3D U^2-Net: A 3D Universal U-Net for Multi-domain Medical Image Segmentation

Chao Huang[1,2], Hu Han[2,3], Qingsong Yao[2], Shankuan Zhu[1(✉)], and S. Kevin Zhou[2,3(✉)]

[1] Chronic Disease Research Institute and Department of Nutrition and Food Hygiene, School of Public Health, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China
{huangchao09,zsk}@zju.edu.cn

[2] Medical Imaging, Robotics, Analytic Computing Laboratory/Engineering (MIRACLE), Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
{hanhu,zhoushaohua}@ict.ac.cn

[3] Peng Cheng Laboratory, Shenzhen, China

Abstract. Fully convolutional neural networks like U-Net have been the state-of-the-art methods in medical image segmentation. Practically, a network is highly specialized and trained separately for each segmentation task. Instead of a collection of multiple models, it is highly desirable to learn a universal data representation for different tasks, ideally a single model with the addition of a minimal number of parameters steered to each task. Inspired by the recent success of multi-domain learning in image classification, for the first time we explore a promising universal architecture that handles multiple medical segmentation tasks and is extendable for new tasks, regardless of different organs and imaging modalities. Our 3D Universal U-Net (3D U^2-Net) is built upon separable convolution, assuming that *images from different domains have domain-specific spatial correlations which can be probed with channel-wise convolution while also share cross-channel correlations which can be modeled with pointwise convolution*. We evaluate the 3D U^2-Net on five organ segmentation datasets. Experimental results show that this universal network is capable of competing with traditional models in terms of segmentation accuracy, while requiring only about 1% of the parameters. Additionally, we observe that the architecture can be easily and effectively adapted to a new domain without sacrificing performance in the domains used to learn the shared parameterization of the universal network. We put the code of 3D U^2-Net into public domain (https://github.com/huangmozhilv/u2net_torch/).

C. Huang and S. Zhu were supported by Cyrus Tang Foundation & Zhejiang University Education Foundation. H. Han was supported by the Natural Science Foundation of China (61732004 and 61672496), External Cooperation Program of CAS (GJHZ1843), and Youth Innovation Promotion Association CAS (2018135). This work was done when C. Huang was an intern in MIRACLE.

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11765, pp. 291–299, 2019.
https://doi.org/10.1007/978-3-030-32245-8_33

Keywords: Universal model · Multi-domain learning · Segmentation

1 Introduction

Image segmentation is crucial for clinical practice and health research. Fully convolutional neural networks (CNNs) like U-Net [15] have been the dominant approach in automatic medical imaging segmentation [4,11]. A practical segmentation model is learned by customizing a neural network architecture for a certain task or dataset and training it from scratch [11,16,18]. [7] learned a single segmentation CNN for brain datasets acquired with different scanners and/or protocols. Notwithstanding being powerful, these models are difficult to extend to new tasks with unseen contents because of the highly specialized design. [6] took one step further by presenting a self-adapting framework for various tasks, yielding mutually independent models for each task. On the contrary, human experts can easily learn to tackle multiple tasks and generalize to new tasks on the basis of acquired skills. Multiple previous works explored multi-task segmentation, wherein all organs of interest appear in the same image [9,17]. Here we consider a more realistic and challenging scenario: for a given dataset, only a local region of the human body is scanned and only one or several anatomical structures within the image are annotated. [12] focused on a similar topic and trained one single CNN on three tasks, however, the trained model was designed as such that it cannot be extended to other tasks. From this point of view, an effective and efficient method for image segmentation remains an open problem.

Bilen et al. [2,13,14] suggested that there might exist a universal data representation across different visual domains. Specifically, they introduced a new competition called Visual Decathlon Challenge[1], aiming to simultaneously model ten visual domains of different styles and contents, e.g., internet images, handwritten characters, sketches, planktons, etc. [13]. They referred to such a new topic as "multi-domain learning" and realized the universal representation by piggybacking parallel residual adapters on the model pre-trained with ImageNet. However, their work exclusively focuses on image classification. Naturally, one question occurs to us: *is it possible to build a single neural network that can deal with medical segmentation tasks from different domains?*

