NETWORK-AGNOSTIC KNOWLEDGE TRANSFER FROM LATENT DATASET FOR MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transfer learning often employs all or part of the weights of a pre-trained network to the problem at hand; this limits the flexibility of new neural architectures. We propose to transfer the knowledge of a neural network (teacher) from a latent dataset to another neural network (student) by training the student on a dataset agent whose annotations are generated by the teacher. The dataset agent requires no manual annotation and is independent of the teacher-training dataset. The student does not need to inherit the weights of the teacher, and such, the proposed algorithm can be flexibly conducted between heterogeneous neural architectures. Extensive experiments on six multi-organ medical image segmentation datasets including public and private in-house have shown that the proposed algorithm was effective for knowledge transfer and easy to be used with fine-tuning. This algorithm has the potential to be employed in novel applications where the teacher-training dataset is not accessible, particularly in medical applications.

1 INTRODUCTION

Deep learning has been extensively studied and has been shown to be as good as or even better than humans in certain areas (Ranjan et al., 2018; Geirhos et al., 2017; Lu & Tang, 2014). In order to achieve superior performance, deep learning architectures often require a sufficiently large training dataset. These training sets are expensive to build and are not easily shared between users. For example, one of the big challenges with semantic medical image segmentation is that the annotated data is quite limited (Litjens et al., 2017). Due to ethical concerns, the medical datasets are not often released with the trained networks. This necessitates the study of knowledge transfer between neural networks, wherein the original training dataset need not be accessed. Since deep learning-based methods operate in a black- box fashion, it becomes increasingly difficult to transfer knowledge between heterogeneous neural networks. To solve this problem, algorithms have been proposed by reusing or sharing the knowledge of neural networks, such as transfer learning (Tan et al., 2018), federated learning (Yang et al., 2019), and self-training (Xie et al., 2020b).

Transfer learning is used to leverage the knowledge that is obtained by conducting a specific task to solve different problems. Some works directly transfer the weights of standard large models, trained on natural image datasets, to solve the problem of medical image segmentation (Kang & Gwak, 2019; Motamed et al., 2019; Jodeiri et al., 2019). For example, Iglovikov and Shvets (Iglovikov & Shvets, 2018) proposed the use of VGG11, which was pre-trained on ImageNet, as the encoder of U-Net for 2D image segmentation. Similarly, convolutional 3D (Tran et al., 2015), trained on natural video datasets, was utilized as the pre-trained encoder of 3D U-Net for the 3D MR image segmentation (Zeng et al., 2017). In contrast to previous studies, Raghu et al. (2019) evaluated the performance of transfer learning by standard architectures for natural images (ImageNet) on medical imaging tasks and found transfer learning did not significantly improve the performance.

There are two challenges that are hard to reconcile for transfer learning. First, transfer learning methods aim to reuse all or part of the pretrained weights; this requires the architecture of the new model to be compatible with that of the pretrained network. The flexibility of new network is therefore limited. Second, the dataset used to tune the new model has to be similar to the dataset

previously used for training, otherwise, transfer learning may fail to work. New tasks are not always congruent with old tasks which often results in disparate datasets.

Another technique that relies on knowledge transfer is federated learning (Yang et al., 2019). Federated learning has received great attention for its capability to train a large-scale model in a decentralized manner without needing to access users' data, which is often needed to protect the users' private data. This kind of method can be used to tackle non-IID data from heterogeneous sources, such as those on smartphones and healthcare institutes. Usually, federated learning approaches adopt the central model to capture the shared knowledge of all users by aggregating their gradients. Due to the difficulty to transfer knowledge between heterogeneous networks, federated learning often requires that all devices, including central servers and local users, use the same neural architecture (Xie et al., 2020a). To our best knowledge, there has been no federated learning system that uses heterogeneous networks. In fact, the server often has more computation ability and necessity to employ more capable neural networks.

Different from transfer learning and federated learning, self-training has the ability to transfer knowledge from one neural network to another. The self-training framework (Scudder, 1965) has three main steps: train a teacher model on labeled images; use the teacher to generate pseudo labels on unlabeled images; and train a student model on the combination of labeled images and pseudo labeled images. Xie et al. (2020b) proposed self-training with the noisy student for the classification task, which iterates this process a few times by treating the student as a teacher to relabel the unlabeled data and training a new student. These studies required labeled training images for the student; moreover, they implicitly assumed that the pseudo-labeled images would be similar in content to the original labeled images.

