Generalizing Deep Models for Ultrasound Image Segmentation

Xin Yang[1], Haoran Dou[2,3], Ran Li[2,3], Xu Wang[2,3], Cheng Bian[2,3],
Shengli Li[4], Dong Ni[2,3](✉), and Pheng-Ann Heng[1]

[1] Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong
[2] National-Regional Key Technology Engineering Laboratory for Medical
Ultrasound, School of Biomedical Engineering, Health Science Center,
Shenzhen University, Shenzhen, China
nidong@szu.edu.cn
[3] Medical UltraSound Image Computing (MUSIC) Lab, Shenzhen Maternal
and Child Healthcare Hospital of Nanfang Medical University, Shenzhen, China
[4] Department of Ultrasound, Shenzhen Maternal and Child Healthcare Hospital
of Nanfang Medical University, Shenzhen, China

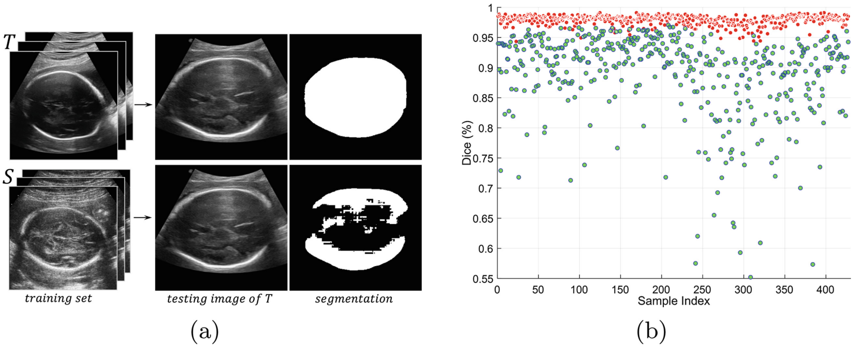
Abstract. Deep models are subject to performance drop when encountering appearance discrepancy, even on congeneric corpus in which objects share the similar structure but only differ slightly in appearance. This performance drop can be observed in automated ultrasound image segmentation. In this paper, we try to address this general problem with a novel online adversarial appearance conversion solution. Our contribution is three-fold. First, different from previous methods which utilize corpus-level training to model a fixed source-target appearance conversion in advance, we only need to model the source corpus and then we can efficiently convert each single testing image in the target corpus on-the-fly. Second, we propose a self-play training strategy to effectively pretrain all the adversarial modules in our framework to capture the appearance and structure distributions of source corpus. Third, we propose to explore a composite appearance and structure constraints distilled from the source corpus to stabilize the online adversarial appearance conversion, thus the pre-trained models can iteratively remove appearance discrepancy in the testing image in a weakly-supervised fashion. We demonstrate our method on segmenting congeneric prenatal ultrasound images. Based on the appearance conversion, we can generalize deep models at-hand well and achieve significant improvement in segmentation without re-training on massive, expensive new annotations.

1 Introduction

With massive annotated training data, deep networks have brought profound change to the medical image analysis field. However, retraining on newly annotated corpus is often compulsory before generalizing deep models to new imaging conditions [1]. Retraining is even required for congeneric corpora in which

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11073, pp. 497–505, 2018.
https://doi.org/10.1007/978-3-030-00937-3_57

objects share similar structures but only differ slightly in appearances. As shown in Fig. 1(a), there are two congeneric copora S and T, representing a similar anatomical structure, i.e. fetal head, with recognizable appearance difference, like intensity, speckle pattern and structure details. However, a deep model trained on S performs poor in segmenting images from T (Fig. 1).



(a) (b)

Fig. 1. Segmentation performance drop. (a) the model trained on T segments testing image in T well (red dots in (b)), while the model trained on S gets poor result in segmenting image in T (green dots in (b)). Better view in color version.

In practice, retraining is actually infeasible, because the data collection and expert annotation are expensive and sometimes unavailable. The situation becomes even worse when images are acquired at different sites, experts, protocols and even time points. Ultrasound is a typical imaging modality which suffers from these varying factors. Building a corpus for specific cases and retraining models for these diverse cases turn to be intractable. Unifying the image appearance across different imaging conditions to relive the burden of retraining is emerging as an attractive choice.

