Transfer Learning with segmentation using GAN in medical imaging

Nattapon Jaroenchai, Edward Tang, Srivardhan Sajja, Beatrice Lovely

CS 543: Computer Vision

Prof. Svetlana Lazebnik

December 13, 2021

1. Introduction

X-ray imagery is a crucial imaging procedure for diagnosing, screening, and monitoring lung illnesses, with over 79 million exams performed annually (the United States, 2015) (Organization for Economic Cooperation, 2013). Finding and accurately segmenting organs like the lung is critical (Mansoor et al., 2015), especially in ML, to eliminate confounders outside the organ (such as respiratory gear, implants, or comorbidities) (Zech et al., 2018).

Automated lung segmentation methods are typically developed and evaluated on small datasets with minimal variability (Goksel et al., 2015) or patients with a single disease class (Yang et al., 2018). These specialized approaches and ML models struggle to generalize to unknown cohorts when used for segmentation. As a result, image processing investigations still rely on semiautomatic segmentation or manual assessment of automated organ masks (Oakden-Rayner et al., 2017; Stein et al., 2016). Human inspection or any human engagement with single data items is not practicable for large-scale data analysis based on thousands of cases. However, disease-specific models are limited in their relevance to unidentified instances, such as in computer-aided diagnosis or cross-sectional data.

Numerous ways to X-ray lung segmentation have been proposed. There are several types of approaches: rule-based (Korfiatis et al., 2007; Hu et al., 2001; Pulagam et al., 2016), atlas-based (Zhang et al., 2005; Iglesias et al., 2015), and machine learning (ML)-based (Sofka et al., 2011; Chen et al., 2019; Agarwala et al, 2017). Because the lung has a low density and a high contrast on x-ray images, thresholding, and atlas segmentation techniques are effective for mild or low-density illnesses such as emphysema. These approaches are complicated by disease-related lung abnormalities such as effusion, atelectasis, consolidation fibrosis, or pneumonia. Multi-atlas registration and hybrid techniques that incorporate multiple atlases, shape models, and other post-processing procedures overcome this issue. However, such intricate processes are not easily repeatable without the source code and underlying set of atlases. However, trained machine learning models can be shared without requiring access to the training data. They are extremely fast to infer and scale effectively with additional training data. As Harrison et al. (2017) demonstrate, deep learning-based segmentation outperforms specialized techniques in the diagnosis of interstitial lung diseases. Unless there are unusual circumstances, trained lung segmentation models are rarely made publicly available, impeding research development. Additionally, machine learning algorithms are bound by the quantity and quality of available training data.

1.1 Transfer Learning

Transfer learning is used to improve a learner from one domain by transferring information from a related domain. It builds on the concept of “prior experience” and its role in learning. It can be understood with the example of a person with an extensive music background who would be able to learn a new instrument faster and more efficiently than someone who has no musical knowledge. Transfer learning breaks the assumption about machine learning and data mining algorithms that the training and future data must be in the same feature space and have the same distribution (Pan et al. 2009).

Transfer learning is extremely useful in domains where there is little labeled training data available, for example, medical imagery. Transfer learning is also much cheaper in terms of computational requirements. We can leverage the existing representations learned in a neural network and re-train it for a much shorter time than would be required if training from scratch.

Typically transfer learning performs best when the new domain is similar to the original domain, eg., two different segmentation tasks or image-to-image translation tasks. However, one well-known shortcoming of transfer learning is the phenomenon known as “catastrophic forgetting” in which the re-trained network performs very well on its new task but “forgets” and performs very badly on the task it was originally trained on. Recent research has suggested different methods for alleviating catastrophic forgetting. (Mallya et al., 2018), (Zhai et al., 2021)

To address the limitation of ML-based lung segmentation, in this study, we proposed a transfer-learning-based method for lung segmentation. Firstly, we look at the possible initial pretrained generative adversarial networks (GAN) that perform well on the semantic segmentation task. We found that Cycle GAN (Zhu et al., 2017) and Pix2Pix (Isola et al., 2017) are the most suitable networks. We, then, took advantage of the knowledge learned from the different tasks which are day-to-day object segmentation, and adapt it to be able to segment general lung chest images in our dataset. This adaptation can be achieved by the transfer learning method.

Our goal in this project is to examine the performance of transfer learning Generative Adversarial Networks (GAN) in semantic segmentation tasks in CMP Facade Database (Tyleček & Šára, 2013) symmetric segmentation in remote sensing imagery such as Cityscapes Dataset by Cordts et al., (2016). Specifically, we are interested in using the transfer learning technique called Piggyback proposed by Mallya et al., (2018).

