Data augmentation for multi-organ detection in medical images

Maryam Hammami, Denis Friboulet, Razmig Kéchichian

Data augmentation for multi-organ detection in medical images

Maryam Hammami
Univ. Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm CREATIS UMR 5220,U1206, F-69266 Lyon, France
maryam.hammami@creatis.insa-lyon.fr

Denis Friboulet
Univ. Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm CREATIS UMR 5220,U1206, F-69266 Lyon, France
denis.friboulet@creatis.insa-lyon.fr

Razmig Kechichian
Univ. Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm CREATIS UMR 5220,U1206, F-69266 Lyon, France
razmig.kechichian@creatis.insa-lyon.fr

***Abstract*—We propose a deep learning solution to the problem of object detection in 3D medical images, i.e. the localization and classification of multiple structures. Supervised learning methods require large annotated datasets that are usually difficult to acquire. We thus develop a Cycle Generative Adversarial Network (CycleGAN) and You Only Look Once (YOLO) combined method for data augmentation from one modality to another via CycleGAN and organ detection from generated images via YOLO. This results in a fast and accurate detection with a mean average distance of 7.95 mm for CT modality and 16.18 mm for MRI modality, which is significantly better than detection without data augmentation. We show that the approach compares favorably to state-of-the-art detection methods for medical images on CT data.**

***Index Terms*—multi-organ detection, image synthesis, data augmentation, medical imaging**

I. Introduction

Object detection consists in localizing structures in images using bounding boxes and classifying them. Many approaches have been proposed relying on statistical [1]–[3] or deep learning [4]–[6] techniques. In spite of the success of existing methods, only a few works have employed deep learning for multi-organ detection in medical images [7], [8]. Object detection is a prerequisite in many radiological procedures such as patient screening and diagnosis which implies localizing anatomical structures or lesions. Most of object detection models in medical images are designed for single-object detection [9], [10].

Our method aims at detecting multiple organs in 3D medical images. Detection should furthermore be time efficient when processing large datasets. We choose the YOLO detector [5] as a basis in our work. It has been shown to offer a good precision vs speed trade-off for natural images compared to other deep detectors. Supervised deep learning requires large training datasets which are not always available for medical images. This is because such images are expensive to obtain, compared to natural images. Furthermore, manual annotation of medical images, especially in 3D, is very time consuming. We propose to expand a baseline training dataset via data augmentation. Most data augmentation approaches apply transformations to images such as rotations and translations [11]. Transformed data are then added to the training set. As opposed to this traditional technique, we use a CycleGAN [12] as an unsupervised method that synthesizes images from annotated source images of a different modality. We show that the CycleGAN+YOLO combination yields an efficient approach to augment and detect multiple structures.

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

II. Related works

A. Object detection

Supervised object detection aims at classifying object instances from predefined annotations and localizing them in images. Deep learning-based detection methods can be categorized into two approaches: *two-stage* and *one-stage*.

Models in the former approach are trained separately for two tasks: detection of regions of interest and classification/localization of objects. Region-Based Convolutional Neural Network (R-CNN) methods [4], [13], [14] are among the best-preforming ones. They use modules for feature extraction, classification/regression and region proposal, the latter being a separate convolutional network in [4].

In the one-stage detection approach, a single network is trained simultaneously for classification and localization, no region proposals are created. You Only Look Once (YOLO) [5] and Single Shot multi-box Detector (SSD) [6] are both popular detection methods in this category. YOLO is a fast real-time object detector with an optimized network architecture. SSD introduces multi-reference and multi-resolution techniques for added precision.

