Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-Ray Segmentation

Cheng Chen[1(✉)], Qi Dou[1], Hao Chen[1,2], and Pheng-Ann Heng[1]

[1] Dept. of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
cchen@cse.cuhk.edu.hk
[2] Imsight Medical Technology Co., Ltd., Shenzhen, China

Abstract. In spite of the compelling achievements that deep neural networks (DNNs) have made in medical image computing, these deep models often suffer from degraded performance when being applied to new test datasets with domain shift. In this paper, we present a novel unsupervised domain adaptation approach for segmentation tasks by designing semantic-aware generative adversarial networks (GANs). Specifically, we transform the test image into the appearance of source domain, with the semantic structural information being well preserved, which is achieved by imposing a nested adversarial learning in semantic label space. In this way, the segmentation DNN learned from the source domain is able to be directly generalized to the transformed test image, eliminating the need of training a new model for every new target dataset. Our domain adaptation procedure is unsupervised, without using any target domain labels. The adversarial learning of our network is guided by a GAN loss for mapping data distributions, a cycle-consistency loss for retaining pixel-level content, and a semantic-aware loss for enhancing structural information. We validated our method on two different chest X-ray public datasets for left/right lung segmentation. Experimental results show that the segmentation performance of our unsupervised approach is highly competitive with the upper bound of supervised transfer learning.

1 Introduction

Deep neural networks (DNNs) have achieved great success in automated medical image computing [4,11], attributing to their learned highly-representative features. However, due to domain shift, DNNs would suffer from performance degradation when being applied to new datasets, which are acquired with different protocols or collected from different clinical centers [6,9]. Actually, domain adaptation has been an important research topic in medical image computing and the traditional automated methods also encountered the same poor generalization problem. For example, Philipsen *et al.* [10] studied the influence of data

distribution variations across chest radiography datasets on traditional segmentation methods based on k-nearest neighbor classification as well as active shape modeling.

To generalize DNNs trained on a *source domain* to a *target domain*, researches have been emerging for domain adaptation of deep learning models. A typical method is supervised transfer learning (STL), which fine-tunes the pre-trained source domain model with additional labeled target domain data. Remarkably, Ghafoorian *et al.* [6] studied on the number of fine-tuned layers to reduce the required amount of annotations for brain lesion segmentation across MRI datasets. However, the STL approaches still rely on extra labeled data, which is expensive or sometimes impractical to obtain.

Instead, unsupervised domain adaptation (UDA) is more appealing to generalize models in clinical practice. Early works have employed histogram matching to make test data resemble the intensity distribution of source domain data [15]. Recently, the generative adversarial networks (GANs) have made great achievements in generating realistic images and adversarial learning excels in mapping data distributions for domain adaptation [1,5,14]. In medical field, adversarial frameworks have been proposed to align feature embeddings between source and target data and presented promising results on cross-protocol brain lesion segmentation [9] and cross-modality cardiac segmentation [3]. Recent works adopted CycleGAN [16] as a data augmentation step to synthesize images from source domain to target domain, and used the pair of synthetic image and corresponding source label to train a segmentation model for target domain [2,7]. However, the synthetic images can be distorted on semantic structures, because the pure CycleGAN did not explicitly constrain the output of each single generator inside the cycle.

In this work, we propose a semantic-aware generative adversarial networks for unsupervised domain adaptation (named *SeUDA*) of medical image segmentation. Our method detaches the segmentation DNN from the domain adaptation process, and does not require any label from the test set. Given a test image, our *SeUDA* framework conducts image-to-image transformation to generate a source-like image which is directly forwarded to the established source DNN. To enhance the preservation of structural information during image transformation, we improve CycleGAN with a novel semantic-aware loss by embedding a nested adversarial learning in semantic label space. We validated our *SeUDA* on two different chest X-ray public datasets for lung segmentation. The performance of our unsupervised method exceeds the UDA baseline and is highly competitive with that of the supervised transfer learning. Last but not least, our transformed image results are visually observable, which sheds the light on the explicit intuition of our proposed method.