Inspired by self-training, we propose a network-agnostic knowledge transfer algorithm for medical image segmentation. This algorithm transfers the knowledge of a neural network(teacher) from a latent dataset to another neural network (student) by training the student on a dataset agent whose annotations are generated by the teacher. The dataset agent (e.g. thyroid nodule images) is independent of the teacher-training dataset (e.g. breast lesion dataset), and it requires no manual annotation. Moreover, the student doesn't need to reuse the weights of the teacher, so that the proposed method can be flexibly conducted between heterogeneous neural architectures. Experiments showed that the proposed method performed well on medical image segmentation. This method has the potential to be employed to train novel models when the teacher-training dataset is not accessible by the student model, which is common in medical applications and federated learning.

We conducted extensive experiments to test the proposed algorithm, using six datasets for semantic segmentation and five state-of-the-art deep neural networks. Out of the six datasets, three public datasets involve breast lesion, nerve structure, and skin lesion, and three XXX datasets involve breast lesion (for two datasets) and thyroid nodule. The deep neural networks include Deeplab v3+ (Chen et al., 2018), U-Net (Ronneberger et al., 2015), Attention U-Net (Oktay et al., 2018), SDU-Net (Wang et al., 2020), and Panoptic Segmentation (Kirillov et al., 2019).

2 ALGORITHM

As Figure 1 shows, the proposed algorithm consists of two main steps for knowledge transfer and one optional step for downstream tasks:

Step 1. Teacher model training: Train Model-A (teacher model) on Dataset-1 (teacher-training dataset) with ground truth, i.e., manual annotations.

Step 2. Student model training: Generate pseudo annotations for the images in Dataset-2 using trained Model-A; and then train Model-B (student model) on Dataset-2 (dataset agent) with the pseudo annotations;

Step 3. Student model tuning (Optional): Fine-tune the trained Model-B for a downstream task on Dataset-3 (fine-tuning dataset) with manual annotations.

In this algorithm, Model-A and Model-B are independent, and Dataset-1, Dataset-2, and Dataset-3 are also independent. For example, Dataset-1 is a dataset for breast lesion segmentation while Dataset-2 can be a thyroid nodule dataset. So these three steps can be conducted by different parties without accessing the dataset of each other.

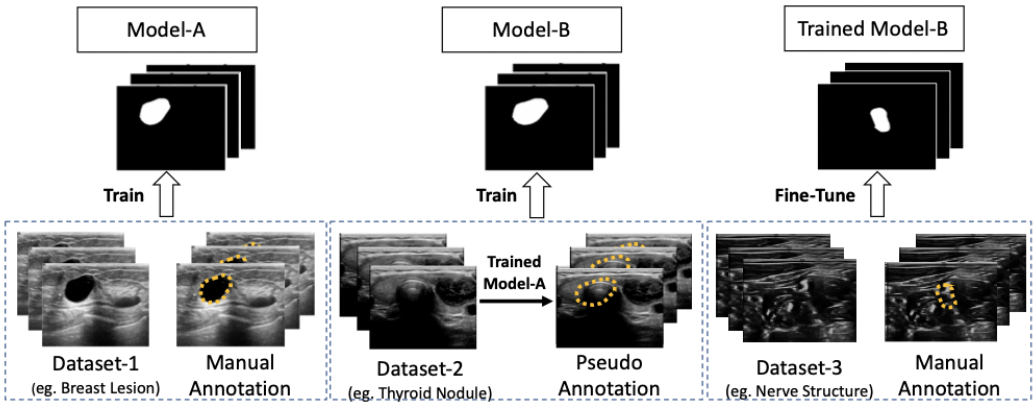

Figure 1: Diagram of the proposed algorithm. Model-A is a teacher model and Model-B is a student model. Dataset-1 is the teacher-training dataset. Dataset-2, whose annotations are generated by the teacher model, is used to transfer the teacher knowledge of dataset-1 to the student model. Dataset-3 is used to fine-tune the trained Model-B for the downstream task.

It is worth to mention that the dataset used to train Model-B, i.e., dataset agent, does not need any manual annotations. Therefore, it is often cheap and easy to build a dataset agent. In step 2, the student model gains the knowledge of the teacher by the dataset agent, ensuring that the student model has similar performance as the teacher model. **Note that the student model is trained to work well on datasets similar to the teacher-training dataset, while it is not aimed to predict the ground truth of the dataset agent.**

The downstream tasks in step 3 can be independent of the knowledge transfer between the teacher and the student. In the case that the downstream task is predefined, it is also practical to carry out step 2 and step 3 collaboratively by building a dataset agent that has characteristics similar to the fine-tuning dataset, which can help inform the fine-tuning performance.