To achieve this goal, we draw inspiration from previous studies [3,5], particularly [5] which won the first place in the Visual Decathlon Challenge to date. [5] believed that [14] ignored the structural heterogeneity of various domains and attempted to address the issue by leveraging depthwise separable convolution. While standard convolution conducts the spatial and channel-wise computation at once, such convolution factors the computation into two sequential steps: first, depthwise convolution applies an independent convolutional filter per input channel, and then a pointwise convolution follows to linearly combine the output across all channels for every spatial location. The basic building block of their multi-domain network comprises a cohort of parallel channel-wise convolutions,

[1] https://www.robots.ox.ac.uk/~vgg/decathlon/.

one per domain, followed by one pointwise convolution shared by all domains. The insight is that the former is better to capture domain-specific spatial patterns while the latter probes the sharable cross-channel interdependencies. In this paper, we claim to note "depthwise separable convolution" as "separable convolution" and "depthwise convolution" as "channel-wise convolution" to avoid confusion with the depth dimension of the image volume.

Based on the separable convolution as introduced above, our work proposes a universal architecture for multi-domain medical image segmentation. The main idea behind is rather intuitive yet powerful: a basic network is first designed on the ground of 3D U-Net [4,15] (or V-Net [11]), and then any $3 \times 3 \times 3$ standard convolution with a stride of 1 is substituted by separable convolution similar to [5]. However, our approach substantially differ from [5] as following: (1) their work focuses on image classification which is fundamentally different from image segmentation here. (2) they obtain the ultimate multi-domain architecture in three steps: First, pre-training a ResNet-26 modified with separable convolution on ImageNet; Second, freezing and transferring the pointwise convolution weights to new network; Thirdly, training the new network on each domain separately and stacking the channel-wise convolutions together while sharing the pointwise convolution weights from the pre-trained model. Nevertheless, we manage to train across the domains together to obtain the final model. (3) we further adapt our universal network to a new domain by simply adding new channel-wise convolutions. To the best of our knowledge, this is the first time to learn an extendable universal network for multi-domain medical image segmentation.

2 Methods

2.1 Problem Definition

Let $\{D_1, D_2, \cdots, D_T\}$ be a set of T image domains, among which domain D_t consists of two paired image spaces of $\{X_t, Y_t\}$. $X_t \in \mathbb{R}^{C_t \times D \times H \times W}$ is the input image space and $Y_t \in \mathbb{R}^{C_t' \times D \times H \times W}$ is the output image space, i.e., segmentation masks. D, H and W are the spatial depth, height and width. C_t and C_t' are the numbers of imaging modalities and segmentation classes specific to each domain. To work well on all domains, our universal network contains domain-specific parameters as well as shared parameters. Let θ_t be the domain-specific parameters for domain D_t and θ_u be the universally shared parameters by all domains. Assuming $\{x_{t,i}, y_{t,i}\}$ as the i^{th} training pair of domain D_t, then the output $\hat{Y}$ of the neural network $F(X)$ is

$$\hat{y}_{t,i} = F(x_{t,i}; \theta_u, \theta_t). \tag{1}$$

2.2 Domain Adapter

Domain adapter, the key component to ensure the success of our universal network, consists of both domain-specific parameters and shared parameters and is built upon separable convolution in place of standard convolution.

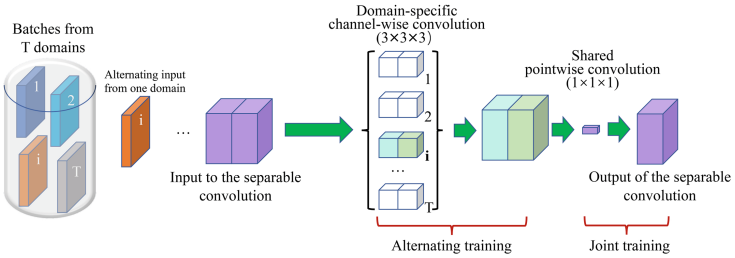

Fig. 1. Domain adapter based on separable convolution.

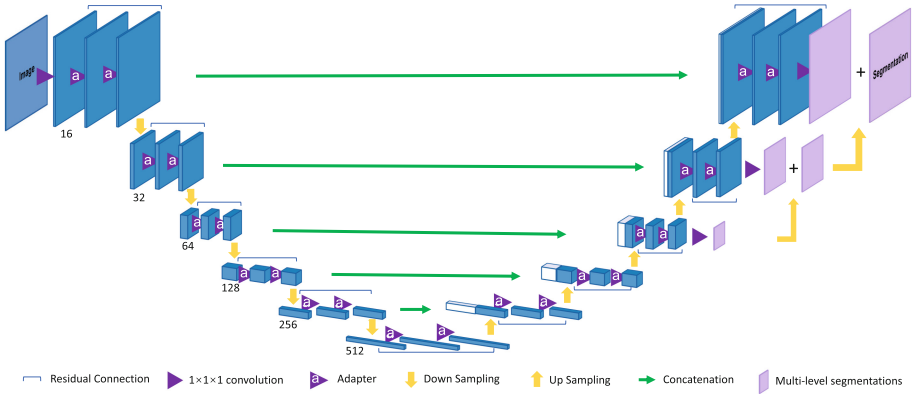

Fig. 2. The proposed 3D Universal U-Net (3D U^2-Net).