2   Method

Given the source domain images $x^s \in \mathcal{X}^s$ and the corresponding labels $y^s \in \mathcal{Y}$, we train a DNN model, denoted by f^s, which learns to segment the input images.
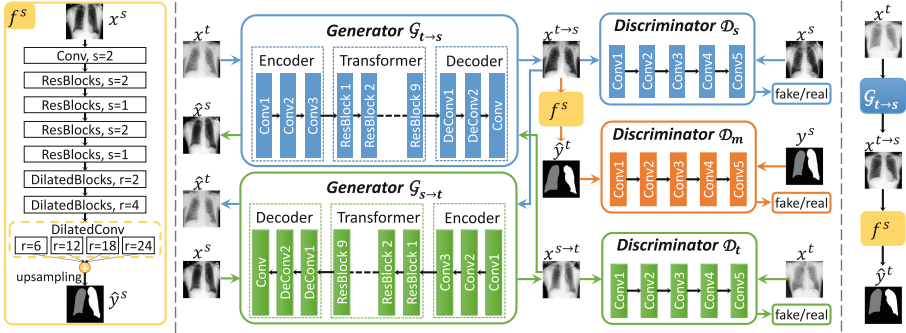
Fig. 1. The overview of our unsupervised domain adaptation framework. **Left**: the segmentation DNN learned on source domain; **Middle**: the *SeUDA* where the paired generator and discriminator are indicated with the same color, the blue/green arrows illustrate the data flows from original images (x^t/x^s) to transformed images ($x^{t \to s}/x^{s \to t}$) then back to reconstructed images ($\hat{x}^t/\hat{x}^s$) in cycle-consistency loss, the orange part is the discriminator for the semantic-aware adversarial learning; **Right**: the inference process of *SeUDA* given a new target image for testing.

In UDA, we have unlabeled target images $x^t \in \mathcal{X}^t$ whose intensity distributions (or visual appearances) are not the same as the source domain data. Figure 1 presents an overview of our proposed *SeUDA* framework, which adapts the appearance of x^t to source image space $\mathcal{X}^s$, so that the established f^s can be directly generalized to the transformed image.

2.1   Segmentation Network Established on Source Domain

Our segmentation DNN model (referred as segmenter) is detached from the learning of our domain adaptation GANs. Compared with the integrated approaches, this advantage of an independent segmenter enables much more flexibility when designing a high-performance network architecture. In this regard, we establish a state-of-the-art segmentation network which makes complementary use of the residual connection, dilated convolution and multi-scale feature learning [11].

The backbone of our segmenter is modified ResNet-101. We replace the standard convolutional layers in the high-level residual blocks with the dilated convolutions. To leverage features with multi-scale receptive fields, we replace the last fully-connected layer with four parallel 3×3 dilated convolutional branches, with a dilation rate of {6, 12, 18, 24}, respectively. An upsampling layer is added in the end to produce dense predictions for the segmentation task. We start with 32 feature maps in the first layer and double the number of feature maps when the spatial size is halved or the dilation convolutions are utilized. The segmenter is optimized by minimizing the pixel-wise multi-class cross-entropy loss of the prediction $f^s(x^s)$ and ground truth y^s with standard stochastic gradient descent.

2.2   Image Transformation with Semantic-Aware GANs

After obtaining f^s which maps the source input space $\mathcal{X}^s$ to the semantic label space $\mathcal{Y}$, our goal is to make it generally applicable to new target datasets. Given that annotating medical data is quite expensive, we conduct the domain adaptation in an unsupervised manner. Specifically, we map the target images towards the source data space. The generated new image $x^{t\to s}$ appears to be drawn from $\mathcal{X}^s$ while the content and semantic structures remain unchanged. In this way, we can directly apply the well-established model f^s on $x^{t\to s}$ without re-training and get the segmentation result for x^t.