Recently, we witnessed many works on medical image appearance conversion. From a corpus level, Lei et al. proposed the convolutional network based low-dose to standard-dose PET translation [11]. With the surge of generative adversarial networks (GANs) [4] for medical image analysis [7], Wolterink et al. utilized GAN to reduce noise in CT images [10]. GAN also enables the realistic synthesis of ultrasound images from tissue labels [9]. Segmentation based shape consistency in cycled GAN was proposed in [5,13] to constrain the translation between CT and MR. Corpus-level conversion models can match the appearance distributions of different corpora from a global perspective. However, these models tend to be degraded on images which have never been modeled during training. From a single image level, style transfer [3] is another flexible and appealing scheme for appearance conversion between any two images. Whereas, it is subjective in choosing the texture level to represent the style of referring ultrasound image and preserve the structure of testing image. Leveraging the well-trained model in source corpus and avoiding the building of heavy target corpus, i.e. just using a single testing image, to realize structure-preserved appearance conversion is still a nontrivial task.

In this paper, we try to address this problem with a novel solution. Our contribution is three-fold. First, different from previous methods which model a corpus-level source-target appearance conversion in advance, our method works in an extreme case. The case is also the real routine clinic scenario where we are blinded to the complete target corpus and only a single testing image from target corpus is available. Our framework only needs to model the source corpus and then it can efficiently convert each testing image in target corpus on-the-fly. Second, under the absence of complete target corpus, we propose a self-play training strategy to effectively pre-train all adversarial modules in our framework to capture both the appearance and structure distributions of source corpus. Third, we propose to explore the mixed appearance and structure constraints distilled from the source corpus to guide and stabilize the online adversarial appearance conversion, thus the pre-trained models can iteratively remove appearance discrepancy in the testing image in a weakly-supervised fashion. We demonstrate the proposed method on segmenting congeneric prenatal ultrasound images. Extensive experiments prove that our method is fast and can generalize the deep models at-hand well, plus achieving significant improvement in segmentation without the re-training on massive, expensive new annotations.

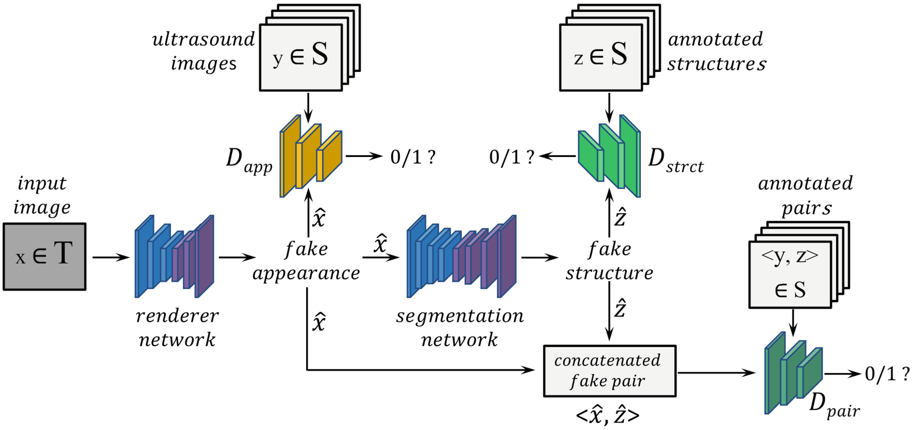
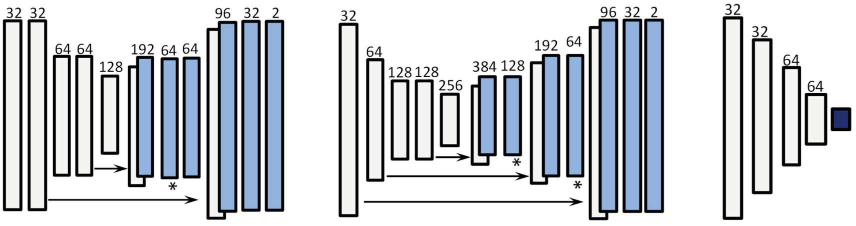


Fig. 2. Schematic view of our proposed framework.

2 Methodology

Figure 2 is the schematic view of our proposed adversarial framework for appearance conversion. System input is a single testing image from the blinded target corpus T. Renderer network renders the testing image and generates fake substitute with the appearance that can not be distinguished by appearance discriminator (D_{app}) from the appearance distribution of source corpus S.

Segmentation network then generates the fake structure on the fake appearance. Fake structure is also expected to fool the structure discriminator (D_{strct}) w.r.t the annotated structures in S. Structure here means shape. To enforce the appearance and structure coherence, the pair of fake appearance and structure is further checked by a pair discriminator (D_{pair}). During the adversarial training, the appearance of testing image and its segmentation will be iteratively fitted to the distributions of S. System outputs the final fake structure as segmentation.



(a) (b) (c)

Fig. 3. Architecture of the sub-networks in our framework. Star denotes the site to inject the auxiliary supervision. Arrow denotes skip connection for concatenation.