In summary, this algorithm can be flexibly and effectively used to transfer the knowledge of pre-trained models to newly developed models, and it is also easy to be used with fine-tuning together.

3 Related Neural Networks

In this work, five state-of-the-art deep neural networks for semantic segmentation were used to conduct the experiment.

DeepLab v3+: Deeplab v3+ (Chen et al., 2018) combines the advantages of spatial pyramid pooling with an encoder-decoder structure. Spatial pyramid pooling is used at the final convolution stage of the encoder path, and captures rich contextual information by applying atrous convolution. Parallel atrous convolutions at different rates are performed and then fused together to create dense features, before being fed into the decoder path. The decoder path is used to gradually recover object boundaries, using the rich semantic information in the final feature map of the encoder.

U-Net: U-Net (Ronneberger et al., 2015) is one of the most widely used segmentation models in medical imaging. The U-Net architecture consists of an encoder-decoder structure with convolutions and downsampling in the encoder path to generate rich semantic information, and upsampling in the decoder path to increase resolution and enable contextual information. To preserve relevant local information, higher resolution features from each parallel convolution layer of the encoder are concatenated with their corresponding upsampled layer in the decoder (skip connections).

Attention U-Net: Attention U-Net (Oktay et al., 2018) is a modification of U-Net. Attention gates control the information that is passed through the skip connections from the parallel encoder layer to the decoder layer. Attention gates make the model focus on target structures of varying shapes and sizes and suppress irrelevant features while highlighting important features.

SDU-Net: Considering the downsampling convolutions in the U-Net reduces the field of view and degrades pixel correlation, SDU-Net (Wang et al., 2020) processes feature maps at each convolution layer using parallel atrous convolutions at different rates and element-wise concatenate feature maps to generate an output for the next layer. Concatenating atrous convolutions is effective to capture features in both small and large receptive fields. SDU-Net achieves high performance using only 40 percent of the parameters in U-Net.

Panoptic: Panoptic segmentation (Kirillov et al., 2019) merges semantic segmentation and object detection (instance segmentation) such that all pixels in the image, including background pixels, are given a class as in semantic segmentation, and each unique object within the image is given a unique ID as in object detection. Therefore, panoptic segmentation method provides a more rich and complete segmentation which is a step closer to real-world vision problems.

4 DATASETS

As shown in Table 1, our experiments employed six image datasets named XXX (XXX is used for Anonym) Breast Lesion-1, XXX Breast Lesion-2, Baheya Breast Lesion Al-Dhabyani et al. (2020), Nerve Structure, Thyroid Nodule, and Skin Lesion (ISIC2019). XXX Breast Lesion-1, XXX Breast Lesion-2, and Thyroid Nodule were collected at XXX while Baheya Breast Lesion, Nerve Structure, and Skin Lesion are publicly available. The example image for each dataset is present in Figure 2.