To achieve this, we first construct a generator $\mathcal{G}_{t\to s}$ and a discriminator $\mathcal{D}_s$. The generator aims to produce realistic transformed image $x^{t\to s} = \mathcal{G}_{t\to s}(x^t)$. The discriminator competes with the generator by trying to distinguish between the fake generated data $x^{t\to s}$ and the real source data x^s. The GAN corresponds to a minimax two-player game and is optimized via the following objective:

$$\mathcal{L}_{\mathrm{GAN}}(\mathcal{G}_{t\to s}, \mathcal{D}_s) = \mathbb{E}_{x^s}[\log\mathcal{D}_s(x^s)] + \mathbb{E}_{x^t}[\log(1 - \mathcal{D}_s(\mathcal{G}_{t\to s}(x^t)))], \qquad (1)$$

where the discriminator tries to maximize this objective to correctly classify the $x^{t\to s}$ and x^s, while the generator tries to minimize $\log(1-\mathcal{D}_s(\mathcal{G}_{t\to s}(x^t)))$ to learn the data distribution mapping from $\mathcal{X}^t$ to $\mathcal{X}^s$.

Cycle-consistency adversarial learning. For image transformation, the generated $x^{t\to s}$ should also preserve the detailed contents in the original x^t. Inspired by the CycleGAN [16] which sets the state-of-the-art for unpaired image-to-image transformation, we employ the cycle-consistency loss during the adversarial learning to maintain the contents with clinical clues of the target images.

Inversely, we build a source-to-target generator $\mathcal{G}_{s\to t}$ and a target discriminator $\mathcal{D}_t$, so that the transformed image can be translated back to the original image. This pair of models are trained with a same-way GAN loss $\mathcal{L}_{\mathrm{GAN}}(\mathcal{G}_{s\to t}, \mathcal{D}_t)$ following the Eq. (1). In this regard, we derive the cycle-consistency loss which encourages $\mathcal{G}_{s\to t}(\mathcal{G}_{t\to s}(x^t)) \approx x^t$ and $\mathcal{G}_{t\to s}(\mathcal{G}_{s\to t}(x^s)) \approx x^s$ in the transformation:

$$\mathcal{L}_{\mathrm{cyc}}(\mathcal{G}_{t\to s}, \mathcal{G}_{s\to t}) = \mathbb{E}_{x^t}[||\mathcal{G}_{s\to t}(\mathcal{G}_{t\to s}(x^t)) - x^t||_1] + \mathbb{E}_{x^s}[||\mathcal{G}_{t\to s}(\mathcal{G}_{s\to t}(x^s)) - x^s||_1], \qquad (2)$$

where the L1-Norm is employed for reducing blurs in the generated images. This loss imposes the pixel-level penalty on the distance between the cyclic transformation result and the input image.

Semantic-aware adversarial learning. In our proposed *SeUDA*, we apply the established f^s to $x^{t\to s}$ which is obtained by inputting x^t to $\mathcal{G}_{t\to s}$. The image quality of $x^{t\to s}$ and the stability of $\mathcal{G}_{t\to s}$ are crucial for the effectiveness of our method. Therefore, besides the cycle-consistency loss which composes both generators and constraints the cyclic input-output consistency, we further try to explicitly enhance the intermediate transformation result $x^{t\to s}$. Specifically, for our segmentation domain adaptation task, we design a novel semantic-aware loss which aims to prevent the semantic distortion during the image transformation.

In our unsupervised learning scenario, we establish a nested adversarial learning module by adding another new discriminator $\mathcal{D}_m$ into the system. It distinguishes between the source domain ground truth lung mask y^s and the predicted lung mask $f^s(x^{t \rightarrow s})$ obtained by applying the segmenter on the source-like transformed image. Our underlying hypothesis is that the shape of anatomical structure is consistent across multi-center medical images. The prediction of $f^s(x^{t \rightarrow s})$ should follow the regular semantic structures of the lung to fool the $\mathcal{D}_m$, otherwise, the generator $\mathcal{G}_{t \rightarrow s}$ would be penalized by the semantic-aware loss:

$$\mathcal{L}_{\text{sem}}(\mathcal{G}_{t \rightarrow s}, \mathcal{D}_m) = \mathbb{E}_{y^s}[\log \mathcal{D}_m(y^s)] + \mathbb{E}_{x^t}[\log(1 - \mathcal{D}_m(f^s(\mathcal{G}_{t \rightarrow s}(x^t))))]. \quad (3)$$

This loss imposes an explicit constraint on the intermediate result of the cyclic transformation. Its gradients can assist the update of the generator $\mathcal{G}_{t \rightarrow s}$, which benefits the stability of the entire adversarial learning procedure.