In standard convolution with filter $W \in \mathbb{R}^{3\times3\times3\times C\times C'}$ applied to an input tensor $U \in \mathbb{R}^{C\times D\times H\times W}$, the output tensor $\hat{U} \in \mathbb{R}^{C'\times D\times H\times W}$ is obtained by applying C' filters $w \in \mathbb{R}^{3\times3\times3\times C}$ on the input in parallel and concatenating the C' output feature maps. A simple calculation tells that the total number of filter parameters in the above filters is $27 * C * C'$. Also, when training the models for the T domains separately, the number of parameters grows T times!

In separable convolution, the computation is factorized into two sequential steps. The first step applies C channel-wise filters $w \in \mathbb{R}^{3\times3\times3}$ to each channel of the input in parallel and concatenate the C output feature maps together. Here, *each domain has its own channel-wise filters*. The second step then applies C' pointwise filters $w \in \mathbb{R}^{1\times1\times1\times C}$ to output the final feature maps of C' channels. Here, *all domains share the same pointwise filters*. A simple calculation tells that the total number of weights in the above filters is $27 * C * T + C * C'$. How to assemble the domain-specific channel-wise convolutions and the shared pointwise convolution to form a domain adapter is illustrated in Fig. 1.

2.3 3D Universal U-Net (3D U^2-Net)

As shown in Fig. 2, our universal network architecture is based on a basic network with six components: (1) input; (2) encoder path; (3) bottleneck block;

(4) decoder path; (5) deep supervision branch; and (6) output. Channels of the input and output could vary according to the number of imaging modalities and classes of different domains. In general, the input layer uses 16 filters. The encoder and decoder paths both contain five levels at different resolutions. Residual connection is applied within each level. Skip connection is employed to preserve more contextual information from the encoder counterpart for decoder path [15]. Inspired by [8], we incorporate a deep supervision branch alongside the end of decoder path via element-wise sum of multi-level segmentation maps to boost the final localization performance. To construct the universal network, domain adapters detailed above are inserted into basic network to replace any standard $3 \times 3 \times 3$ convolution with a stride of 1.

2.4 Loss Function

A hybrid loss function is employed by combining Lovász-Softmax loss [1], capable of improving intersection-over-union segmentation scores, and focal loss [10], aimed to alleviate class imbalance. During training the universal model, we sample a batch from each dataset in a round-robin fashion, allowing each domain to contribute to the shared parameters. Assuming that for the nth iteration the batch data pair $\{x_t, y_t\}$ is from domain D_t, the corresponding loss L_n is

$$L_n = L_L(x_t, y_t; \theta_u, \theta_t) + L_f(x_t, y_t; \theta_u, \theta_t), \tag{2}$$

where θ_t be the domain-specific parameters for domain D_t and θ_u be the universally shared parameters of the neural network. L_L is the Lovász-Softmax loss and L_f is the focal loss counterpart.

3 Experimental Results

In this section, we present extensive experiments to evaluate the proposed 3D U^2-Net in dealing with medical multi-organ segmentation: (1) independent models, aimed to reproduce the traditional methods, are obtained by training the basic network for each base domain separately; (2) shared model, which aims at investigating whether all parameters of a model can be shared by all domains and thus is gained by training the single basic network with all base domains together; and (3) universal model, which is our ultimate goal and is achieved by training the universal architecture with all base domains simultaneously. Notably, the first two represent two extreme multi-organ segmentation approaches and serve as baselines for the universal model. Additionally, we test the generalizability of both the shared model and universal model on one new domain.

Datasets: We use six public datasets from the Medical Segmentation Decathlon challenge[2] as introduced by [19]. The first five datasets are considered as base

[2] https://decathlon.grand-challenge.org/.

Table 1. Basic characteristics of the datasets.