In medical image analysis, best-performing traditional detection approaches are based on regression forests [2] which are applied in a cascaded, global-to-local fashion in [1], [3] augmented by a shape prior in the latter work for improved precision. Deep methods however are quickly gaining ground. YOLO has been used for detection on retinal images [15], and SSD for liver lesion detection in CT [9]. Among recent works investigating deep multi-organ detection we mention [7] where two convolutional networks, one for classification and another for bounding-box regression, were trained and

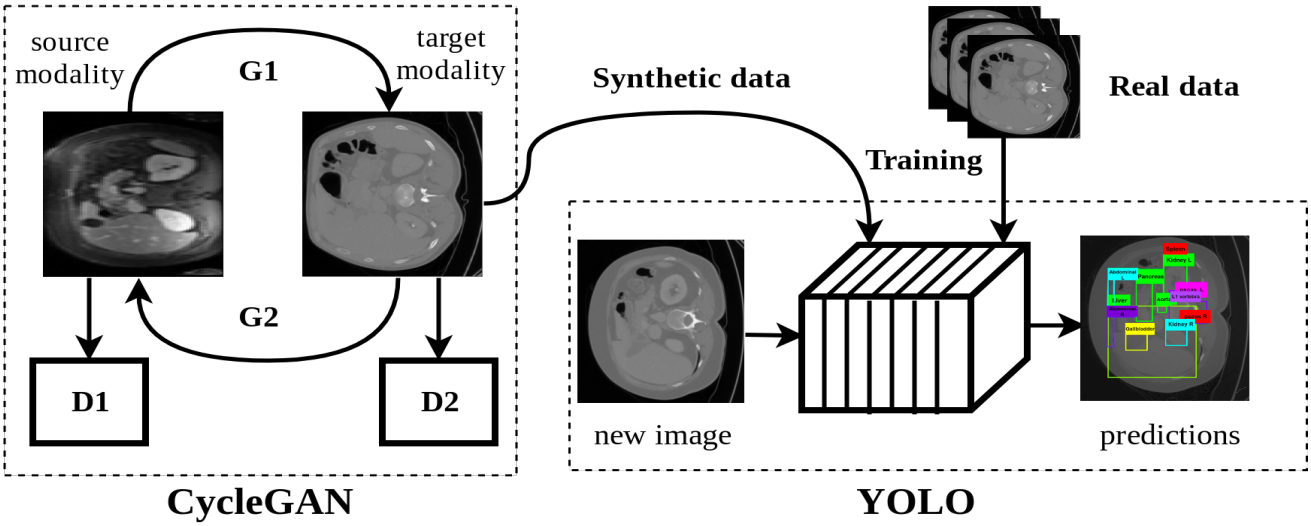


Fig. 1. The proposed framework : CycleGAN (image synthesis) + YOLO (multi-organ detection).

tested separately for few dozen structures in 2D slices of 3D CT images, and [8] which employs a convolutional network, trained and tested on chest CT structures simultaneously, which is augmented with spatial pyramid pooling to analyze 2D slices of different sizes.

B. Cross-modality Image Synthesis

A rich body of work that uses Generative Adversarial Networks (GAN) for synthesizing images in one modality from those in another has been proposed [11]. CycleGAN was proposed by [12] and became one of the commonly used approaches in synthetic medical image applications. Importantly, it can be used in the case of unpaired data, which is particularly useful in our application, since it is usually not possible to have images of different modalities for the same patient under the same conditions. In other words, instances are not mutually mapped between source and target domains, and therefore no registration is required.

CycleGAN was employed in [16] to generate brain CT images from MRI images, and in [17] to generate lung MRI images from CT images in order to segment lung tumors. Our work is inspired by [18] where the goal is to segment a single organ (the liver) without having ground-truth annotations for the target modality. A CycleGAN is used to generate target modality images from labeled source images. Source labels are then transferred to the target. As already stated, our work aims at multi-organ detection.

III. METHOD

As shown in Fig. 1, our workflow has 2 stages: cross-modality synthesis with CycleGAN [12] and multiple-organ detection with the YOLO algorithm [5].

YOLO was selected as the detector due to its speed and precision, as mentioned in the previous section. This approach starts by splitting an image into $S \times S$ cells. Each grid cell predicts three components: (1) coordinates (x, y, w, h) of B bounding boxes, (2) a confidence score $P(object)$, and (3) a class probability for C categories conditioned on the presence of an object in the bounding box. In our work, we use the third version of YOLO (YOLOv3) [19]. Its architecture is composed of 53 convolutional layers (Darknet-53). It makes predictions at three different scales. For a more stable prediction, scale-dependent box priors are used. These are learned from the training dataset. YOLOv3 also adds cross-layer connections between each two prediction layers except for the output layer. In the experimental work, we chose to apply YOLO on 2D full-resolution axial cross-sections of 3D images due to GPU memory restrictions and the complexity of the network model. A low resolution 3D approach would incur significant loss of precision for small structures. We train YOLO with 450 epochs and a decreasing learning rate.

