Semi-supervised Contrastive Learning for Label-Efficient Medical Image Segmentation

Xinrong Hu[1], Dewen Zeng[1], Xiaowei Xu[2], and Yiyu Shi[1](✉)

[1] University of Notre Dame, Notre Dame, IN 46556, USA
{xhu7,dzeng2,yshi4}@nd.edu
[2] Guangdong Provincial People's Hospital, Guangdong 510000, China

Abstract. The success of deep learning methods in medical image segmentation tasks heavily depends on a large amount of labeled data to supervise the training. On the other hand, the annotation of biomedical images requires domain knowledge and can be laborious. Recently, contrastive learning has demonstrated great potential in learning latent representation of images even without any label. Existing works have explored its application to biomedical image segmentation where only a small portion of data is labeled, through a pre-training phase based on self-supervised contrastive learning without using any labels followed by a supervised fine-tuning phase on the labeled portion of data only. In this paper, we establish that by including the limited label information in the pre-training phase, it is possible to boost the performance of contrastive learning. We propose a supervised local contrastive loss that leverages limited pixel-wise annotation to force pixels with the same label to gather around in the embedding space. Such loss needs pixel-wise computation which can be expensive for large images, and we further propose two strategies, downsampling and block division, to address the issue. We evaluate our methods on two public biomedical image datasets of different modalities. With different amounts of labeled data, our methods consistently outperform the state-of-the-art contrast-based methods and other semi-supervised learning techniques.

Keywords: Semi-supervised learning · Contrastive learning · Semantic segmentation · Label efficient learning

1 Introduction

Accurate semantic segmentation result is of great value to medical application, which provides physicians with the anatomical structural information for disease diagnosis and treatment. With the emergence of convolutional neural network,

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-87196-3_45) contains supplementary material, which is available to authorized users.

© Springer Nature Switzerland AG 2021
M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12902, pp. 481–490, 2021.
https://doi.org/10.1007/978-3-030-87196-3_45

supervised deep learning has achieved state-of-the-art performance in many biomedical image segmentation tasks, including different organs and different modalities [6,13,15,16]. These methods all need abundant labeled data with the class of every pixel or voxel known to guide the training. However, it is difficult to collect densely labeled medical images, because labeling medical images require domain-specific knowledge and pixel-wise annotations can be time-consuming.

To deal with this issue, various methods have been proposed. One branch of them is data augmentation, which expands labeled dataset by generative adversarial networks (GAN) [9] for data synthesis [3,5,8] or by simple linear combination[17]. However, data augmentation only leads to limited improvement of segmentation performance due to the defect of artificially generated images and labels. As a better alternative, semi-supervised learning based approaches [2,12,14,18] can utilize both labeled data and unlabeled data efficiently. Recently, contrastive learning [7,10], by forcing the embedding features of similar images to be close in the latent space and those of dissimilar ones to be apart, achieved state-of-the-art in self-supervised classification problems. The powerful ability of extracting features that are useful for downstream tasks from unlabeled data makes it a great candidate for label efficient learning.

However, most existing contrastive learning methods target image classification tasks. Semantic segmentation, on the other hand, requires pixel-wise classification. [4] first attempted at using contrastive learning to improve segmentation accuracy when only part of the data is labeled. In the pre-training phase, a two-step self-supervised contrastive learning scheme is used to learn both global and local features from unlabeled data. Specifically, it first projects a 2D slice to latent space with the encoder path only and computes a global contrastive loss, similar to what has been used for image classification problem. Then based on the trained encoder, decoder blocks are added and the intermediate feature maps are used to perform local contrastive learning at pixel level. Only a small number of points are selected from fixed positions in the feature maps. Next, in the fine-tuning phase the pre-trained model from contrastive learning is trained with supervision on the labeled part of the data. In such a framework, the label information is never utilized in the pre-training phase. Yet intuitively, by including such information to provide certain supervision, the features can be better extracted by contrastive learning.

Following this philosophy, in this work we propose a semi-supervised framework consisting of self-supervised global contrast and supervised local contrast to take advantage of the available labels. Compared with the unsupervised local contrast in the literature, the supervised one can better enforce the similarity between features within the same class and discriminate those that are not. In addition, we use two strategies, downsampling and block division, to address the high computation complexity associated with the supervised local contrastive loss. We conduct experiments on two public medical image segmentation datasets of different modalities. We find that, for different amount of labeled data, our framework can always generate segmentation results with higher Dice than the state-of-the-art methods.