Task	Modality	Data size	Image shape	Voxel spacing
Base01_Heart	MRI	20	(90∼130) × 320 × 320	1.37 × 1.25 × 1.25
Base02_Liver	CT	131	(74∼987) × 512 × 512	(0.7∼5) × (0.557∼1) × (0.557∼1)
Base03_Hippocampus	MRI	260	(24∼47) × (40∼59) × (31∼43)	1 × 1 × 1
Base04_Prostate	T2, ADC	32	(11∼24) × (256∼384) × (256∼384)	(3∼4) × (0.6∼0.75) × (0.6∼0.75)
Base05_Pancreas	CT	281	(37∼751) × 512 × 512	(0.7∼7.5) × (0.605∼0.977) × (0.605∼0.977)
New_Spleen	CT	41	(31∼168) × 512 × 512	(1.25∼7.5) × (0.535∼0.977) × (0.535∼0.977)

domains and are used to train the universal model. On the other hand, the last dataset is treated as the new domain and is used to test the adaptiveness of the universal model. Basic characteristics of the datasets are shown (Table 1). For each dataset, 80% of the samples are randomly extracted for training, while the remaining 20% are used as testing data.

Preprocessing: The datasets are highly diverse in terms of modality, image size and voxel spacing. Pre-processing procedures are conducted as below: (1) all images are cropped to the region of nonzero values, thereby reducing the image size to alleviate computation burden; (2) all images are resampled to the median voxel spacing of the corresponding dataset to retain spatial semantics; (3) for each patient, the image is clipped to the [2.0, 98.0] percentiles of the intensity values of the entire image, followed by Z-score normalization with the mean and standard deviation of the image for each modality; and (4) the following data augmentation are applied: random elastic deformation, random rotation, random scaling and random mirroring. Data augmentation is done "on-the-fly" during training with batch generators[3], a python package maintained by the Division of Medical Image Computing at the German Cancer Research Center.

To accommodate the limited GPU memory, we train the network with patches randomly sampled from the whole images. While for inference, the patches are generated with a sliding window moving across the entire image with a stride of half patch size. As for the shared model and universal model, the input batch is of two patches with a size of $128 \times 128 \times 128$ and the number of down-sampling operations is set to 6. However, for the independent models, we adjust the input patch size and the resolution levels for each domain considering the image size in order to maximize the utilization of computation resources. If the median shape is smaller than $128 \times 128 \times 128$, we toggle between the input patch size and batch size to have the patch size of the same aspect ratio as the

[3] https://github.com/MIC-DKFZ/batchgenerators/.

Table 2. Quantitative results on base domains.

	Base01_ Heart	Base02_ Liver	Base03_ Hippocampus		Base04_ Prostate		Base05_ Pancreas	
(Dice%)	Left_atrium	Liver	Anterior	Posterior	PZ	TZ	Pancreas	Mean
Independent	93.26	95.02	89.62	87.74	58.39	87.18	78.78	84.28
Shared	92.73	93.40	89.25	87.30	68.38	89.30	57.57	82.56
Universal	91.98	93.54	89.34	87.05	68.50	89.21	62.08	83.10

median shape. The number of down-sampling operations per axis is set until the feature map size of the deepest layer reaches as small as 8. Specifically, to prepare the patches for shared model and universal model, we first extract a patch of size as in the independent model and then resize it to the above target patch size.

Implementation Details: The network is implemented in Pytorch 1.0.1 on an NVIDIA V100 GPU. The ADAM optimizer is applied with an initial learning rate of 3×10^{-4} and a weight decay of 10^{-5}. An epoch is defined as an iteration over 250 batches. Exponential moving average, l^t_{MA}, is monitored for training loss for every 30 epochs. The learning rate is reduced by a factor of 5 as long as l^t_{MA} does not decrease by 5×10^{-4}. We terminate the training once the learning rate is below 10^{-8}. During training the shared and universal models, we apply a round-robin fashion to feed the network sample batches from each domain in turn, so as to allow all the domains to contribute to the final model equally. The results are presented on the testing data.

Quantitative Results of Base Domains: Table 2 lists the mean Dice scores of the three models on each base domain. Comparing along the columns, we observe that the independent models obtain the highest scores on most domains and yield the highest overall mean score. However, strikingly both the shared model and the universal model achieve moderate performance for most domains comparable to the independent models, and gains significant increase regarding to peripheral zone (PZ) and transition zone (TZ) of Base04_Prostate. Compared to the shared model, we further observe that the universal model is better in the segmentation of pancreas for Base05_Pancreas. Besides, the universal model gets an overall higher mean score across all domains in comparison to the shared model. The increase in overall performance could be attributed to the use of domain-specific parameters that can agree with each domain well.