2.1 Architecture of Sub-networks

We adapt the renderer and segmentor network from U-net [8] featured with skip connections. Renderer (Fig. 3(a)) is designed to efficiently modify the image appearance, thus its architecture is light weighted with less convolutional and pooling layers compared with the segmentor (Fig. 3(b)). Auxiliary supervisions [2] are coupled with renderer and segmentor. Discriminators D_{app}, D_{strct} and D_{pair} share the same architecture design for fake/real classification (Fig. 3(c)), except that D_{pair} gets 2-channel input for the pairs. Definition of objective functions to tune parameters in these 5 sub-networks are elaborated below.

2.2 Objective Functions for Online Adversarial Rendering

Our system is firstly fully trained on the source corpus S to capture both appearance and structure distributions. Then the system iteratively renders a single testing image in corpus T with online updating. In this section, we introduce the diverse objectives we use during the full training and online updating.

Renderer Loss. With a renderer, our goal is to modulate the intensity represented appearance of ultrasound image x into $\hat{x}$ to fit the appearance in S. Severely destroying the content information in x is not expected. Therefore, there is an important *L1 distance* based objective for renderer to satisfy the content-preserved conversion (Eq. 1). α_i is the weight for auxiliary losses.

$$\mathcal{L}_{rend} = \sum_i \alpha_i \parallel x - \hat{x} \parallel_1, i = 0, 1. \tag{1}$$

Appearance Adversarial Loss. Renderer needs to preserve the content in x, but at the same time, it still needs to enable the fake $\hat{x}$ fool the appearance discriminator D_{app} which is trying to determine whether the input is from corpus S or T. Therefore, the adversarial loss for D_{app} is shown as Eq. 2.

$$\mathcal{L}_{D_{app}} = \mathbb{E}_{y \sim S}[\log D_{app}(y)] + 1 - \log(D_{app}(\hat{x})). \tag{2}$$

Segmentor Loss. Segmentor extracts fake structure $\hat{z}$ from $\hat{x}$. Built on limited receptive field, convolutional networks may lose power in boundary deficient areas, like acoustic shadow, in ultrasound images. Therefore, based on classic cross-entropy loss, we adapt the hybrid loss $\mathcal{L}_{seg}$ as proposed in [12] to get Dice coefficient based shape-wise supervision in order to combat boundary deficiency.

Structure Adversarial Loss. Renderer is trying to keep content of x while cheat the D_{app} by minimizing both Eqs. 1 and 2. However, the renderer may stick to x or, on the contrary, collapse on a average mode in S. Structure discriminator D_{strct} here is beneficial to alleviate the problem, since it requires that the structure $\hat{z}$ extracted from $\hat{x}$ must further fit the structure distribution of $z \in S$. The adversarial loss for D_{strct} is shown as Eq. 3.

$$\mathcal{L}_{D_{strct}} = \mathbb{E}_{z \sim S}[\log D_{strct}(z)] + 1 - \log(D_{strct}(\hat{z})). \tag{3}$$

Pair Adversarial Loss. Inspired by the conditional GAN [6], as illustrated in Fig. 2, we further inject a discriminator D_{pair} to determine whether the $\hat{x}$ and $\hat{z}$ in the $<\hat{x}, \hat{z}>$ pair can match each other. Pair adversarial loss for D_{pair} is shown in Eq. 4.

$$\mathcal{L}_{D_{pair}} = \mathbb{E}_{<x,z> \sim S}[\log D_{pair}(<x, z>)] + 1 - \log(D_{pair}(<\hat{x}, \hat{z}>)). \tag{4}$$

Our full objective function is therefore defined as:

$$\mathcal{L}_{full} = \mathcal{L}_{rend} + \mathcal{L}_{D_{app}} + \mathcal{L}_{seg} + \mathcal{L}_{D_{strct}} + \mathcal{L}_{D_{pair}}. \tag{5}$$

2.3 Optimization and Online Rendering

Self-play Full Training. With the images and labels in S, we can only train the segmentor for image-to-label mapping in a supervised way. How to train other adversarial networks without fake samples and further convey the distilled appearance and structure constraints of S to online testing phase? In this section, we propose a self-play scheme to train all sub-networks in a simple way.