2 Methods

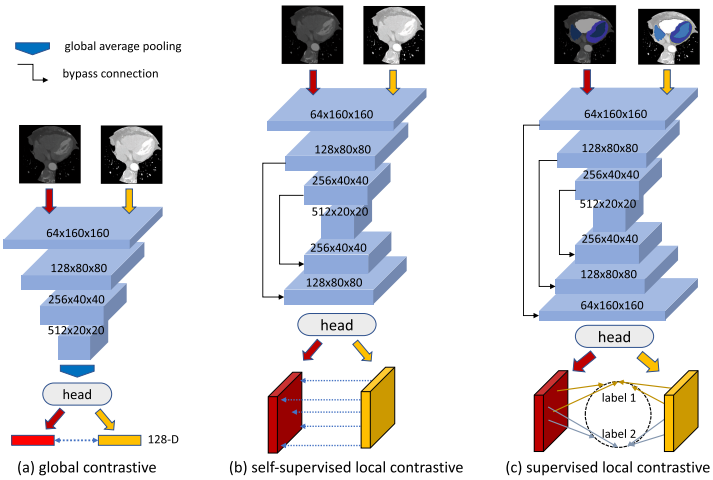

Fig. 1. Illustration of (a) self-supervised global contrastive loss, (b) self-supervised local contrastive loss used in existing works, and (c) our proposed supervised local contrastive loss.

2.1 Overview of the Semi-supervised Framework

The architecture backbone of a deep segmentation network usually consists of two parts, the encoder E and the decoder D. The encoder comprises several downsampling blocks to extract features from input images and the decoder path deploys multiple upsampling blocks to restore the resolution for pixel-wise prediction. In the pre-training stage, our framework contains a self-supervised global contrastive learning step which only uses unlabeled data to train E and captures image-level features, followed by a supervised local contrastive learning step which uses the limited labeled data to further train both E and D and captures pixel-level features.

The global contrastive learning is similar to what is used in the literature (e.g. SimCLR [7]) and for the sake of completeness we briefly describe it here with an illustration in Fig. 1(a). Given an input batch $B = \{x_1, x_2, ..., x_b\}$, in which x_i represents an input 2D slice, random combinations of image transformations $aug(\cdot)$ are performed twice on each image in B to form a pair of augmented images, same as the existing contrastive learning methods. These augmented images form an augmented image set A. Let $a_i, i \in I = \{1...2b\}$ represent an augmented image in A and let j(i) be the index of the other image in A that is

derived from the same image in B as a_i, i.e., a_i and $a_{j(i)}$ is an augmented pair. The global contrastive loss then takes the following form

$$L_g = -\frac{1}{|A|}\sum_{i\in I} log\frac{exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{k\in I-\{i\}} exp(z_i \cdot z_k)/\tau)}, \tag{1}$$

where z_i is the normalized output feature by adding a header $h_1(\cdot)$ to encoder E. The header is usually a multi-layer perceptron that enhances the representing ability of extracted features from E. Mathematically put, $z_i = |h_1(E(a_i))|$ and $z_{j(i)} = |h_1(E(a_{j(i)}))|$. τ is the temperature constant.

Only teaching the encoder to extract image-level features is not enough, since segmentation requires the prediction of all pixels in the input image. After the global contrastive learning, we attach the decoder path D to the encoder and retrain the whole network with supervised local contrastive loss. This is where our work differs from existing works and the details will be described in the following section. The two-step pre-training stage will eventually generate a model that can be used as an initialization for the subsequent fine-tuning stage on the labeled data to boost segmentation accuracy.

2.2 Supervised Local Contrastive Learning

The goal of local contrastive learning is to train the decoder to extract distinctive local representations. Let $f^l(\tilde{x_i}) = h_2(D_l(E(a_i)))$ be the output feature map of the l-th uppermost decoder block D_l for an augmented input a_i, where h_2 is a two-layer point-wise convolution. For feature map $f(a_i)$, the local contrastive loss is defined as

$$loss(a_i) = -\frac{1}{|\Omega|}\sum_{(u,v)\in\Omega}\frac{1}{|P(u,v)|}log\frac{\sum_{(u_p,v_p)\in P(u,v)} exp(f^l_{u,v}\cdot f^l_{u_p,v_p}/\tau)}{\sum_{(u',v')\in N(u,v)} exp(f^l_{u,v}\cdot f^l_{u',p'}/\tau)} \tag{2}$$

$f^l_{u,v} \in \mathbb{R}^c$, where c is the channel number, stands for the feature at the u-th column and v-th row of the feature map. Ω is the set of points in $f_{u,v}$ that are used to compute the loss. $P(u,v)$ and $N(u,v)$ denote the positive set and negative set of $f_{u,v}(a_i)$, respectively. Then the overall local contrastive loss is

$$L_l = \frac{1}{|A|}\sum_{a_i\in A} loss(a_i) \tag{3}$$

In self-supervised setting, [4] only selects 9 or 13 points from the feature map as illustrated by Fig. 2(b). Without annotation information, $P(u,v)$ only contains points from the same position at the feature maps of the paired images, that is $f_{u,v}(a_{j(i)})$, and the negative set is the union of Ω of all images in A minus the positive set.