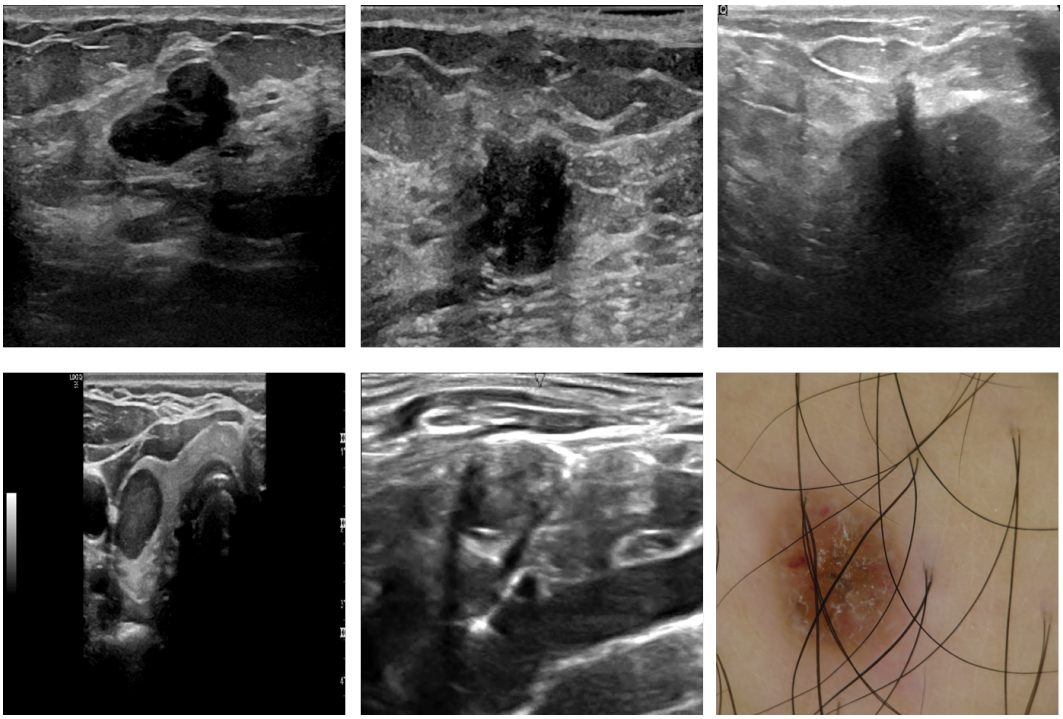

Figure 2: Image examples for the datasets. From left to right, the top row presents the examples for XXX Breast Lesion-1, XXX Breast Lesion-2, and Baheya Breast Lesion; the second row presents the examples for Thyroid Nodule, Nerve Structure, and Skin Lesion.

XXX Breast Lesion: Consecutive patients who were evaluated and referred for Ultrasound (US)-guided biopsy (BI-RADS 4 or 5 mass) at XXX Institute between September 2017 and October 2019 were recruited. US evaluation of breast lesions was performed as per standard practice, with a high-frequency linear transducer on a GE Logiq E9 system (General Electric Healthcare, Chicago, IL, USA). Eligible subjects were adult patients with breast masses detected by the US. Those who had masses without definite pathological diagnoses were excluded. Finally, a total of 1385 B-mode US images from 660 breast lesions (292 malignant; 368 benign) were adopted. These images were

Table 1: Datasets.

Dataset	Image Number	Modality	Data Source
XXX Breast Lesion-1	1,237	Ultrasound	XXX
XXX Breast Lesion-2	148	Ultrasound	XXX
Baheya Breast Lesion	697	Ultrasound	Baheya Hospital
Thyroid Nodule	3,830	Ultrasound	XXX
Nerve Structure	11,143	Ultrasound	Kaggle Challenge
Skin Lesion	25,331	Dermoscopy	ISIC Challenge

randomly assigned to XXX Breast Lesion-1 (n=1237) and XXX Breast Lesion-2 (n=148). Two radiologists manually annotated the borders of each lesion.

Baheya Breast Lesion: Baheya Breast Lesion was released by Al-Dhabyani et al. (2020) containing breast US images acquired by GE Logiq E9 US and Logiq E9 Agile US system. The data was collected in 2018 including breast US images among women in ages between 25 and 75 years old. The dataset consists of 780 images from 600 female patients. The images are categorized into three classes, which are normal (n=133), benign (n=487), and malignant (n=210). Radiologists from Baheya Hospital, Egypt manually annotated the lesion boundaries. In our experiment, we only used images with benign or malignant lesions, a total of 697 images.

Thyroid Nodule: Over two years, 3,830 US images of thyroid nodules were retrospectively collected to form the XXX Thyroid Nodule Dataset. Data from patients (aged ¿ 18 years old) who underwent US examination before their thyroid biopsy between April 2010 and April 2012 were included. Patients who underwent a biopsy for thyroid evaluation were excluded. All thyroid imaging was performed by a sonographer with more than 20 years of thyroid US experience using a SuperSonic Imagine Aixplorer (Aix-en-Provence, France) US imaging system with a 15-4–MHz broadband linear transducer. Manual delineation of thyroid nodules by two expert radiologists were the reference standard.

Nerve Structure: Hosted on Kaggle, the US Nerve segmentation dataset was made available for a prediction competition in segmenting nerve structures in US images of the neck, more specifically focusing on a collection of nerves called the Brachial Plexus (BP). The large dataset was manually annotated by human experts, which are available at https://www.kaggle.com/c/ultrasound-nerve-segmentation. Some other characteristics of the dataset include potential noise, artifacts, and mistakes when labeling the ground truth as well as near-identical images due to the inherent image acquisition technique. We combined the original images of the training set and the test set in our experiment, a total of 11,143 images.