2.3 Learning Procedure and Implementation Details

The configurations of the generators and discriminators follow the practice of [16]. Specifically, both generators have the same architecture consisting of an encoder (3 convolutions), a transformer (9 residual blocks) and a decoder (2 deconvolutions and 1 convolution). All the three discriminators process 70×70 patches and produce real/fake predictions via 3 stride-2 and 2 stride-1 convolutional layers. The overall objective for the generators and discriminators is as follows:

$$\mathcal{L}(\mathcal{G}_{s \rightarrow t}, \mathcal{G}_{t \rightarrow s}, \mathcal{D}_s, \mathcal{D}_t, \mathcal{D}_m) = \mathcal{L}_{GAN}(\mathcal{G}_{s \rightarrow t}, \mathcal{D}_t) + \alpha \mathcal{L}_{GAN}(\mathcal{G}_{t \rightarrow s}, \mathcal{D}_s) +$$
$$\beta \mathcal{L}_{\text{cyc}}(\mathcal{G}_{t \rightarrow s}, \mathcal{G}_{s \rightarrow t}) + \lambda \mathcal{L}_{\text{sem}}(\mathcal{G}_{t \rightarrow s}, \mathcal{D}_m), \quad (4)$$

where the $\{\alpha, \beta, \lambda\}$ denote trade-off hyper-parameters adjusting the importance of each component, which is empirically set to be $\{0.5, 10, 0.5\}$ in our experiments. The entire framework is optimized to obtain:

$$\mathcal{G}_{s \rightarrow t}^*, \mathcal{G}_{t \rightarrow s}^* = \arg \min_{\substack{\mathcal{G}_{s \rightarrow t} \\ \mathcal{G}_{t \rightarrow s}}} \max_{\mathcal{D}_s, \mathcal{D}_t, \mathcal{D}_m} \mathcal{L}(\mathcal{G}_{s \rightarrow t}, \mathcal{G}_{t \rightarrow s}, \mathcal{D}_s, \mathcal{D}_t, \mathcal{D}_m). \quad (5)$$

In our *SeUDA*, the generators $\{\mathcal{G}_{t \rightarrow s}, \mathcal{G}_{s \rightarrow t}\}$ and discriminators $\{\mathcal{D}_s, \mathcal{D}_t, \mathcal{D}_m\}$ are optimized altogether and updated successively. Note that the segmenter f^s is not updated in the process of image transformation. In practice, when training the generative adversarial networks, we followed the strategies of [16] for reducing model oscillation. Specifically, the negative log likelihood in $\mathcal{L}_{GAN}$ was replaced by a least-square loss to stabilize the training. The discriminator loss was calculated using one image from a collection of fifty previously generated images rather than the one produced in the latest training step. We used the Adam optimizer with an initial learning rate of 0.002, which was linearly decayed every 100 epochs. We implemented our proposed framework on the TensorFlow platform using an Nvidia Titan Xp GPU.

Table 1. Quantitative evaluation results of domain adaptation methods for both lung segmentations from chest X-ray images.

| Methods	Right lung				Left lung			
	Dice	Recall	Precision	ASD	Dice	Recall	Precision	ASD
S-test	95.98	97.98	94.23	2.23	95.23	96.56	94.01	2.45
T-noDA	82.29	98.40	73.38	10.68	76.65	95.06	69.15	11.40
T-HistM [15]	90.05	92.96	88.05	5.72	91.03	94.35	88.45	4.66
T-FeatDA [9]	94.85	93.66	96.42	3.26	92.93	91.67	94.46	3.80
T-STL [6]	96.91	98.47	95.46	1.93	95.84	97.48	94.29	2.20
CyUDA	94.09	96.31	92.28	3.88	91.59	92.28	91.70	4.57
SeUDA	95.59	96.55	94.77	2.85	93.42	92.40	94.70	3.51

3   Experimental Results

Datasets and Evaluation Metrics. We validated our unsupervised domain adaptation method for lung segmentations using two public Chest X-ray datasets, i.e., the Montgomery set (138 cases) [8] and the JSRT set (247 cases) [13]. Both the datasets are typical X-ray scans collected in clinical practice, but their image distributions are quite different in terms of the disease type, intensity, and contrast (see the first and fourth column in Fig. 2(a)). The ground truth masks of left and right lungs are provided in both datasets. We randomly split each dataset into 7:1:2 for training, validation and test sets. All the images were resized to 512×512, and rescaled to $[0, 255]$. The prediction masks were post-processed with the largest connected-component selection and hole filling.