CycleGAN [12] is an unsupervised deep learning method which allows bidirectional translation between the source X and the target domain Y. It uses two generator networks G_1, G_2 such as $G_1 : X \rightarrow Y$ and $G_2 : Y \rightarrow X$, each associated with a discriminator network, D_1 and D_2 following an adversarial training. G and D networks compete against each other. D works as a binary classifier attempting to distinguish between the synthetic and the real target image, while G seeks to deceive the discriminator by improving the quality of the synthetic output image. The input of the generator network G is a source domain image $x \in X$ and its output is a synthetic image, $\hat{y} = G(x)$. The inputs of a discriminator D are the synthetic output $\hat{y}$ and an unpaired random image from the target domain $y \in Y$. As for the architectures, the generator has an encoder, a transformer (a Residual Network in practice) and a decoder. The discriminator model is implemented as a PatchGAN model [12] which aims at classifying images as real or synthetic. The CycleGAN was trained using 200 epochs. We fixed the learning rate on 0.0002 for the first 100 epochs, then we linearly decay it until reaching zero over the rest.

Synthetic images that are generated using CycleGAN are

then used, along with the annotations of source images, to augment the training datasets for YOLO detectors.

IV. EXPERIMENTAL RESULTS

A. Datasets and pre-processing

The data used in this study comes from the Visceral Anatomy Benchmarks [20] and involves 2 datasets: (1) a *Gold* dataset, the annotations of which were created using manual segmentation, and (2) a *Silver* dataset, the labels of which were obtained by merging the segmentations produced by the algorithms of benchmark participants.

Both datasets consist of unpaired 3D contrast-enhanced thoracic-abdominal CT and abdominal MRI images providing 15 and 12 structure annotations respectively. Mean image dimensions are $512\times512\times438$ voxels for CT and $312\times72\times384$ voxels for MRI. The Gold dataset provides 20 patients per modality. We chose 30 patients per modality in the Silver dataset. To reduce the computational cost, all our experiments are carried out in 2D axial slices of original 3D images. We use the provided segmentation annotations to define 2D bounding-box annotations to train YOLO detectors.

For the experiments on cross-modality image synthesis, we crop CT images in both datasets around the abdomen because the thorax is absent in MRI images. We resize the images to 320×320 pixels for CT modality and 256×256 for MRI modality. These resolutions were chosen so that the dimensions of synthetic images are proportional to those of source images.

B. Performance metrics

We perform quantitative evaluations for two tasks: multiple-organ detection and cross-modality image synthesis. We use mean Average Precision (mAP) to select the best detector model over a validation set in a k-fold cross-validation procedure. This metric is conventionally computed as the area under the precision-recall curve. The best detector is then used to create the 2D bounding-box predictions on the test set, from which 3D bounding-boxes are constructed simply by taking maximum coordinates. We measure the 3D detection precision with respect to ground truth as the average distance over the 6 faces of the reconstructed 3D detection and the annotation bounding boxes.

For experiments on image synthesis, as in [21], we evaluate reconstruction fidelity as an indicator of synthesis quality via the Structural SIMilarity (SSIM) metric.

C. Cross-modality image synthesis

We perform image translations using CycleGAN across both modalities, i.e. from MRI to CT and from CT to MRI. An example of MRI to CT translation is presented in Fig. 2. It shows the consistency of translated structures with CT enhancement patterns, e.g. bright vertebra, kidneys brighter than muscles etc. To perform a quantitative evaluation of CycleGAN performance, we compute the SSIM between a source image x and its reconstruction $G_1(G_2(x))$. For the MRI to CT translation (Fig. 2), we have a mean SSIM of 0.97 (averaged over all patients). Conversely, for the CT to MRI translation (Fig. 3), we have a mean of SSIM of 0.91. This evaluation was performed on Gold dataset images with a model trained on those of the Silver dataset.