Although all samples in S are supposed to share an appearance distribution, the intra-class variation still exists (Fig. 4). Our self-play training scheme roots in this observation. Before training, we can assume that every randomly selected sample from S has the same chance to be located far from the appearance distribution center of S. Thus, in each training epoch, we randomly take a sample from S as a fake sample and the rest as real samples to train our sub-networks. The result of this self-play training is that renderer can learn to convert all samples in S into a more concentrated corpus S' so that the objective $\mathcal{L}_{full}$ can be minimized. Also, segmentor can learn to extract structures from the resulted S'. D_{app}, D_{strct} and D_{pair} also capture the appearance and structure knowledge of S' for classification in online rendering stage. As shown in Fig. 4, with the self-play full training, ultrasound samples in S' present more coherent appearance and enhanced details than that in S. S' will replace S and be used as real samples to tune adversarial modules in the online testing phase.

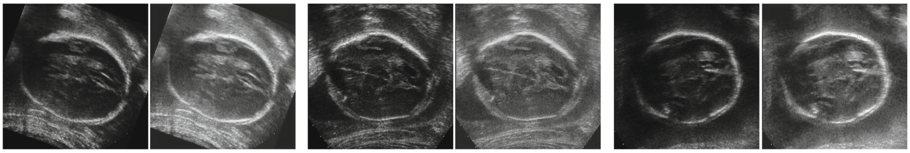


Fig. 4. Illustration of the self-play training based appearance unification on S. In each group, original image in S (left), intensity unified image in S' (right).

Online Rendering for a Single Image. In testing phase, we apply the pre-trained renderer to modify the appearance of testing image to fit S'. D_{app}, D_{strct} and D_{pair} try to distinguish the fake appearance, structure and pair from any randomly selected images or pairs in S' to ensure that the renderer generates reliable conversion. Testing phase is iterative and driven by the minimization of the objective $\mathcal{L}_{full}$ discarding the $\mathcal{L}_{seg}$. The optimization is fast and converges in few iterations. As depicted, our online appearance conversion is image-level, since we can only get a single image from the blinded target corpus. All the adversarial procedures are thus facing a 1-to-many conversion problem, which may cause harmful fluctuations during rendering. However, three designs of our framework alleviate the risk: (i) the composite constraints imposed by D_{app}, D_{strct} and D_{pair} from complementary perspectives, (ii) the loss $\mathcal{L}_{rend}$ restricts the appearance change within a limited range, (iii) S' provides exemplar samples with low intra-class appearance variation (Fig. 4), which is beneficial to smooth the gradient flow in rendering. Detailed ablation study is shown in Sect. 3.

3 Experimental Results

Materials and Implementation Details. We verify our solution on the task of prenatal ultrasound image segmentation. Ultrasound images of fetal head are collected from different ultrasound machines and compose two congeneric datasets. 1372 images acquired using a Siemens Acuson Sequoia 512 ultrasound scanner serve as corpus S with the gestational age from 24 w to 40 w. 1327 images acquired using a Sonoscope C1-6 ultrasound scanner serve as corpus T with the gestational age from 30 w to 34 w. In both S and T, we randomly take 900 images for training, the rest for testing. S and T are collected by different experts and present distinctive image appearance. Experienced experts provide boundary annotations for S and T as ground truth. To avoid unrelated factors to image appearance, like scale and translation, we cropped all images to center around the fetal head region and resize them to the size as 320 × 320. Segmentation model trained on S drops severely on T, as Fig. 1 and Table 1 show.

We implement the whole framework in *Tensorflow*, using a standard PC with an NVIDIA TITAN Xp GPU. *Code is online available*[1]. In full training, we update the weights of all sub-networks with an Adam optimizer (batch size = 2, initial learning rate is 0.001, momentum term is 0.5, total iteration = 6000). During the online rendering, we update the weights of all sub-networks with smaller initial learning rate 0.0001. Renderer and segmentor are updated twice as often as the discriminators. We only need less than 25 iterations (about 10 s for each iteration) before achieving a satisfying and stable online rendering.

Table 1. Quantitative evaluation of our proposed framework

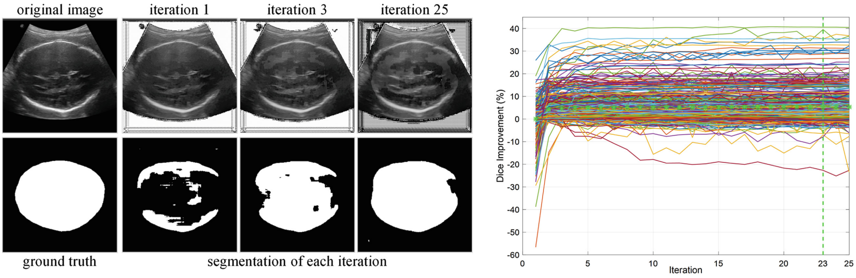
Method	Metrics						
	Dice [%]	Conf [%]	Adb [pixel]	Hdb [pixel]	Jaccard [%]	Precision [%]	Recall [%]
Orig-T2T	97.848	95.575	3.7775	25.419	95.799	96.606	99.148
Orig-S2T	88.979	73.493	21.084	73.993	80.801	94.486	84.737
S2T-sp	92.736	84.075	13.917	62.782	86.688	94.971	91.267
S2T-p	93.296	85.127	12.757	58.352	87.619	95.115	**93.262**
S2T	**93.379**	**85.130**	**11.160**	**53.674**	**87.886**	**95.218**	92.633