1.2 Image Segmentation using GANs (Generative Adversarial Networks)

Separate from classification and object detection, segmentation is the ability for a network to accurately segment the contents of an image according to what the object is (eg., windows, streets, trees). Semantic segmentation information can be used to make decisions about how to respond to the environment (eg., a self-driving car navigating urban streets). Instance Segmentation is an interesting use case with applications in medical imagery (for example accurately predicting the bounds of a tumor or cell structures in CT scans for treatment planning).

Segmentation is a difficult and resource-intensive task. It is memory-intensive as the output needs to be the same size as the input image. It is also much more difficult for a model to label every individual pixel. By comparison, the relatively simpler task of image classification only requires predicting a single class for the entire image and outputting a single vector of the same dimension as the number of classes.

Generative Adversarial Networks (GAN) have seen significant adoption for the image segmentation task. GANs are a combination of two neural networks, a Generator and a Discriminator that are trained simultaneously. The Generator is tasked with generating an image (in this case a segmentation mask) from some input. In the case of unpaired image generation, the Generator is fed with random noise, in segmentation or image-to-image translation tasks, the input is an image (eg the image to be segmented). The task of the Discriminator is then to distinguish the generated image as “real” or “fake”. The Discriminator is trained with target examples that are labeled as “real” and the generated images from the Generator, labeled as “fake” and its task is to distinguish between the two. During initial training, only the Discriminator is trained and the weights for the Generator are kept fixed and loss is only backpropagated through the Discriminator. As the Discriminator improves its ability to distinguish “real” from “fake”, we begin training the Generator. Loss is backpropagated also through the Generator, which is penalized for failing to “fool” a discriminator. The goal is a Generator powerful enough to fool the Discriminator with its generated images. GAN training is a difficult balance to train a Generator good enough to fool a well-trained Discriminator, while not allowing the Discriminator to become too good at distinguishing fake images, in which case the Generator would never be rewarded for fooling the Discriminator, and would fail to learn anything useful.

2. Materials

2.1 Datasets

2.1.1 The CMP Facade Database (Tyleček & Šára, 2013)

is a dataset of facade images assembled at the Center for Machine Perception, which includes 606 rectified images of facades from various sources, which have been manually annotated. The facades are from different cities around the world and have diverse architectural styles.[1]

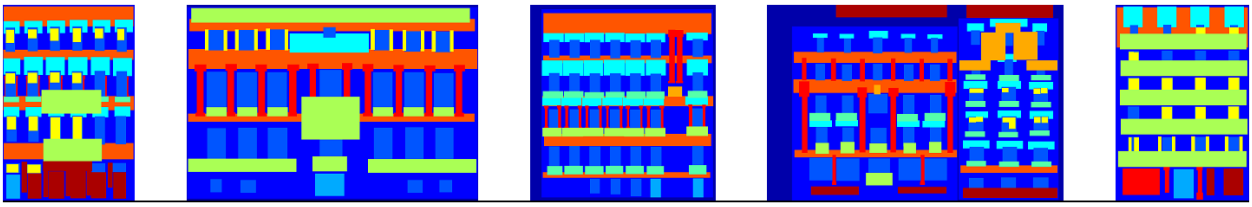


Figure 1: Example image from CMP Facade Database

2.1.2 The Montgomery County Chest X-ray Database

The Montgomery County Chest X-ray Database is the standard digital image database for Tuberculosis created by the National Library of Medicine in collaboration with the Department of Health and Human Services, Montgomery County, Maryland, USA. The set contains data from X-rays collected under Montgomery County`s Tuberculosis screening program. There are a total of 138 X-rays: - 58 cases with the manifestation of tuberculosis, and 80 normal cases.[2] This set contains 138 posterior-anterior x-rays, of which 80 x-rays are normal and 58 x-rays are abnormal with manifestations of tuberculosis. All images are de-identified and available in DICOM format. The set covers a wide range of abnormalities, including effusions and miliary patterns. The data set includes radiology readings available as a text file. (Antani et al. 2020).

2.4. Computational Resources

We use Google Colab Pro from 3 google accounts for the model training. Therefore, the hardware that we used to train the model is NVidia K80, P100, or T4 with 25GB (high memory VMs) memory. We run the model using Pytorch 1.10.0 and Numpy 1.19.5.