Skin Lesion: The skin-lesion dataset was from a challenge hosted by the International Skin Imaging Collaboration (ISIC) in 2019 (available at https://challenge2019.isic-archive.com/) where the task was to perform multiclass classification to distinguish dermoscopic images among nine distinct diagnostic categories of skin cancers. This dataset contains 25,331 images for training across 8 classes and a test dataset containing an additional outlier class that is not represented in the training dataset.

5 EXPERIMENT

The experiment evaluated the proposed method on US image segmentation, which consists of two parts: the segmentation evaluation for student models with knowledge transfer and the segmentation evaluation for student models with further fine-tuning.

5.1 MODEL TRAINING

We adopted DeepLab v3+ as the teacher model and employed U-Net, Attention U-Net, SDU-Net, and Panoptic as the student models. And we used XXX Breast Lesion-1 as the teacher-training dataset, and the other four datasets, including Baheya Breast Lesion, Thyroid Nodule, Nerve Structure, and Skin Lesion, as data agents for training the student models.

Table 2: Average Dice scores (±std, %) for student models on XXX Breast Lesion-2. Direct Learning indicates student models were directly trained on XXX Breast Lesion-1, and the rows below are for student models trained on pseudo-annotated datasets.

	U-Net	**Attention U-Net**	**SDU-Net**	**Panoptic**
Direct Learning	82.03±20.56	83.53±20.51	86.75±17.12	85.95±14.82
Baheya Breast Lesion	72.64±20.93	74.72±22.28	85.00±14.48	81.12±22.52
Thyroid Nodule	66.28±29.23	71.17±25.10	86.12±14.11	82.77±18.70
Nerve Structure	45.07±37.40	58.66±33.61	78.53±23.24	80.06±23.41
Skin Lesion	38.92±22.12	34.93±22.16	50.10±31.64	55.16±33.85

As mentioned above, the knowledge transfer from the teacher model to the student model includes two main steps. In the first step, DeepLab v3+ was trained on XXX Breast Lesion-1, with the training and validation split ratio of 8:2. In the second step, the trained DeepLab v3+ was used to process the image in the dataset agents to generate the mask, i.e., the pseudo annotation. Each of the four student models was trained on each of the four pseudo-annotated datasets, with the training and validation split ratio set to 9:1. For a fair comparison, we also tested the performance of U-Net, Attention U-Net, SDU-Net, and Panoptic by training them from scratch on XXX Breast Lesion-1, with the training and validation split ratio set to 8:2.

For downstream tasks, the student models were further fine-tuned and tested on Baheya Breast Lesion and Thyroid Nodule, respectively. Baheya Breast Lesion was separated into a train-val set of 50 images (training:validation=8:2) and a test set of 595 images. Thyroid Nodule was separated into a train-val set of 50 images (training:validation=8:2) and a test set of 3780 images.

For simplicity, we name the model trained on the teacher-training dataset as a direct-trained model, the model trained on pseudo-annotated datasets as an educated model. Please refer to section A.1 for more training details.

5.2 Performance of Knowledge Transfer

The educated student models were tested on XXX Breast Lesion-2 and Baheya Breast Lesion. Meanwhile, we also tested the performance of the teacher model as well as the direct-trained student models. Note that the performance were evaluated against manual annotations, i.e., ground truth.

1) Test on XXX Breast Lesion-2

The teacher model, DeepLab v3+, resulted in an average Dice score (±std) of 87.04%(±16.01%) on XXX Breast Lesion-2. The Dice scores of the student models are shown in Table 2. The educated student models obtained impressive performance. For example, the Dice scores of SDU-Net trained on pseudo-annotated Thyroid Nodule and Baheya Breast Lesion are quite similar to the teacher model and the direct-trained SDU-Net, although thyroid nodule and breast lesion are two different kinds of objects and Baheya Breast Lesion is a quite small dataset. Also, we can see the performance of Panoptic trained on pseudo-annotated Thyroid Nodule, Baheya Breast Lesion, and even Nerve is close to the teacher model and the direct-trained Panoptic.

Even trained on pseudo-annotated Skin Lesion, these four student models obtained moderate Dice scores, although the image of Skin Lesion is of a different modality. As we can see from Figure 2, the skin lesion is visually much different from the breast lesion.

2) Test on Baheya Breast Lesion

Similarly, the models were tested on Baheya Breast Lesion. The teacher model, DeepLab v3+, resulted in an average Dice score (±std) of 65.93%(±32.55%). The Dice scores of the student models on Baheya Breast Lesion are shown in Table 3. By being directly trained on XXX Breast Lesion-1, the students models U-Net, Attention U-Net, SDU-Net, and Panoptic obtained average Dice scores (±std, %) of 66.83±33.93, 61.98±33.61, 62.92±33.00, and 69.12±30.68.

