Craniomaxillofacial Bony Structures Segmentation from MRI with Deep-Supervision Adversarial Learning

Miaoyun Zhao[1], Li Wang[1], Jiawei Chen[1], Dong Nie[1], Yulai Cong[2], Sahar Ahmad[1], Angela Ho[3], Peng Yuan[3], Steve H. Fung[3], Hannah H. Deng[3], James Xia[3(✉)], and Dinggang Shen[1(✉)]

[1] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, USA
dgshen@med.unc.edu
[2] Department of Electrical and Computer Engineering, Duke University, Durham, USA
[3] Houston Methodist Hospital, Houston, TX, USA
jxia@houstonmethodist.org

Abstract. Automatic segmentation of medical images finds abundant applications in clinical studies. Computed Tomography (CT) imaging plays a critical role in diagnostic and surgical planning of craniomaxillofacial (CMF) surgeries as it shows clear bony structures. However, CT imaging poses radiation risks for the subjects being scanned. Alternatively, Magnetic Resonance Imaging (MRI) is considered to be safe and provides good visualization of the soft tissues, but the bony structures appear invisible from MRI. Therefore, the segmentation of bony structures from MRI is quite challenging. In this paper, we propose a cascaded generative adversarial network with deep-supervision discriminator (Deep-supGAN) for automatic bony structures segmentation. The first block in this architecture is used to generate a high-quality CT image from an MRI, and the second block is used to segment bony structures from MRI and the generated CT image. Different from traditional discriminators, the deep-supervision discriminator distinguishes the generated CT from the ground-truth at different levels of feature maps. For segmentation, the loss is *not only* concentrated on the voxel level *but also* on the higher abstract perceptual levels. Experimental results show that the proposed method generates CT images with clearer structural details and also segments the bony structures more accurately compared with the state-of-the-art methods.

1 Introduction

Generating a precise three-dimensional (3D) skeletal model is an essential step during craniomaxillofacial (CMF) surgical planning. Traditionally, computed tomography (CT) images are used in CMF surgery. However, a patient has to be exposed under radiation [1]. Magnetic Resonance Imaging (MRI), on the other hand, provides a safer scanning without radiation and non-invasive way to render CMF anatomy. However, it is extremely difficult to accurately segment CMF bony structures from MRI due to the

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11073, pp. 720–727, 2018.
https://doi.org/10.1007/978-3-030-00937-3_82

confusing boundaries between bones and air (both appearing to be black in MRI), low signal-to-noise ratio, and partial volume effect.

Recently, deep learning has demonstrated outstanding performance in a wide range of computer vision and image analysis applications. With a properly designed loss function, deep learning methods can automatically learn complex hierarchical features for a specific task. In particular, fully convolutional neural network (FCN) [2] was proposed to perform image segmentation by down-sampling and up-sampling streams. U-Net based methods further proposed skip connections to concatenate the lower fine feature maps to the higher coarse feature maps [3]. Nie *et al.* proposed a 3D deep-learning based cascade framework, in which a 3D U-Net is used to train a coarse segmentation and then a CNN is cascaded for fine-grained segmentation [4]. However, most of the previous works typically perform segmentation on the original MRI with low contrast for bony structures. Inspired by great success of Generative Adversarial Network (GAN) [5] in generating realistic images, we hypothesize that the segmentation problem can also be treated as an estimation problem, i.e., generating realistic CT images from MRIs and performing segmentation from the generated CT images. In this paper, we propose a framework of deep-supervision adversarial learning for CMF structure segmentation on the MR images. Our proposed framework consists of two major steps: (1) a simulation GAN to estimate a CT image from an MR image, and (2) a segmentation GAN to segment CMF bony structures based on both the original MR image and the generated CT image. Specifically, a CT image is first generated from a given MR image by a deep-supervision discriminative GAN, where a perceptive loss strategy is developed to obtain the knowledge from the real CT image *in terms of* both local detailed information and global structures. Furthermore, in segmentation task, with the proposed perceptive loss strategy, the discriminative GAN evaluates the segmentation results with the feature maps at different layers and the feedback structure information from both the original MR image and the generated CT image.