According to SimCLR, when applying contrastive learning to image classification tasks, the data augmentation that boosts the accuracy most is the spatial

transformation, such as crop and cutout. Yet in self-supervised setting, the transformation $aug(\cdot)$ can only include intensity transformation such as color distort, Gaussian noise and Gaussian blur. It is because after spatial transformation, features at the same position of $f(a_i$ and $f(a_{j(i)})$ could originate from different points or areas but would still be treated as similar when calculating (2). Moreover, without any supervision, chances are that the selected points in Ω might all correspond to the background area, in which case the model should fail to effectively learn local features.

Inspired by [11], since the dataset contains limited labeled data anyway to be used in the fine-tuning stage, we can make use of it in the pre-training stage through a supervised local contrastive learning. Observing that the last point-wise convolution in U-Net functions similarly as the fully connected layer of classification network, we view the feature maps $f^1 \in \mathbb{R}^{c\times h\times h}$ just before the 1×1 convolutional layer as the projections of all pixels into the latent space, where c equals the channel number that is also the dimension of latent space and h is the input image dimension. Given a point (u, v) in the feature map, the definition of positive set $P(u, v)$ is all features in the feature maps of A that share the same annotation as (u, v). The negative set is then all features in the feature maps of A that are labeled as different classes. In addition, two embedding features of background pixels make up the majority of positive pairs and the subsequent segmentation task benefits less from comparing features as these pairs. Hence, for supervised setting, we change the definition of Ω to only contain points with non-background annotation. Please note that background piexls are still included in the negative set $N(u, v)$. Then the supervised contrastive loss takes the same form as (2), except for the new meaning of $\Omega, P(u, v), N(u, v)$ as discussed above. With the guidance of labels, supervised contrastive loss provides additional information on the similarity of features derived from the same class and dissimilarity of inter-class features.

2.3 Downsampling and Block Division

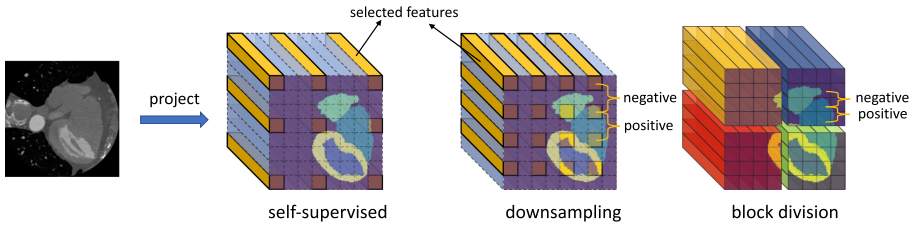

Fig. 2. Illustration of three strategies to select features from feature maps for local contrastive learning. Left: self-supervised as in [4]. Middle: downsampling where the contrastive loss is computed with features separated with fixed stride; Right: block division where contrastive loss is computed within each block and then averaged.

As we try to utilize the segmentation annotation, the feature map is of the same size as the input image. Though we only consider the points that do not belong to background, $|P(u,v)|$ can still be very large, that is $O(h^2)$ with input image dimension $h \times h$. The overall computation complexity for the superivsed local contrastive loss is then $O(h^4)$. For example, the image size in MMWHS is 512×512. Even after we reshape the slices to 160×160, the size of negative set is on the order of 10^4 and the total multiplications needed is on the order of 10^8 to evaluate the loss. To address this issue, we propose two strategies to reduce the size of the negative set as illustrated in Fig. 2. The first one is to downsample the feature map with fixed stride. The idea behind this method is that neighboring pixels contain redundant information and their embedded features appear quite similar in the latent space. With stride number s, the size of $P(u,v)$ and the computation complexity of the loss function are reduced by a factor of s^2 and s^4 respectively. As larger stride leads to fewer negative pairs and the number of negative pairs is important for contrastive learning [10]. In this work, we set the stride to 4, the minimum value that will not cause out-of-memory (OOM) issue on our experiment platform.