From Table 3, we can see all students models trained on pseudo-annotated Baheya Breast Lesion resulted in higher Dice scores than the teacher model that was trained on XXX Breast Lesion-1 with

Table 3: Average Dice scores (±std, %) for each student model on Baheya Breast Lesion. Direct Learning indicates student models were directly trained on XXX Breast Lesion-1 with ground truth, and the rows below are for student models educated by corresponding pseudo-annotated datasets.

	U-Net	**Attention U-Net**	**SDU-Net**	**Panoptic**
Direct Learning	62.92±33.00	61.98±33.61	66.83±33.93	69.12±30.68
Baheya Breast Lesion	65.58±31.22	65.14±31.07	69.36±30.82	67.89±31.93
Thyroid Nodule	47.41±35.20	54.63±33.18	63.71±32.80	62.30±33.27
Nerve Structure	35.47±36.60	42.86±36.92	59.62±34.10	62.39±32.28
Skin Lesion	28.16±22.57	27.64±22.77	30.17±33.53	40.94±37.19

Table 4: Average Dice scores (±std, %) for student models fine-tuned or trained from scratch on Baheya Breast Lesion. Direct Learning indicates student models were directly trained on XXX Breast Lesion-1 with ground truth, and the rows below are for student models trained on corresponding pseudo-annotated datasets.

	U-Net	**Attention U-Net**	**SDU-Net**	**Panoptic**
Scratch	37.20±24.65	37.15±24.74	46.88±28.87	52.88±30.31
Direct Transfer	59.29±29.37	58.96±30.19	70.13±31.05	72.63±29.37
Baheya Breast Lesion	59.29±29.04	65.31±29.68	70.03±31.25	70.61±29.64
Thyroid Nodule	57.08±28.51	60.09±30.04	70.48±30.26	68.04±29.50
Nerve Structure	56.58±26.94	63.21±30.61	66.55±31.83	71.19±28.29
Skin Lesion	45.03±26.08	43.29±27.52	53.45±29.92	62.18±31.91

ground truth. Referring to the direct-trained students models, all student models trained on pseudo-annotated Baheya Breast Lesion resulted in better or comparable performance. It is partly because the training and test of the student models used the same dataset, although the training only has pseudo annotations. This is a unique advantage of the proposed algorithm. Meanwhile, SDU-Net and Panoptic trained on Thyroid Nodule and Nerve also obtained comparable performance. When Skin Lesion was used as data agent, although the performance is not comparable to that with other data agents, all student models learn considerable knowledge from Skin Lesion.

5.3 Performance with Fine-Tuning

The student models that were trained on pseudo-annotated datasets were fine-tuned on the train-val sets of Baheya Breast Lesion and Thyroid Nodule, respectively. Meanwhile, we conducted two more types of training for comparison, indicated by Scratch and Direct Transfer. Scratch: the student models were trained from scratch on the train-val set. Direct Transfer: the direct-trained student models were fined-tuned on the train-val set. All the models were tested on the test sets of Baheya Breast Lesion and Thyroid Nodule.

1) Fine-tuning on Baheya Breast Lesion

The fine-tuned teacher model, DeepLab v3+, resulted in an average Dice score (±std) of 69.90%±31.01% on Baheya Breast Lesion. The results of the student models are presented in Table 4. All fine-tuned student models outperformed the models trained from scratch by a large margin. The models trained on the pseudo-annotated Baheya Breast Lesion, Thyroid Nodule, and Nerve can well be used for fine-tuning, which resulted in similar performance as the teacher model and the direct-trained student models. The results demonstrate that the proposed method can be well used with transfer learning.

2) Fine-tuning on Thyroid Nodule

The fine-tuned teacher model, DeepLab v3+, resulted in an average Dice score (±std) of 66.56%±27.78% on Thyroid Nodule. The results of the student models are presented in Table 5. The student models trained from scratch did not perform well, as the tuning-dataset is small. The educated student models had similar or even better performance by comparing to the direct-trained

Table 5: Average Dice scores (±std, %) for student models fine-tuned or trained from scratch on Thyroid Nodule. Direct Learning indicates student models were directly trained on XXX Breast Lesion-1 with ground truth, and the rows below are for student models trained on corresponding pseudo-annotated datasets.

