX") and the baseline's Dice and Hausdorff values. However, what is remarkable is rather the rate and trend of convergence of CoordConv models relative to the baseline model. Observing the corresponding validation curves in Fig. 3

[1] https://competitions.codalab.org/competitions/21145

[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/

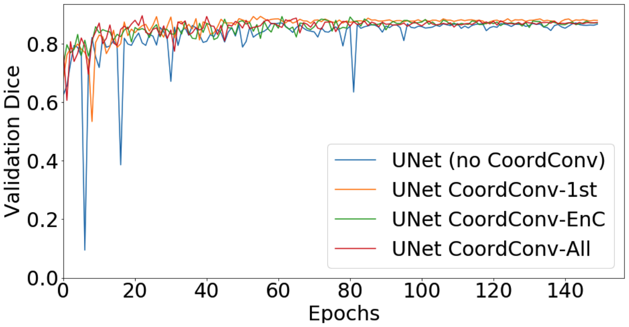

Fig. 3. Evolution of the Dice loss in validation, for the SegTHOR dataset, in fully supervised setting, using different configuration, from no CoordConv layer to all convolutional layers replaced by CoordConv.

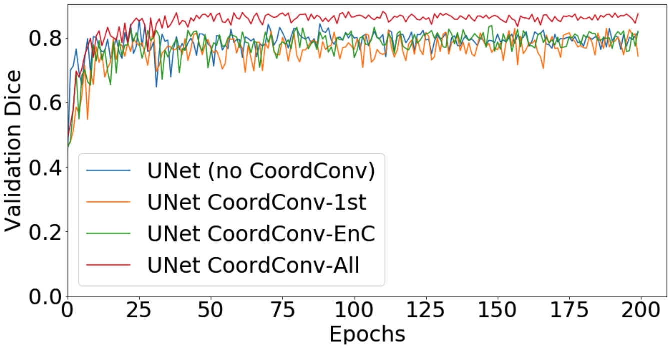

Fig. 4. Evolution of the Dice loss in validation, for the ACDC dataset, in fully supervised setting, using different configuration, from no CoordConv layer to all convolutional layers replaced by CoordConv.

and Fig. 4, we gather that CoordConv model helps regularizing the network training (for the SegTHOR dataset) and learns faster (for the ACDC dataset). CoordConv layers allow more stable as well as faster convergence evading performance dropout realized by the "no CoordConv" model.

Computational complexity: Regarding the computational overhead induced by using CoordConv, it can be quantified as follows. Let us denote by c_{in} the number of input channels and by c_{out} the number of output channels of a standard convolutional layer, and by k the square kernel size. A convolutional layer has $c_{in}c_{out}k^2$ weights. If we denote by d the number of dimensions taken into account into the CoordConv layer (in our case $d = 2$), a convolutional CoordConv layer has $(c_{in} + d)c_{out}k^2$ weights, the additional computational cost is thus reduced. Note that we did not count the bias weights since their number are unchanged by CoordConv.

V. Conclusion and Perspectives

In this paper, we investigated the role of CoordConv on model performance and convergence for organ segmentation in anatomical image. Our results show that CoordConv layers may have an effect on model convergence at almost no additional computational cost, consistently as in [5]. Future investigations include exploring the network weights with sensitivity maps, so as to gain some insight into what is learnt at the coordinates level and adding other layers in CoordConv, such as the "r" layer based on polar coordinates as advocated in [5].

References

[1] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep Semantic Segmentation of Natural and Medical Images: A Review," *arXiv:1910.07655 [cs, eess]*, Oct. 2019, arXiv: 1910.07655. [Online]. Available: http://arxiv.org/abs/1910.07655

[2] D. Grosgeorge, C. Petitjean, and S. Ruan, "Multilabel statistical shape prior for image segmentation," *IET Image Processing*, vol. 10, no. 10, pp. 710–716, 2016.

[3] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed, "Constrained-CNN losses for weakly supervised segmentation," *Medical Image Analysis*, vol. 54, pp. 88–99, May 2019, arXiv: 1805.04628. [Online]. Available: http://arxiv.org/abs/1805.04628

[4] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah, "Organ Segmentation in CT Images With Weak Annotations: A Preliminary Study," in *27th GRETSI Symposium on Signal and Image Processing*, Lille, France, Aug. 2019.