3. Methodology

We experiment with transfer learning between two quite different domains: images of facades and medical images (x-rays). The first domain, facades, presents lots of different colors and shapes and has mostly larger details, whereas the x-ray images are grey-scale and contain very fine detail and imprecise borders that may confuse a ML model. We apply transfer learning on the two different pre-trained models listed in 2: pix2pix trained on the CMP Facades dataset (facade-to-label segmentation task) and CycleGAN (Zhu et al., 2017), trained on the horse2zebra dataset. Both models are provided by the authors on GitHub.[3]

3.1.1 Pix2Pix Network

Pix2pix is a conditional GAN model developed by Isola et al., (2017). Conditional means the generator is fed an input image as a “condition” rather than a random noise vector. It was originally trained for and tested on several paired image-to-image translation tasks (eg aerial images-to-maps, facades-to-label segmentation) The Generator architecture used in the pix2pix model is the UNet which follows the common encoder-decoder

[1] https://cmp.felk.cvut.cz/~tylecr1/facade

[2] https://bit.ly/3uHdxAk

[3] junyanz/pytorch-CycleGAN-and-pix2pix: Image-to-Image Translation in PyTorch (github.com)

architecture with added skip-connections in which intermediate outputs in the encoder are fed directly to the corresponding layer in the decoder, along with the data that has traveled through the entire network.

3.1.2 CycleGAN Network

developed by Zhu et al., (2017) for unpaired image-to-image translation (eg image-to-painting, horse-to-zebra). The model learns to translate images from the input domain X to output domain Y where there are no ground-truth images to compare to, only examples of eg paintings or zebras. The goal is that the distribution of output images G(X) be indistinguishable from the distribution of target domain Y. The main contribution of this model is the introduction of a cycle-consistent loss. This introduces an inverse mapping function F such that $F(G(X)) \sim X$, eg that the inverse from output to input should produce something very similar to the original input.

We fine-tuned both models for the Montgomery County Chest X-ray Dataset and compared the performance of each model. We experiment with different hyperparameters (training length in epochs and learning rate decay). See results in 3.

4. Results:

We trained on 603 training images and tested on a provided test set of 30 images in the Montgomery County Chest X-ray dataset. We experimented with different hyperparameters: number of epochs and learning rate decay. See results below in Figures 3, 4.

4.1 Metrics

We quantitatively evaluate our results in Table 1. DICE score is a common metric to evaluate the similarity of an output segmentation to ground truth. We chose to also compute precision, recall, and IoU (intersection over union) scores on the binary segmentation task. DICE score is a good metric because it provides a measure of the overall segmentation quality balancing precision and recall, DICE and IoU are the two most common metrics to evaluate segmentation tasks. An argument can be made that for a binary segmentation task in a medical setting, eg. segmenting cancerous tissue from healthy tissue, a slight drop in precision in favor of accurately recalling cancerous tissue would be allowable, and preferable to the opposite case. If a segmentation misses cancerous tissue it may have a detrimental effect on treatment planning (eg radiation therapy) and ultimate success of treatment, whereas small sections of healthy tissue misclassified and therefore treated would be less detrimental.

Most outputs are visually quite good, with the exception of some missing patches. Training the CycleGAN for 4x longer (120 epochs) only yielded an absolute improvement of 2 percentage points in DICE score, and 3 percentage points in IoU. This is consistent with most deep learning approaches that only see very gradual improvement after the initial epochs.

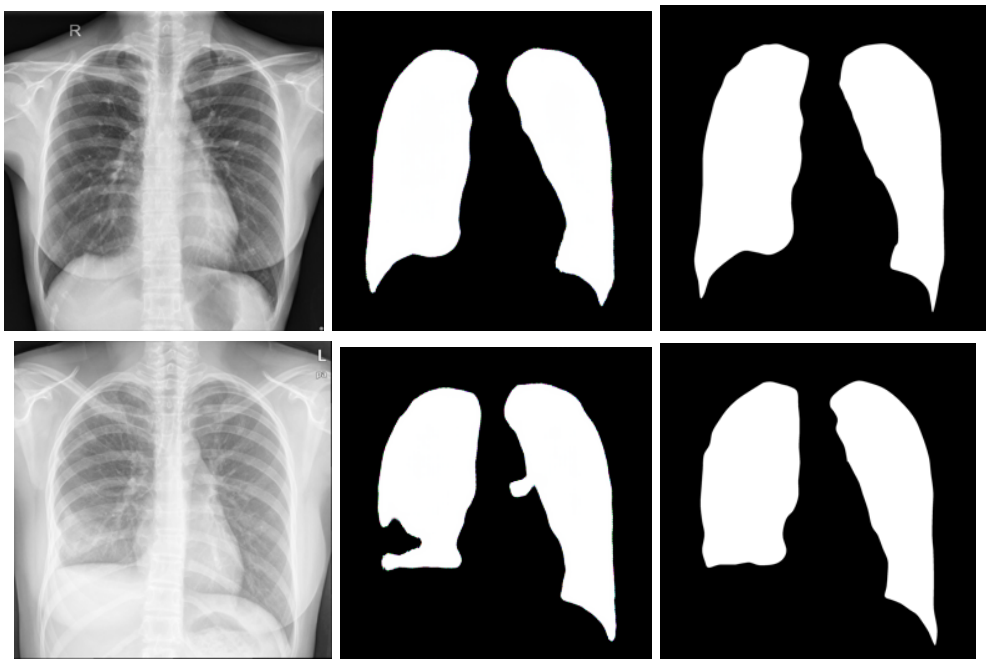


Figure 3: Output of the pix2pix model trained for 120 epochs on the lung dataset, column 1 is input, column 2 is output, column 3 is ground truth.