	U-Net	**Attention U-Net**	**SDU-Net**	**Panoptic**
Scratch	34.02±23.89	35.85±24.51	46.00±27.08	48.96±26.92
Direct Transfer	49.69±30.08	52.55±30.60	58.25±31.62	67.21±28.33
Baheya Breast Lesion	50.93±27.95	41.28±47.74	57.99±30.00	64.92±28.20
Thyroid Nodule	55.41±27.69	57.18±28.82	60.15±31.34	65.65±29.32
Nerve Structure	44.93±27.00	45.51±30.80	53.05±33.07	65.37±28.23
Skin Lesion	41.32±23.37	34.41±24.68	49.97±29.69	54.30±30.29

models, which outperformed the models trained from scratch. This further verified that the effect of knowledge transfer between heterogeneous neural networks.

It is encouraging that the student model trained on pseudo-annotated Thyroid Nodule and Nerve Structure, which are visually much different from teacher-training dataset, achieved similar or even superior Dice score to the direct-trained model. Even trained on the pseudo-annotated Skin Lesion, which is of a different modality, the student model obtained moderate performance. Please refer to Tables 6- 9 for more comparison by the ratio of Dice scores between the educated student model and the direct-trained model. In summary, these two groups of experiments demonstrated the effectiveness of knowledge transfer between heterogeneous models using pseudo-annotated datasets.

6 Conclusion

In this work, we proposed a simple but effective algorithm for transferring knowledge from one neural network to another independent one, through the use of pseudo annotations only. This algorithm is particularly useful in training neural networks in fields (e.g. medical image segmentation) where training datasets with ground truth are expensive and cumbersome to obtain. It is also straightforward to combine the proposed algorithm with fine-tuning.

We conducted a detailed experiment to assess the performance of the algorithm. The tests were conducted with 5 distinct segmentation models and 6 multi-organ datasets. Dice scores were used to compare the results obtained using the pseudo annotations and those obtained using ground-truth annotations. We found that even with pseudo annotations, the student network could achieve similar or even higher Dice score than that obtained with ground truth annotations. This is really encouraging, as this is the first work that studies medical image segmentation using pseudo annotations only, and there are various interesting avenues of future work to further improve the algorithm. For instance, one possibility is to optimize the dataset agent and explore certain metrics to build good dataset agents. Another possibility is to ensemble multiple teacher models to train a student model that may be better generalized and more robust. Lastly, it is interesting to study how the neural architectures affect the performance of knowledge transfer.

References

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018.

Robert Geirhos, David HJ Janssen, Heiko H Schütt, Jonas Rauber, Matthias Bethge, and Felix A Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*, 2017.

Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018.

Ata Jodeiri, Reza A Zoroofi, Yuta Hiasa, Masaki Takao, Nobuhiko Sugano, Yoshinobu Sato, and Yoshito Otake. Region-based convolution neural network approach for accurate segmentation of pelvic radiograph. In *2019 26th National and 4th International Iranian Conference on Biomedical Engineering (ICBME)*, pp. 152–157. IEEE, 2019.

Jaeyong Kang and Jeonghwan Gwak. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*, 7:26440–26447, 2019.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9404–9413, 2019.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88, 2017.

Chaochao Lu and Xiaoou Tang. Surpassing human-level face verification performance on lfw with gaussianface. *arXiv preprint arXiv:1404.3840*, 2014.

Saman Motamed, Isha Gujrathi, Dominik Deniffel, Anton Oentoro, Masoom A Haider, and Farzad Khalvati. A transfer learning approach for automated segmentation of prostate whole gland and transition zone in diffusion weighted mri. *arXiv preprint arXiv:1909.09541*, 2019.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, pp. 3347–3357, 2019.

Rajeev Ranjan, Swami Sankaranarayanan, Ankan Bansal, Navaneeth Bodla, Jun-Cheng Chen, Vishal M Patel, Carlos D Castillo, and Rama Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, 35(1): 66–83, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241. Springer, 2015.

H Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pp. 270–279. Springer, 2018.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015.

Shuhang Wang, Szu-Yeu Hu, Eugene Cheah, Xiaohong Wang, Jingchao Wang, Lei Chen, Masoud Baikpour, Arinc Ozturk, Qian Li, Shinn-Huey Chou, et al. U-net using stacked dilated convolutions for medical image segmentation. *arXiv preprint arXiv:2004.03466*, 2020.