2 Method

In this section, we propose a cascaded generative adversarial network with deep-supervision discriminators (Deep-supGAN) to perform CMF bony structures segmentation from the MR image and generated CT image. The proposed framework is shown in Fig. 1. It includes two parts: (1) a simulation GAN that estimates a CT image from an MR image and (2) a segmentation GAN that segments the CMF bony structures based on both the original MR image and the generated CT image. The simulation GAN consists of the deep-supervision discriminators designed at each convolution layer to evaluate the quality of the generated image. In segmentation GAN, the deep-supervision perception loss is employed to evaluate the segmentation at multiple levels. Note that, for the discriminators of both parts, we utilize the first four convolution layers of a VGG-16 network [6] pre-trained on the ImgeNet dataset to extract the feature maps.

2.1 Simulation GAN

The simulation GAN for generating CT from MRI is shown in the upper portion of Fig. 1. Considering z as a ground-truth MRI patch, x as a ground-truth CT patch, and x' as a generated CT patch, we design a generator $G_c(z)$ to map a given MR image patch into a CT image patch. To make the generated CT image patch similar to the ground-truth CT image *in terms of* both local details and global structures, we design multiple deep-supervision discriminator $D_c^l(x)$, $(l = 1, 2, 3, \cdots)$. Here, $D_c^l(x)$ is a discriminator at the l-th layer of a pre-trained VGG-16 network, where each layer can extract features with different scales, from local details to global structures. Thus, each discriminator compares the generated CT with the ground-truth CT in different scales, resulting in an accurate simulation. To match the generated CT with the ground-truth CT, an adversarial game is played between $G_c(z)$ and $D_c^l(x)$. The loss function for the game is described as:

$$\min_{G_c} \max_{D_c^l} \mathbb{E}_{x\sim p(x)}\left[\sum_l \sum_{i,j} \log\left(\left[D_c^l(x)\right]_{i,j}\right)\right] + \mathbb{E}_{z\sim q(z)}\left[\sum_l \sum_{i,j} \log\left(1 - \left[D_c^l(G_c(z))\right]_{i,j}\right)\right] \tag{1}$$

where $p(x)$ is the distribution of the original CT data, $q(z)$ is the distribution of the original MRI data, $\left[D_c^l(x)\right]_{i,j}$ is the (i, j)-th element in matrix $D_c^l(x)$, and L is the number of layers connected with discriminator.

2.2 Segmentation GAN

Similarly, with the generated CT x' from $G_c(z)$, we can construct a segmentation GAN $G_s\left(z, x'\right)$, which learns to predict a bony structures segmentation y'. Then, the ground-truth y and the predicted segmentation y' are forwarded to the discriminator $D_s(y)$ to get an evaluation. Note that, different from the discriminator D_c^l in the simulation GAN, the discriminator $D_s(y)$ is only designed for the feature map at the last layer of the pre-trained VGG-16 net. The adversarial game for segmentation is as follows:

$$\min_{G_s} \max_{D_s} \mathbb{E}_{y\sim p(y)}[\log D_s(y)] + \mathbb{E}_{z,x'\sim q(z,x')}[\log(1 - D_s(G_s(z, x')))] \tag{2}$$

where $\mathrm{p}(y)$ is the distribution of ground-truth segmentation images, and $q(z, x')$ is the joint distribution of the original MRI and the generated CT data. For the segmentation results, a voxel-wise loss is intuitively considered as follows:

$$\mathcal{L}_{vox} = \mathbb{E}_{z,x'\sim q(z,x')}\left\|G_s(z, x') - y\right\|^2 \tag{3}$$

Moreover, we also consider a perceptual loss L_{percp}^l to encourage the consistence of features maps from generated segmentation and ground-truth segmentation. To this end, the pre-trained part of the discriminator is utilized to extract multi-layer feature maps from the generated segmentation and ground-truth segmentation. Taking $\varphi_l(y)$ as the feature map of input y at the l-th layer of the feature extraction network, and N_l as

the number of voxels in feature map $\varphi_l(y)$, we can obtain the perceptual loss for the l-th layer as follows:

$$\mathcal{L}^l_{percp} = \mathbb{E}_{z,x' \sim q(z,x')}\left[\frac{1}{N_l}\|\varphi_l(G_s(z, x')) - \varphi_l(y)\|^2\right] \tag{4}$$