Model Complexity: When investigating the complexity of the models, we exclude the input layer, last layer and deep supervision branch as they are never shared across domains. The basic network used in the shared model is considered as reference. The number of parameters are computed and displayed in Table 3(a). Obviously the proposed 3D U^2-Net requires the least parameters, indicating that it can perform effectively across various domains. The overall

Table 3. (a) Model complexity. (b) Quantitative results on a new spleen domain.

	(a) #Par	(a) Ratio	(b) New_Spleen – Dice%	(b) #Added Par
Independent	126.7M	4.1×	92.37	30.7M
Shared	30.7M	1×	90.67	0
Universal	1.7M	0.06×	91.60	0.1M

number of parameters from the universal model is around **1%** of that of all independent models, while the two obtain comparable segmentation accuracy.

Quantitative Results of a New Domain: Furthermore, we conduct experiments to illustrate the effectiveness of adapting the trained shared model or universal model to a new task, which are implemented by freezing the corresponding shared pointwise convolutions or standard convolutions and adding and training all other domain-specific modules like input layer and channel-wise convolutions in parallel to the structures of the same kindred for this domain. Table 3(b) shows that the universal model performs better for the new domain 'New_Spleen' in comparison to the shared model, therefore indicating a superior generalization ability over the latter. This adds further evidence of the effectiveness of the domain-specific parameters. The universal model is adaptive to new domain with a few extra parameters, i.e., 0.3% compared to the traditional independent model, which is exactly what we anticipate in this paper.

4 Conclusions

In summary, we present a novel universal neural network named 3D U^2-Net for multi-organ segmentation problem, filling the gap of extendable multi-domain learning in image segmentation. Experimental results demonstrate that the proposed approach, with only a tiny portion of the parameters, obtains the segmentation performance comparable to the independent models trained in the traditional manner. As CT and MRI images are routine images on hand and the amount of human organs is constant, the universal model for multi-organ segmentation can be fully developed soon in the near future. Besides, the proposed framework could extend to many other multi-domain applications and thus facilitate the translation of neural networks to clinical practice.

References

1. Berman, M., Rannen Triki, A., Blaschko, M.B.: The lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: Proceedings of CVPR, pp. 4413–4421 (2018)
2. Bilen, H., Vedaldi, A.: Universal representations: the missing link between faces, text, planktons, and cat breeds. arXiv:1701.07275 (2017)

3. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of CVPR, pp. 1251–1258 (2017)
4. 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, M.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
5. Guo, Y., Li, Y., Feris, R., Wang, L., Rosing, T.: Depthwise convolution is all you need for learning multiple visual domains. arXiv:1902.00927 (2019)
6. Isensee, F., et al.: nnU-Net: self-adapting framework for u-net-based medical image segmentation. arXiv:1809.10486 (2018)
7. Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E.: A lifelong learning approach to brain MR segmentation across scanners and protocols. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 476–484. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_54
8. Kayalibay, B., Jensen, G., van der Smagt, P.: CNN-based segmentation of medical imaging data. arXiv:1701.03056 (2017)
9. Lay, N., Birkbeck, N., Zhang, J., Zhou, S.K.: Rapid multi-organ segmentation using context integration and discriminative models. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 450–462. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38868-2_38
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of ICCV, pp. 2980–2988 (2017)
11. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of 3DV, pp. 565–571 (2016)
12. Moeskops, P., et al.: Deep learning for multi-task medical image segmentation in multiple modalities. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 478–486. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_55
13. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: Proceedings of NIPS, pp. 506–516 (2017)
14. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Efficient parametrization of multi-domain deep neural networks. In: Proceedings of CVPR, pp. 8119–8127 (2018)
15. 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
16. Roth, H.R., et al.: DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 556–564. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_68
17. Roth, H.R., et al.: Hierarchical 3D fully convolutional networks for multi-organ segmentation. arXiv:1704.06382 (2017)
18. Savioli, N., Montana, G., Lamata, P.: V-FCNN: volumetric fully convolution neural network for automatic atrial segmentation. arXiv:1808.01944 (2018)
19. Simpson, A.L., Antonelli, M., Bakas, S., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv:1902.09063 (2019)