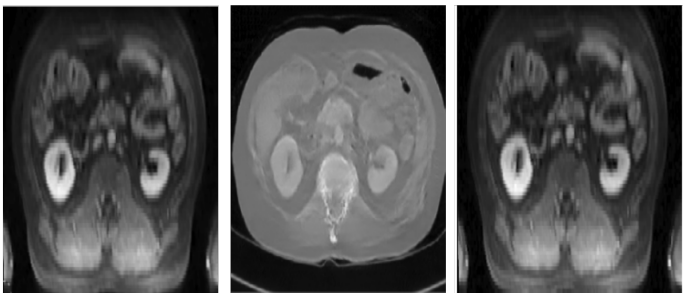


Fig. 2. Qualitative results of cross-modality generation (from MRI to CT image). The real MRI image (left), the generated CT image (center) and the reconstructed MRI image (right).

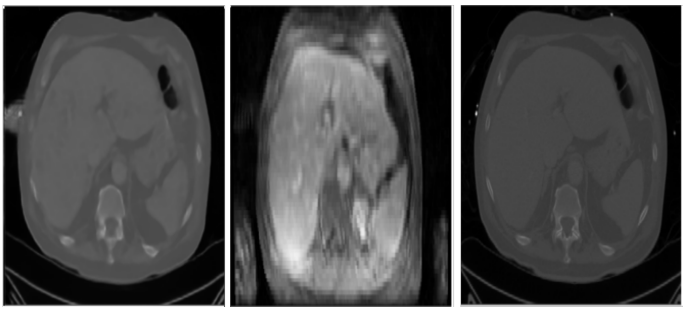
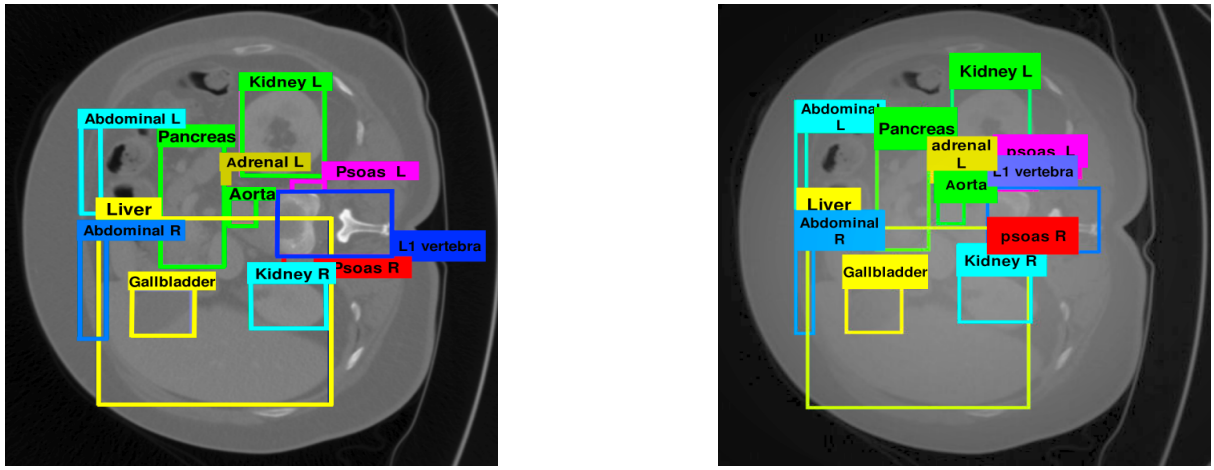


Fig. 3. Qualitative results of cross-modality generation (from CT to MRI image). The real CT image (left), the generated MRI image (center) and the reconstructed CT image (right).