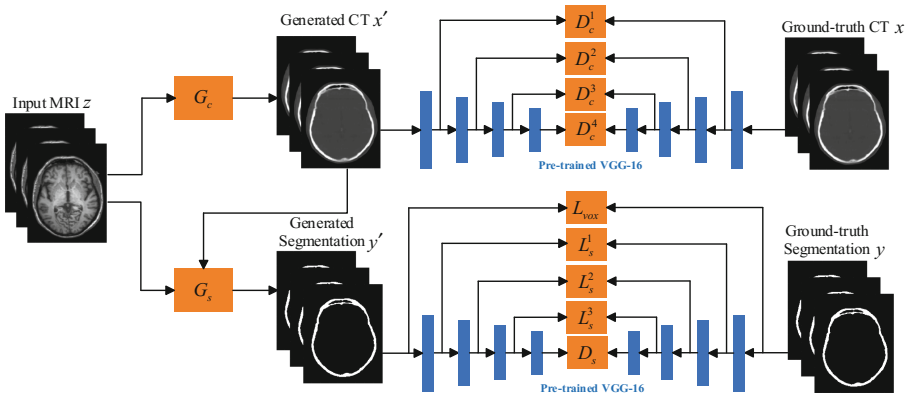

Fig. 1. The overview of the proposed Deep-supGAN. **Top:** CT generation net, where the generator G_c takes MRI patch z as input and generates the corresponding CT patch x', while the discriminator D_c^l takes generated CT patch x' and ground-truth CT patch x as input and produces classification (ground-truth = 1, generated = 0). **Bottom:** segmentation net, where G_s takes MRI patch z and generated CT patch x' as input and then generates the segmentation y', while D_s takes the generated segmentation y' and the ground-truth segmentation y as input and produces classification (ground-truth = 1, generated = 0).

In summary, the total loss function *with respect to* the generator is:

$$\min_{G_s} \mathbb{E}_{z,x' \sim q(z,x')}[-\log D_s(G_s(z, x'))] + \lambda_1 \mathcal{L}_{vox} + \lambda_2 \sum_{l=1}^{L} \mathcal{L}^l_{percp} \tag{5}$$

where parameters λ_1 and λ_2 are utilized to balance the importance of the three loss functions.

3 Experimental Results

3.1 Dataset

The experiments were conducted on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [7]. It consists of 16 subjects with paired MRI and CT scans. The MRI scans were obtained by a Siemens Triotim scanner, with a voxel size of 1.2 × 1.2 1 mm^3, TE 2.95 ms, TR 2300 ms, and flip angle 9. The CT scans were obtained from a Siemens Somatom scanner, with a voxel size of 0.59 × 0.59 × 3 mm^3.

The preprocessing was conducted as follows. Both MRI and CT scans were resampled to size 152 × 184 × 149 with a voxel size of 1 × 1× 1 mm^3. Each CT was aligned with its corresponding MRI. All intensities of MRI and CT were rescaled into [− 1, 1]. To be compatible with VGG-16 net, both MRI and CT data were cropped into patches of size 152 × 184 × 3 for training. The experiments were conducted on the 16 subjects in a leave-one-out cross validation. To measure the quality of the generated CT, we used the mean absolute error (MAE) and peak-signal-to-noise-ratio (PSNR). To measure the segmentation accuracy, we used Dice similarity coefficient (DSC). We adopted TensorFlow to implement the proposed framework. The network was trained using Adam with a learning rate of 1e−4 and a momentum of 0.9. In the experiments, we empirically set the parameters in the proposed method as: $L = 4$, $\lambda_1 = 1$ and $\lambda_2 = 1$.

3.2 Impact of Deep-Supervision Feature Maps

To evaluate the effectiveness of the deep-supervision strategy on the simulation GAN, we train the network with the discriminator in different layers of the pre-trained VGG-16 network. The results are shown in Fig. 2. It is obvious that the lower layer the discriminator is applied, the clearer the results will be. A quantitative comparison is shown in Table 1, indicating that, when the lower layer is connected with discriminator, the PSNR is bigger and the MAE is smaller.