Alternatively, we can divide the feature maps into several blocks of the same size $h' \times h'$ as shown in Fig. 2. The local contrastive loss is calculated within each block only and then averaged across all the blocks. If all the pixels in a block are labeled as background, we will not compute its local contrastive loss. The size of $P(u,v)$ decreases to the number of features in a block $O(h'^2)$ and the total computation complexity is $O((h/h')^2 \cdot h'^4) = O(h^2 h'^2)$. Smaller block size will also lead to fewer negative pairs. Therefore the block size should also be determined based on the acceptable training time/memory consumption. In this paper, we set it as 16×16, the maximum value that will not cause OOM.

3 Experiments and Results

Dataset and Preprocessing. We evaluate our methods on two public medical image segmentation datasets of different modalities, MRI and CT. (1) Hippocampus dataset is from the MICCAI 2018 medical segmentation decathlon challenge[1]. There are 260 3D volumes of mono-modal MRI scans with the annotation of anterior and posterior hippocampus body. (2) MMWHS [19] comes from the MICCAI 2017 challenge, which consists of 20 3D cardiac CT volumes. The annotations include seven structures of heart: left ventricle, left atrium, right ventricle, right atrium, myocardium, ascending aorta, and pulmonary artery. Regarding preprocessing, we normalize the intensity for the whole volume. Then we resize the slices of all volumes in the same dataset to unify them using bilinear interpolation. The resolutions of the two reshaped dataset are 64×64 for Hippocampus and 160×160 for MMWHS.

Experiment Settings. Similar to [4], the segmentation network backbone we choose is 2D U-Net with three blocks for both encoder and decoder path, the implementation of which is based on the PyTorch library. For both sets, we divide the data into three parts, training set X_{tr}, validation set X_{vl}, and test set X_{ts}.

Table 1. Comparison between state-of-the-art methods and the proposed methods w.r.t. subsequent segmentation dice scores on two datasets. Ablation studies of our methods (w/o global contrast) are also included.

Methods	Hippocampus			MMWHS		
	% of data labeled in X_{tr}			% of data labeled in X_{tr}		
	5	10	20	10	20	40
Random	0.788	0.833	0.852	0.328	0.440	0.715
Global [7]	0.817	0.834	0.861	0.359	0.487	0.724
Global+local (self) [4]	0.808	0.843	0.858	0.367	0.490	0.730
Mixup [17]	0.818	0.847	0.861	0.365	0.541	0.755
TCSM [14]	0.796	0.838	0.855	0.347	0.489	0.733
Local(stride)	0.818	0.845	0.860	0.354	0.485	0.743
Local(block)	0.817	0.843	0.862	0.366	0.475	0.736
Global+local(stride)	0.822	**0.851**	0.863	**0.384**	0.525	0.758
Global+local(block)	**0.824**	0.849	**0.866**	0.382	**0.553**	**0.764**

The ratio $|X_{tr}|$:$|X_{vl}|$:$|X_{ts}|$ for Hippocampus is 3:1:1 and 2:1:1 for MMWHS. Then we randomly choose certain number of samples in X_{tr} as labeled volumes and the rest are viewed as unlabeled. We split the dataset four times independently to generate four different folds and the average dice score on X_{ts} is used as the metric for comparison. We choose Adam as the optimizer for both contrastive learning and the following segmentation training. The learning rate is 0.0001 for contrastive learning and is 0.00001 for segmentation training. The training epoch for contrastive learning is 70, and for segmentation training, it stops after 120 epochs. We run experiments of Hippocampus dataset on two NVIDIA GTX 1080 gpus and the experiments of MMWHS dataset on two NVIDIA Tesla P100 gpus.

We implement four variants of the proposed methods: the two-step global contrast and supervised local contrast with downsampling (global+local(stride)) and block division (global+local(block)); and for the purpose of ablation study, these two methods without global contrast (local(stride) and local(block)). For comparison, we implement two state-of-the-art contrast based methods: a global contrastive learning method [7] (global), and a self-supervised global and local contrastive learning method [4] (global+local(self)). Note that the entire training data is used but without any labels for global contrast and self-supervised local contrast, while only labeled training data is used for our supervised local contrast. We further implement three non-contrast baseline methods: training from scratch with random initialization (random), a data augmentation based method Mixup [17], and a state-of-the-art semi-supervised learning method TCSM [14].