For evaluation metrics, we utilized four common segmentation measurements, i.e., the Dice coefficient ([%]), recall ([%]), precision ([%]) and average surface distance (ASD)([mm]). The first three metrics are measured based on the pixel-wise classification accuracy. The ASD assesses the model performance at boundaries and a lower value indicates better segmentation performance.

Experimental Settings. We employed the Montgomery set as source domain and the JSRT set as target domain. We first established the segmenter on source training data independently. Next, we test the segmenter under various settings: (1) testing on source domain (*S-test*); (2) directly testing on target data (*T-noDA*); (3) using histogram matching to adjust target images before testing (*T-HistM*); (4) aligning target features with the source domain as proposed in [9] (*T-FeatDA*); (5) fine-tuning the model on labeled target data before testing on JSRT (*T-STL*); In addition, we investigated the performance of our proposed domain adaptation method with and w/o the semantic-aware loss, i.e., *SeUDA* and *CyUDA*.

Comparison of Experimental Results Between Different Methods. As shown in Table 1, when directly applying the learned source domain segmenter
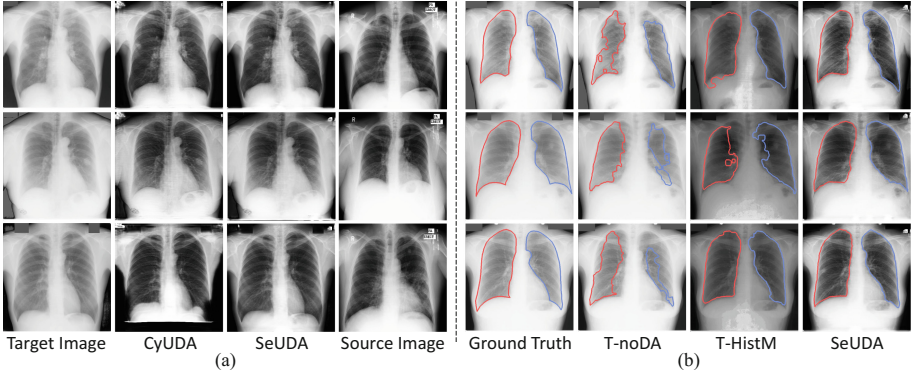
Fig. 2. Typical results for the image transformation and lung segmentation. (a) Visualization of image transformation results, from left to right, are the target images in JSRT set, *CyUDA* transformation results, *SeUDA* transformation results, and the nearest neighbor of $x^{t \to s}$ got from source set; each row corresponds to one patient. (b) Comparison of segmentation results between the ground truth, *T-noDA*, *T-HistM*, and our proposed *SeUDA*; each row corresponds to one patient.

to target data (*T-noDA*), the model performance significantly degraded, indicating that domain shift would severely impede the generalization performance of DNNs. Specifically, the average Dice over both lungs dropped from 95.61% to 79.47%, and the average ASD increased from 2.34 to 11.04 mm.

With our proposed *SeUDA*, remarkable improvements have been achieved by applying the source segmenter on transformed target images. Compared with *T-noDA*, our *SeUDA* increased the average Dice by 15.04%. Meanwhile, the ASDs for both lungs were reduced significantly. Also, our method outperforms the UDA baseline histogram matching *T-HistM* with the average dice increased by 3.97% and average ASD decreased from 5.19 mm to 3.18 mm. Compared with the feature-level domain adaptation method *T-FeatDA*, our *SeUDA* can not only obtain higher segmentation performance, but also provide intuitive visualization of how the adaptation is achieved. Notably, the performance of our unsupervised *SeUDA* is even comparable to the upper bound of supervised *T-STL*. In Table 1, the gaps of Dice are marginal, i.e., 1.32% for right lung and 2.42% for left lung.