To evaluate the effectiveness of the deep-supervision strategy on the segmentation GAN, we train the network with the perception reconstruction loss in different layers of the pre-trained VGG-16 network. As shown in Fig. 3, the results with higher layer connected with perceptual loss, $\mathcal{L}^2_{percp}$ and $\mathcal{L}^3_{percp}$, are more smooth and accurate in thin structures, as shown in the yellow rectangles. The DSC of different layer connected with perceptual loss is provided in Table 2, which again indicates that the deep-supervision perceptual loss enhances the performance greatly.

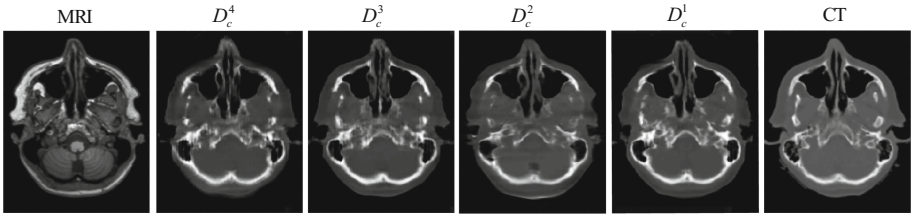

Fig. 2. CT generated by proposed Deep-supGAN with different layers connected with the discriminator. **Left to right:** original MRI, four CT images generated with the fourth, third, second, and first layer respectively connected with the discriminator, and ground-truth CT.

Table 1. PSNR and MAE with different layer connected with the discriminator.

Layer ID	1	2	3	4
PSNR	23.03	23.95	24.40	25.11
MAE (%)	1.99	1.61	1.45	1.23

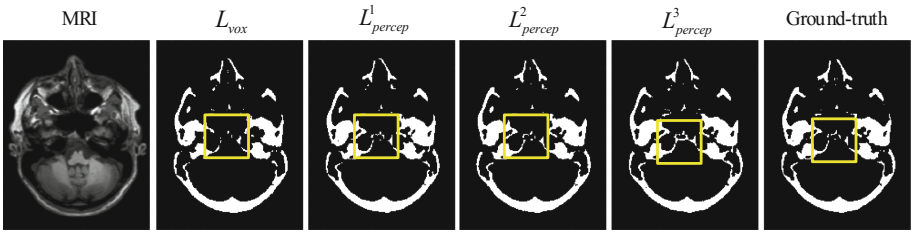

Fig. 3. Segmentation of proposed Deep-supGAN with different layers connected with the perceptual loss.

Table 2. DSC (%) of proposed Deep-supGAN with different layers connected with the perceptual loss.

Layer ID	$\mathcal{L}_{vox}$	$\mathcal{L}^1_{percp}$	$\mathcal{L}^2_{percp}$	$\mathcal{L}^3_{percp}$
DSC (%)	90.52	92.16	94.24	94.46

3.3 Impact of Generated CT

To evaluate the contribution of generated CT to the segmentation results, the segmentation results *only with MRI as input* (denoted as with MRI) is shown in Fig. 4. The segmentation result *with both original MRI and generated CT as input* (denoted as with MRI + CT) is more smooth and complete for thin structures, especially in the regions indicated by the yellow rectangles. The quantitative comparison in terms of DSC is shown in Table 3. It can be seen that the performance is significantly improved with the generated CT.

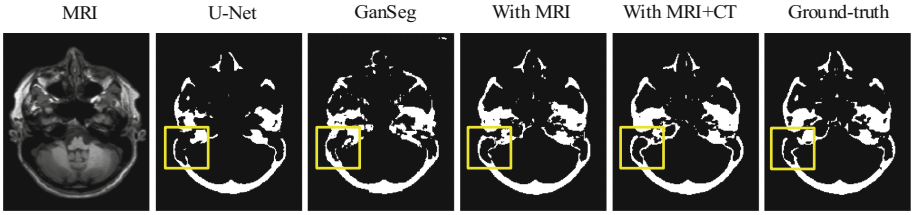

Fig. 4. Segmentation results by comparison methods. **Left to right**: Original MRI, U-Net, GanSeg, our method with MRI, our method with MRI+CT, and ground-truth.