Quantitative and Qualitative Analysis. We adopt 7 metrics to evaluate the proposed framework on segmenting ultrasound images from T, including Dice coefficient (DSC), Conformity (Conf), Hausdorff Distance of Boundaries (Hdb), Average Distance of Boundaries (Adb), Precision and Recall. We firstly trained two segmentors on the training set of corpus T (Orig-T2T) and corpus S (Orig-S2T) respectively with same settings, and then test them on T. From Table 1, we can see that, compared with Orig-T2T, the deep model Orig-S2T is severely degraded (about 10% in Dice) when testing images from T. As we upgrade the Orig-S2T with the proposed online rendering (denoted as S2T), we achieve a significant improvement (4% in DSC) in the segmentation. This proves the efficacy of our renderer in converting the congeneric ultrasound images to the appearance which can be well-handled by the segmentor.

Ablation study is conducted to verify the effectiveness of D_{strct} and D_{pair}. We remove the D_{pair} in S2T to form the S2T-p, and further remove the D_{strct} in S2T-p to form the S2T-sp. As we can observe in Table 1, without the constraints imposed by D_{strct} and D_{pair}, S2T-sp becomes weak in appearance conversion. Compared to S2T-sp, S2T-p is better in appearance conversion, thus D_{strct} takes more important role than D_{pair} in regularizing the conversion. With Fig. 5(a), we show the intermediate results of the online rendering. As the renderer modulates the appearance of input ultrasound image, the segmentation result is also

[1] https://github.com/xy0806/congeneric_renderer.

gradually improved. Figure 5(b) illustrates the Dice improvement curve along with iteration for all the 427 testing images in T. Almost all the rendering come to convergence around 5 iterations (about 50 s in total). The highest averaged Dice improvement (5.378%) is achieved at iteration 23.



(a) (b)

Fig. 5. (a) Intermediate rendering and segmentation result. (b) Dice improvement over iteration 0 for all the 427 testing images in T. Green star is average at each iteration.

4 Conclusions

We present a novel online adversarial appearance rendering framework to fit the input image appearance to the well-modeled distribution of source corpus, and therefore relieve the burden of retraining for deep networks when encountering congeneric images with unseen appearance. Our framework is flexible and renders the testing image on-the-fly, which is more suitable for routine clinic applications. The proposed self-play based full training scheme and the composite adversarial modules prove to be beneficial in realizing the weakly-supervised appearance conversion. Our framework is novel, fast and can be considered as an alternative in more tough tasks, like cross-modality translation.

Acknowledgments. The work in this paper was supported by the grant from National Natural Science Foundation of China under Grant 81270707, and the grants from the Research Grants Council of the Hong Kong Special Administrative Region (Project Nos. GRF 14202514 and GRF 14203115).

References

1. Chen, H., Ni, D., et al.: Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE JBHI **19**(5), 1627–1636 (2015)
2. Dou, Q., Yu, L., et al.: 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. **41**, 40–54 (2017)
3. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR, pp. 2414–2423. IEEE (2016)
4. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)

5. Huo, Y., Xu, Z., et al.: Adversarial synthesis learning enables segmentation without target modality ground truth. arXiv preprint arXiv:1712.07695 (2017)
6. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint (2017)
7. Nie, D., et al.: Medical image synthesis with context-aware generative adversarial networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 417–425. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_48
8. 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
9. Tom, F., Sheet, D.: Simulating patho-realistic ultrasound images using deep generative networks with adversarial learning. arXiv preprint arXiv:1712.07881 (2017)
10. Wolterink, J.M., et al.: Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans. Med. Imaging **36**(12), 2536–2545 (2017)
11. Xiang, L., et al.: Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI. Neurocomputing **267**, 406–416 (2017)
12. Yang, X., Bian, C., Yu, L., Ni, D., Heng, P.A.: Hybrid loss guided convolutional networks for whole heart parsing. In: Pop, M., et al. (eds.) STACOM 2017. LNCS, vol. 10663, pp. 215–223. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75541-0_23
13. Zhang, Z., Yang, L., Zheng, Y.: Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. arXiv preprint arXiv:1802.09655 (2018)