D. Multi-organ detection

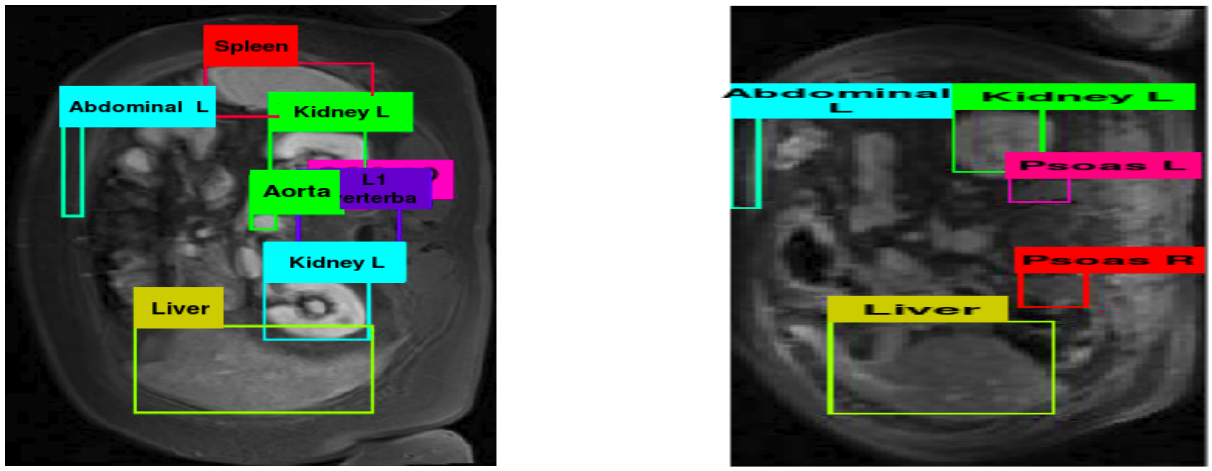
1) Qualitative evaluation: Fig. 4 and Fig. 5 show an example of multi-organ detection using YOLO respectively for an axial CT image and an axial MRI image. Fig. 4 is organized in 2 views; Ground truth and prediction. It shows the similarity of the predicted bounding box to the truth on the ground. Fig. 5 illustrates two 2D prediction MRI images. It shows that the bounding boxes are centered on the organs, even for smaller ones. Furthermore, the detector is able to detect an anomaly such as given in the right view of Fig. 5.b where it was able to detect that the patient has only one kidney.

2) Quantitative evaluation: Multiple-organ detection was evaluated on the Visceral Gold dataset using 10-fold cross-validation under two scenarios, without and with data augmentation. For the latter scenario, a CycleGAN trained on the Silver dataset was used to translate CT images in the Gold dataset into MRI images which were used to augment the training data in each of the 10 folds. The same scenario was used to augment CT image data from MRI. Test data are identical in both scenarios. For each fold, the model that yielded the best detection performance (measured by mAP) is selected.



(a) Ground Truth (b) Prediction

Fig. 4. 2D multi-organ detection on an axial CT image.



(a) Prediction 1 (b) Prediction 2

Fig. 5. 2D multi-organ detections an axial MRI images.

As previously stated, detection precision is measured in 3D on reconstructed bounding boxes, results are averaged over all images. Quantitative evaluation results using average distance are given in Table I and Table II. We observe that distances corresponding to large organs in both modalities such as the spleen (6.8 mm for CT and 11.7 mm for MRI) and the right kidney (5.6 mm for CT and 11.4 mm MRI) are satisfying, as opposed to organs that are difficult to detect such as the pancreas (14.3 mm for CT) and the muscle body of left rectus abdominis (55.9 mm for MRI).

The CycleGAN+YOLO scenario yields significantly better results for most organs in both modalities. For CT modality, the mean average distance is 7.95 mm as compared to YOLO alone 8.66 mm. This improvement is statistically significant ($p = 0.046$ on a paired one-sided t-test). For MRI modality, the mean average distance is 16.18 mm as compared to YOLO alone 22.62 mm. This improvement is statistically significant ($p = 0.050$ on a paired one-sided t-test).

Tables I-II also indicate that the standard deviation is high for several organs (e.g. the average distance of the right kidney 12.9 mm for CT and 15.7 mm for MRI modality). A careful examination of predictions confirms that this is due to outliers in the detection.

Regarding the running time of YOLO, we process an entire CT volume in 8 s, and an entire MRI volume in 3 s. All our models are trained and tested on NVIDIA Tesla V100 GPUs with 32 Gb of memory.