Table 3. DSC (%) of compared methods on 16 subjects using leave-one-out cross validation.

Methods	U-Net [3]	GanSeg [8]	With MRI	With MRI + CT
DSC (%)	85.47	83.30	89.94	94.46

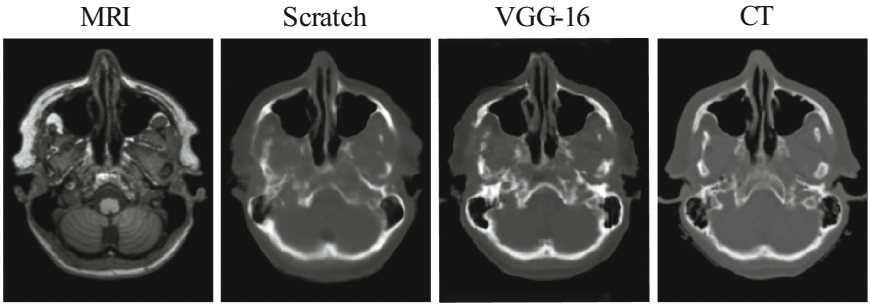

Fig. 5. The impact of pre-trained VGG-16 for generating CT. **Left to right:** original MRI, two CT images generated respectively by our method with discriminator (1) trained from scratch and (2) utilizing a pre-trained VGG-16, and ground-truth CT.

3.4 Impact of Pre-trained VGG-16 Network

Here we compare the generated CT with two different training settings: (1) learning the discriminator from scratch (denoted as Scratch) and (2) utilizing a pre-trained VGG-16 network (denoted as VGG-16) for the discriminator. As shown in Fig. 5, the CT generated with pre-trained VGG-16 is much clearer and more realistic than that trained from scratch.

3.5 Comparison with State-of-the-Art Segmentation Methods

To illustrate the advantage of our method on bony structures segmentation, we also compared it with two widely-used deep learning methods, i.e., U-Net [3] based segmentation method and Generative Adversarial Network based semantic segmentation method [8] (denoted as GanSeg, a traditional GAN with the generator designed as segmentation network). Comparison results on a typical subject are shown in Fig. 4. It can be seen that both U-Net and GanSeg failed to accurately segment bony structures, as indicated by yellow rectangles. Compared with these two methods, our proposed method can achieve more accurate segmentation. The quantitative comparison in terms of DSC is shown in Table 3. It clearly demonstrates the advantage of our proposed method in terms of segmentation accuracy.

4 Conclusion

In this paper, we proposed a cascade GAN network, Deep-supGAN, to segment CMF bony structures from the combination of an original MRI and a generated CT image. A GAN with deep-supervision discriminator is designed to generate a CT image from an MRI. With the generated CT image, a GAN with deep-supervision perceptual loss is designed to perform bony structures segmentation using both original MRI and the generated CT image. The combination of MRI and CT image can provide complementary information about bony structures for the segmentation task. Comparisons with the state-of-the-art methods demonstrate the advantage of our proposed method in terms of segmentation accuracy.

References

1. Brenner, D.J., Hall, E.J.: Computed tomography-an increasing source of radiation exposure. N. Engl. J. Med. **357**(22), 2277–2284 (2007)
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on CVPR, pp. 3431–3440 (2015)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, Mert R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
4. Nie, D., et al.: Segmentation of craniomaxillofacial bony structures from MRI with a 3D deep-learning based cascade framework. In: Wang, Q., Shi, Y., Suk, H.-I., Suzuki, K. (eds.) MLMI 2017. LNCS, vol. 10541, pp. 266–273. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_31
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
7. Trzepacz, P.T., Yu, P., Sun, J., et al.: Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer's dementia. Neurobiol. Aging **35**(1), 143–151 (2014)
8. Luc, P., Couprie, C., Chintala, S.: Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 (2016)