Table 1 lists the average dice scores of all the methods. There are several important observations we can make from the table: First, all contrastive learning methods lead to higher dice scores compared to the random approach, which implies that contrastive learning is a feasible initialization when labels are limited. Second, we notice that the proposed supervised local contrast alone

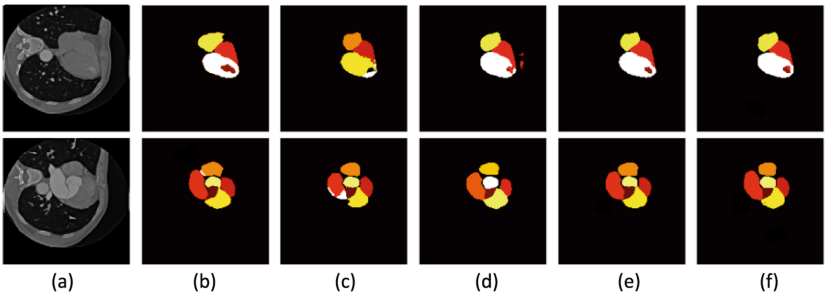

Fig. 3. Visualization of segmentation results on MMWHS with 40% labeled data. (a) and (b) represent input image and ground truth. (c)–(f) are the predictions of random, global+local(self), global+local(stride), and global+local(block). Different color represents different structure categories. (Better to view in colors)

(local(stride) and local(block)) generates segmentation results comparable with those from existing contrast based methods. Considering that the former only makes use of the labeled data while the latter leverage all volumes, the supervised contrastive loss efficiently teaches the model the pixel-wise representations shared by the same class. Lastly, combining the supervised local contrast with the global contrast, the semi-supervised setting gains a considerable improvement of segmentation performance over the state-of-the-art. It demonstrates that the image-level representation extracted from unsupervised global contrast is complementary to the pixel-level representation from the supervised local contrast.

Finally, the segmentation results on MMWHS with 40% labeled data from our methods, the state-of-the-art (global+local(self) [4]), and the random approach are shown in Fig. 3. From the figure we can clearly see that our approaches perform the best. With the same setting, the embedding features after applying t-SNE for these methods are visualized in Fig. 4. From the figure we can see that compared with the random approach, the features for the same class are tightly clustered and those for different classes are clearly separated for the global+local(block) method, while those from the state-of-the-art are more scattered. This aligns with the results in Table 1 where global+local(block) method achieves the highest Dice (rightmost column).

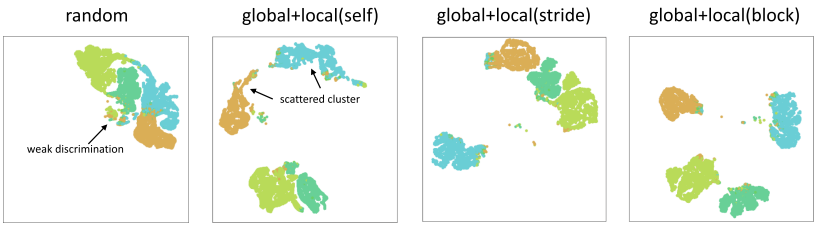

Fig. 4. Visualization of embedding features after applying t-SNE on MMWHS with 40% labeled data. Different color represents different classes. The features are all from the feature map before the last point-wise convolution. (Better to view in colors.)

4 Conclusion

The limited annotation is always a hurdle for the development of deep learning methods for medical image segmentation. Aiming at tackling this challenge, in this work, we propose a supervised local contrastive loss to learn the pixel-wise representation. Further, we put forward two strategies to reduce the computational complexity when computing the loss. Experiments on two public medical image datasets with only partial labels show that when combining the proposed supervised local contrast with global contrast, the resulting semi-supervised contrastive learning achieves substantially improved segmentation performance over the state-of-the-art.

Acknowledgement. This work is partially supported by NSF award IIS-2039538.

References

1. Medical segmentation decathlon challenge (2018). http://medicaldecathlon.com/index.html
2. Bai, W., et al.: Semi-supervised learning for network-based cardiac MR image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 253–260. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_29
3. Bowles, C., et al.: GAN augmentation: augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863 (2018)
4. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. arXiv preprint arXiv:2006.10511 (2020)
5. Chaitanya, K., Karani, N., Baumgartner, C.F., Becker, A., Donati, O., Konukoglu, E.: Semi-supervised and task-driven data augmentation. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 29–41. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_3
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
8. Costa, P., et al.: End-to-end adversarial retinal image synthesis. IEEE Trans. Med. Imag. **37**(3), 781–791 (2017)
9. Goodfellow, I.J., et al.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
11. Khosla, P., et al.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)

12. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3D semantic segmentation for medical images. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 552–561. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_54
13. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imag. **37**(12), 2663–2674 (2018)
14. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Networks Learn. Syst. **42**, 523 (2020)
15. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
18. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 408–416. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_47
19. Zhuang, X., Shen, J.: Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med. Image Anal. **31**, 77–87 (2016)