In Fig. 2(a), we can visualize typical transformed target images, demonstrating that *SeUDA* has successfully adapted the appearance of target data to look more similar to source images. In addition, the positions, contents, semantic structures and clinical clues are well preserved after transformation. In Fig. 2(b) we can observe that without domain adaptation, the predicted lung masks are quite cluttered. With histogram matching, appreciable improvements are obtained but the transformed images cannot mimic the source images very well. With our *SeUDA*, the lung areas are accurately segmented attributing to the successful target-to-source appearance transformation.

Effectiveness of Semantic-aware Loss. We investigated the contribution of our novel semantic-aware loss designed for segmentation domain adaptation. We implemented *CyUDA* by removing the semantic-aware loss from the *SeUDA*. One notorious problem of GANs is that their training would be unstable and sensitive to initialization states [1,12]. In this study, we measured the standard deviation (std) of the *CyUDA* and *SeUDA* by running each model for 10 times under different initializations but with the same hyper-parameters. We observed significant lower variability on the segmentation performance across the 10 *SeUDA* models than the 10 *CyUDA* models, i.e., Dice std: 0.25% v.s. 2.03%, ASD std: 0.16 v.s. 1.19 mm. Qualitatively, we observe that the *CyUDA* transformed images may suffer from distorted lung boundaries in some cases, see the third row in Fig. 2(a). In contrast, adding the semantic-aware loss, the transformed images consistently present a high quality. This reveals that the novel semantic-aware loss contributes to stabilize the image transformation process and prevent the distortion in structural contents, and hence contributes to boost the performance of segmentation domain adaptation.

4 Conclusion

This paper proposes a novel approach *SeUDA* for unsupervised domain adaptation of medical image segmentation. The *SeUDA* leverages GANs to transform the target images to resemble source images and generalize the source segmentation DNN directly on the transformed images. We design a novel objective which composes a GAN loss for mapping data distributions, a cycle-consistency loss to preserve the pixel-level content, and a semantic-aware loss to enhance the structural information. Our method is highly competitive with the supervised transfer learning on the task of lung segmentation in chest X-rays. Our proposed *SeUDA* solution is general and can inspire more researches on domain adaptation problems in medical image computing.

References

1. Bousmalis, K., Silberman, N., Dohan, D., et al.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: CVPR. pp. 95–104 (2017)
2. Chartsias, A., Joyce, T., Dharmakumar, R., Tsaftaris, S.A.: Adversarial image synthesis for unpaired multi-modal cardiac data. In: Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (eds.) SASHIMI 2017. LNCS, vol. 10557, pp. 3–13. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68127-6_1
3. Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.: Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. In: IJCAI, pp. 691–697 (2018)

4. Dou, Q., et al.: Automated pulmonary nodule detection via 3d convnets with online sample filtering and hybrid-loss residual learning. In: MICCAI, pp. 630–638 (2017)

5. Ganin, Y., Ustinova, E., Ajakan, H.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096–2030 (2016)

6. Ghafoorian, M., et al.: Transfer learning for domain adaptation in MRI: application in brain lesion segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 516–524. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_59

7. Huo, Y., Xu, Z., Bao, S., et al.: Adversarial synthesis learning enables segmentation without target modality ground truth. arXiv preprint arXiv:1712.07695 (2017)

8. Jaeger, S.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quant. Imaging Med. Surg. **4**(6), 475 (2014)

9. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_47

10. Philipsen, R.H., Maduskar, P., Hogeweg, L., Melendez, J., Sánchez, C.I., van Ginneken, B.: Localized energy-based normalization of medical images: application to chest radiography. IEEE Trans. Med. Imaging **34**(9), 1965–1975 (2015)

11. 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

12. Salimans, T., Goodfellow, I., et al.: Improved techniques for training gans. Adv. Neural Inf. Process. Syst. 2234–2242 (2016)

13. Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am. J. Roentgenol. **174**(1), 71–74 (2000)

14. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR, pp. 2962–2971 (2017)

15. Wang, L.: Correction for variations in mri scanner sensitivity in brain studies with histogram matching. Magn. Reson. Med. **39**(2), 322–327 (1998)

16. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV, pp. 2242–2251 (2017)