Table III compares our performances with state-of-the-art methods [1]–[3] applied to abdominal organs in contrast-enhanced CT images. This comparison is indicative as it was not possible to evaluate our method on the same datasets. Table III shows that YOLO and CycleGAN+YOLO yield best performances for the majority of studied organs in comparison with other methods.

TABLE I
YOLO (MEAN DIST. 8.66 MM) VS CYCLEGAN+YOLO (MEAN DIST. 7.95 MM) COMPARISON ON PER ORGAN MEAN DIST. FOR CT MODALITY.

	Pancreas	Gallbladder	Bladder
YOLO	14.3 ± 10.4	6.9 ± 10.9	4.0 ± 1.2
CycleGAN + YOLO	10.6 ± 5.2	7.4 ± 11.2	4.5 ± 1.6
	Verterba L1	Kidney R	Kidney L
YOLO	6.2 ± 3.5	5.6 ± 12.9	4.7 ± 5.8
CycleGAN + YOLO	5.8 ± 3.3	5.9 ± 12.4	4.3 ± 4.8
	Adrenal R	Adrenal L	Psoas R
YOLO	6.6 ± 6.5	8.1 ± 8.4	16.6 ± 13.4
CycleGAN + YOLO	6.3 ± 5.9	7.8 ± 8.7	11.8 ± 6.9
	Psoas L	Abdominal R	Abdominal L
YOLO	12.7 ± 7.1	13.6 ± 12.1	11.9 ± 7.3
CycleGAN + YOLO	12.8 ± 5.7	11.9 ± 6.7	12.2 ± 7.7
	Aorta	Liver	Spleen
YOLO	4.0 ± 3.0	7.4 ± 4.4	6.8 ± 7.0
CycleGAN + YOLO	3.9 ± 2.6	6.9 ± 3.4	6.5 ± 6.2

TABLE II
YOLO (MEAN DIST. 22.62 MM) VS CYCLEGAN+YOLO (MEAN DIST. 16.18 MM) COMPARISON ON PER ORGAN MEAN DIST. FOR MRI MODALITY.

	Pancreas	Gallbladder	Bladder
YOLO	17.9 ± 8.3	17.7 ± 10.4	15.1 ± 17.7
CycleGAN + YOLO	14.9 ± 5.6	13.9 ± 5.9	11.3 ± 12.0
	Verterba L1	Kidney R	Kidney L
YOLO	13.5 ± 5.0	11.4 ± 15.7	8.8 ± 11.9
CycleGAN + YOLO	9.7 ± 3.1	10.0 ± 15.6	10.1 ± 13.4
	Psoas R	Psoas L	Abdominal L
YOLO	12.0 ± 5.2	13.8 ± 7.1	55.9 ± 5.0
CycleGAN + YOLO	12.9 ± 5.6	12.5 ± 6.3	35.8 ± 34.3
	Aorta	Liver	Spleen
YOLO	81.1 ± 0.0	12.2 ± 7.5	11.7 ± 17.7
CycleGAN + YOLO	37.6 ± 20.6	14.0 ± 10.5	10.9 ± 6.9

TABLE III
COMPARISON OF CT IMAGES DETECTION WITH STATE-OF-THE-ART METHODS BASED ON MEAN DISTANCES PER ORGAN.

Method	Liver	Kidney R	Kidney L	Spleen	Gallbladder
Cuingnet [1]	12.2 ± 4	6.4 ± 4	6.8 ± 6	9.0 ± 5	11.8 ± 8
Criminsi [2]	14.0 ± 5	13.2 ± 6	12.3 ± 7	14.2 ± 6	15.5 ± 8
Gauriau [3]	10.7 ± 4	$\mathbf{5.6 \pm 3}$	5.5 ± 4	7.9 ± 4	9.5 ± 4
YOLO	7.4 ± 4	5.6 ± 12	4.7 ± 5	6.8 ± 7	$\mathbf{6.9 \pm 10}$
CycleGAN +Y.	$\mathbf{6.9 \pm 3}$	5.9 ± 12	$\mathbf{4.3 \pm 4}$	$\mathbf{6.5 \pm 6}$	7.4 ± 11