Ming Xie, Guodong Long, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. Multi-center federated learning. *arXiv preprint arXiv:2005.01026*, 2020a.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020b.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Guodong Zeng, Xin Yang, Jing Li, Lequan Yu, Pheng-Ann Heng, and Guoyan Zheng. 3d u-net with multi-level deep supervision: Fully automatic segmentation of proximal femur in 3d mr images. In *International Workshop on Machine Learning in Medical Imaging*, pp. 274–282. Springer, 2017.

A APPENDIX

A.1 TRAINING DETAILS

Two common loss functions were used for the segmentation tasks, the Dice loss function and binary cross entropy loss (BCE). The Dice loss is defined as:

$$L_{dice} = 1 - \frac{2\sum p_{h,w} \cdot \hat{p}_{h,w} + \epsilon}{\sum p_{h,w} + \sum \hat{p}_{h,w} + \epsilon} \tag{1}$$

where (h, w) represents the pixel coordinate, $p_{h,w} \in \{0, 1\}$ is the mask ground truth, $p_{h,w} = 1$ indicates the pixel belonging to the target, $0 \leq \hat{p}_{h,w} \leq 1$ is the prediction probability for the pixel belonging to the target, ϵ is a small real number.

The binary cross entropy loss is defined as:

$$L_{bce} = -\frac{1}{N}\sum p_{h,w} \cdot log(\hat{p}_{h,w}) + (1 - p_{h,w}) \cdot log(1 - \hat{p}_{h,w}) \tag{2}$$

where (h, w) represents the pixel coordinate, $p_{h,w} \in \{0, 1\}$ is the mask ground truth, $p_{h,w} = 1$ indicates the pixel belonging to the target, $0 \leq \hat{p}_{h,w} \leq 1$ is the prediction probability for the pixel belonging to the target, N is the number of pixels.

Adam optimizer was used for training all the networks. The learning rate was set to 0.0001, and all the other parameters were set to the default values in PyTorch. The batch size was set to 4. The epoch number was set to 500 for the training on the teacher-training dataset. Inversely proportional to the image number, the epoch numbers were set to 500, 175, 60, and 30 to for the training on pseudo-annotated Baheya Breast Lesion, Thyroid Nodule, Nerve Structure, and Skin Lesion. Moreover, the training epoch number for fine-tuning was set to 50.

A.2 PERFORMANCE OF KNOWLEDGE TRANSFER

Tables 6 and 7 present the ratio of average dice scores between educated student models and direct-trained student models. It is easy to see that the educated student models, especially SDU-Net and Panoptic, achieved similar performance as the direct-trained student models.

Table 6: Ratio of average Dice scores on XXX Breast Lesion-2 between educated student models and direct-trained student models.

	U-Net	**Attention U-Net**	**SDU-Net**	**Panoptic**
Baheya Breast Lesion	0.89	0.89	0.98	0.94
Thyroid Nodule	0.81	0.85	0.99	0.96
Nerve Structure	0.55	0.70	0.91	0.93
Skin Lesion	0.47	0.42	0.58	0.64

Table 7: Ratio of average Dice scores on Baheya Breast Lesion between educated student models and direct-trained student models.

	U-Net	**Attention U-Net**	**SDU-Net**	**Panoptic**
Baheya Breast Lesion	1.04	1.05	1.04	0.98
Thyroid Nodule	0.75	0.88	0.95	0.90
Nerve Structure	0.65	0.69	0.89	0.90
Skin Lesion	0.45	0.45	0.45	0.59

Table 8: Ratio of average Dice scores on XXX Breast Lesion-2 between fine-tuned educated student models and fine-tuned direct-trained student models.

	U-Net	**Attention U-Net**	**SDU-Net**	**Panoptic**
Baheya Breast Lesion	1.00	1.11	1.00	0.97
Thyroid Nodule	0.96	1.02	1.01	0.94
Nerve Structure	0.95	1.07	0.95	0.98
Skin Lesion	0.76	0.73	0.76	0.86

A.3 Performance with Fine-Tuning

Tables 8 and 9 present the ratio of average dice scores between fine-tuned educated student models and fine-tuned direct-trained student models. Some of the educated student models, even achieved higher Dice scores than the direct-trained student models.

Table 9: Ratio of average Dice scores on Baheya Breast Lesion between fine-tuned educated student models and fine-tuned direct-trained student models.

	U-Net	**Attention U-Net**	**SDU-Net**	**Panoptic**
Baheya Breast Lesion	1.03	0.79	1.00	0.97
Thyroid Nodule	1.12	1.09	1.03	0.98
Nerve Structure	0.90	0.87	0.91	0.97
Skin Lesion	0.83	0.65	0.86	0.81