[5] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution," *arXiv:1807.03247 [cs, stat]*, Dec. 2018, arXiv: 1807.03247. [Online]. Available: http://arxiv.org/abs/1807.03247

[6] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end Training of Deep Visuomotor Policies," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, Jan. 2016. [Online]. Available: http://dl.acm.org/citation.cfm?id=2946645.2946684

[7] H. Qi, S. Collins, and J. A. Noble, "UPI-Net: Semantic Contour Detection in Placental Ultrasound," in *Visual Recognition for Medical Images (VRMI), ICCV 2019 workshop*, Sep. 2019, arXiv: 1909.00229.

[8] X. Yao, H. Yang, Y. Wu, P. Wu, B. Wang, X. Zhou, and S. Wang, "Land Use Classification of the Deep Convolutional Neural Network Method Reducing the Loss of Spatial Features," *Sensors*, vol. 19, no. 12, p. 2792, Jan. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/12/2792

[9] P.-A. Ganaye, M. Sdika, B. Triggs, and H. Benoit-Cattin, "Removing segmentation inconsistencies with semi-supervised non-adjacency constraint," *Medical image analysis*, vol. 58, p. 101551, 2019.

[10] X. Hu, L. Fuxin, D. Samaras, and C. Chen, "Topology-Preserving Deep Image Segmentation," *NIPS 2019*, Jun. 2019, arXiv: 1906.05404. [Online]. Available: http://arxiv.org/abs/1906.05404

[11] J. R. Clough, I. Oksuz, N. Byrne, J. A. Schnabel, and A. P. King, "Explicit topological priors for deep-learning based image segmentation using persistent homology," *arXiv:1901.10244 [cs]*, Jan. 2019, arXiv: 1901.10244. [Online]. Available: http://arxiv.org/abs/1901.10244

[12] X. Chen, B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams, and Y. Zheng, "Learning Active Contour Models for Medical Image Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 11 624–11 632. [Online]. Available: https://ieeexplore.ieee.org/document/8953484/

[13] F. Caliva, C. Iriondo, A. M. Martinez, S. Majumdar, and V. Pedoia, "Distance map loss penalty term for semantic segmentation," in *International Conference on Medical Imaging with Deep Learning*, London, UK, 2019.

[14] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker, and D. Rueckert, "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 384–395, Feb. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8051114/

[15] Z. Yan, X. Yang, and K.-T. Cheng, "A deep model with shape-preserving loss for gland instance segmentation," in *MICCAI*, 2018.

[16] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, T. Dawes, D. P. O'Regan, B. Kainz,

B. Glocker, and D. Rueckert, “Anatomically constrained neural networks : Application to cardiac image enhancement and segmentation,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 384–395, Feb 2018.

[17] R. Trullo, C. Petitjean, S. Ruan, B. Dubray, D. Nie, and D. Shen, “Joint segmentation of multiple thoracic organs in CT images with two collaborative deep architectures,” *MICCAI’17 workshop Deep Learning in Medical Image Analysis*, 2017.

[18] H. Oda, H. R. Roth, K. Chiba, J. Sokolić, T. Kitasaka, M. Oda, A. Hinoki, H. Uchida, J. A. Schnabel, and K. Mori, “Besnet: Boundary-enhanced segmentation of cells in histopathological images,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 228–236.

[19] C. Zotti, Z. Luo, O. Humbert, A. Lalande, and P. Jodoin, “Gridnet with automatic shape prior registration for automatic MRI cardiac segmentation,” in *Statistical Atlases and Computational Models of the Heart STACOM, Held in Conjunction with MICCAI, Quebec City, Canada*, ser. LNCS, vol. 10663, 2017, pp. 73–81.

[20] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah, “Bb-unet: U-net with bounding box prior,” *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2020.

[21] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. Van Uden, C. Sanchez, G. Litjens, F.-E. Leeuw, B. Ginneken, E. Marchiori, and B. Platel, “Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities,” *Scientific Reports*, vol. 7, 10 2016.

[22] T. M. Quan, D. G. C. Hildebrand, and W. Jeong, “Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics,” *CoRR*, vol. abs/1612.05360, 2016.

[23] Z. Lambert, C. Petitjean, B. Dubray, and S. Ruan, “Segthor: Segmentation of thoracic organs at risk in CT images,” *CoRR*, vol. abs/1912.05950, 2019.

[24] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P. Jodoin, “Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?” *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, Nov 2018.

[25] H. Kervadec, “Constrained-CNN losses for weakly supervised segmentation,” https://github.com/LIVIAETS/SizeLoss$_WSS$, 2019.