V. CONCLUSION AND FUTURE WORK

In this study, we proposed a CycleGAN+YOLO combination for data augmentation to train a multi-organ detector for multi-modality images. Our work tackles the scarcity of labeled medical data which hinders the supervised learning of deep networks by using a CycleGAN to generate synthetic images to augment training data. We showed that this approach achieves accurate detection with mean average distance of 7.95 ± 6.2 mm for CT modality, and mean average distance of 16.18 ± 11.6 mm for MRI modality. Further improvement of our results implies the development of a strategy that rejects detection outliers. This can be done by encoding anatomical constraints of proximity or adjacency as new terms in the loss function of the detector, to be optimized simultaneously with regression and class-probability terms.

REFERENCES

[1] R. Cuingnet, R. Prevost, D. Lesage, L.D. Cohen, B. Mory, and R. Ardon, "Automatic detection and segmentation of kidneys in 3d ct images using random forests," in *International Conference on Medical Image Computing and Computer–Assisted Intervention*. Springer, 2012, pp. 66–74.

[2] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Medical image analysis*, vol. 17, no. 8, pp. 1293–1303, 2013.

[3] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, "Multi–organ localization with cascaded global–to–local regression and shape prior," *Medical image analysis*, vol. 23, no. 1, pp. 70–83, 2015.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r–cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real–time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[7] J. Onieva, G.G. Serrano, T.P. Young, G.R. Washko, M.J. Ledesma-Carbayo, and Raúl S.J., "Multiorgan structures detection using deep convolutional neural networks," in *Medical Imaging 2018: Image Processing*. International Society for Optics and Photonics, 2018, vol. 10574, p. 1057428.

[8] B.D. De Vos, J.M. Wolterink, P.A. De Jong, T. Leiner, M.A. Viergever, and I. Išgum, "Convnet–based localization of anatomical structures in 3–d medical images," *IEEE transactions on medical imaging*, vol. 36, no. 7, pp. 1470–1481, 2017.

[9] S.-g. Lee, J.S. Bae, H. Kim, J.H. Kim, and S. Yoon, "Liver lesion detection from weakly-labeled multi–phase ct volumes with a grouped single shot multibox detector," in *International Conference on Medical Image Computing and Computer–Assisted Intervention*. Springer, 2018, pp. 693–701.

[10] J. Liu, D. Wang, L. Lu, Z. Wei, L. Kim, E.B. Turkbey, B. Sahiner, N.A. Petrick, and R.M. Summers, "Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks," *Medical physics*, vol. 44, no. 9, pp. 4630–4642, 2017.

[11] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, p. 101552, 2019.

[12] J.-Y. Zhu, T. Park, P. Isola, and A.A Efros, "Unpaired image–to–image translation using cycle–consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[14] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[15] T. Araújo, G. Aresta, A. Galdran, P. Costa, A.M. Mendonça, and A. Campilho, "Uolo–automatic object detection and segmentation in biomedical images," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 165–173. Springer, 2018.

[16] J.M. Wolterink, A.M. Dinkla, M.HF Savenije, P.R Seevinck, C.AT van den Berg, and I. Išgum, "Deep mr to ct synthesis using unpaired data," in *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2017, pp. 14–23.

[17] J. Jiang, Y.-C. Hu, N. Tyagi, P. Zhang, A. Rimner, G.S Mageras, J.O Deasy, and H. Veeraraghavan, "Tumor–aware, adversarial domain adaptation from ct to mri for lung cancer segmentation," in *International Conference on Medical Image Computing and Computer–Assisted Intervention*. Springer, 2018, pp. 777–785.

[18] Y. Huo, Z. Xu, S. Bao, A. Assad, R.G. Abramson, and B.A. Landman, "Adversarial synthesis learning enables segmentation without target modality ground truth," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1217–1220.

[19] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[20] Oscar Jimenez-del Toro et al., "Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2459–2475, 2016.

[21] Qi Y. Wu S. Jin, X., "Cyclegan face-off," *arXiv preprint arXiv:1712.03451*, 2017.