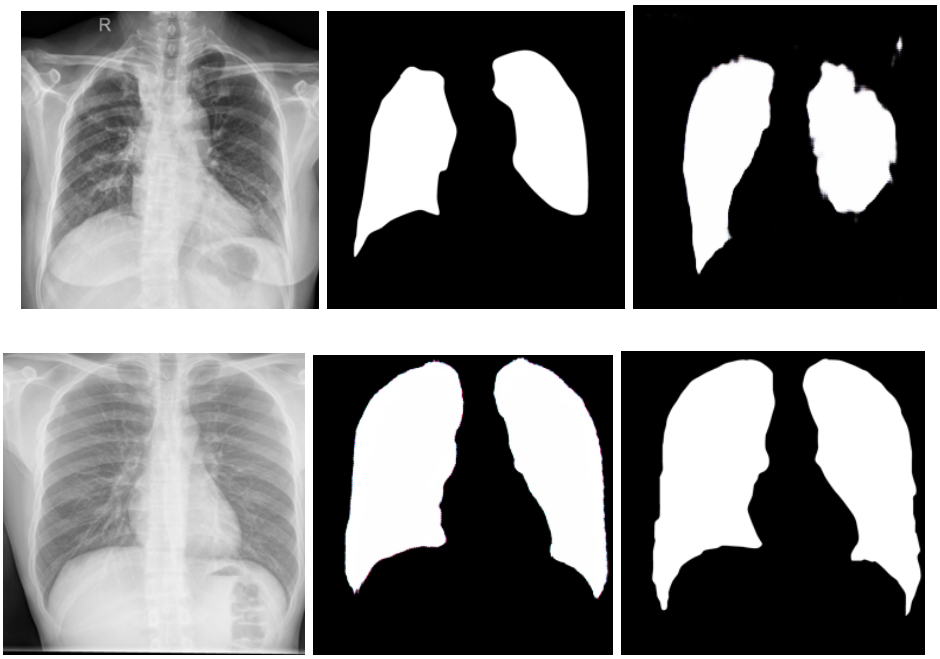


Figure 4: Output of the pix2pix model trained for 30 epochs on the lung dataset, column 1 is input, column 2 is output, column 3 is ground truth.

Table 1: Results of each model fine-tuned on Montgomery County Chest X-ray Dataset

Metrics (mean of test set)	CycleGAN 120 epochs, lr fixed at 2e-4	CycleGAN 30 epochs, Lr fixed at 2e-4	Pix2pix 30 epochs (lr decay for the last 20 epochs)	Pix2pix 30 epochs (lr decay for all 30 epochs)	benchmark/SOTA (Islam, J. and Zhu, Y., 2018)
IoU	0.908	0.88	0.72	0.91	-
per-pixel-accuracy	0.97	0.96	0.83	0.91	-
Precision	0.94	0.93	0.83	0.84	-
Recall	0.96	0.94	0.83	0.91	-
DICE score	0.95	0.93	0.83	0.91	0.98

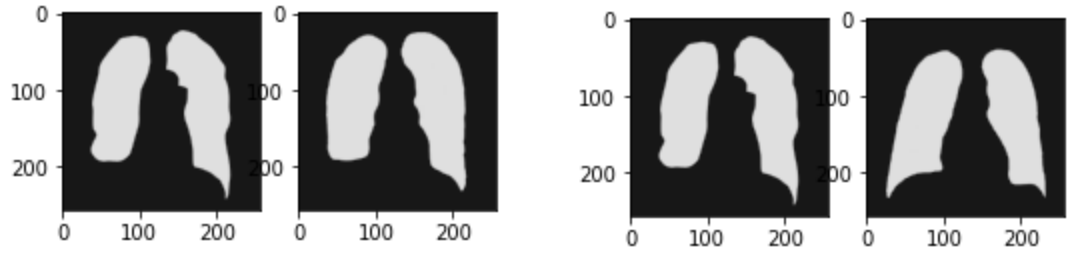


Figure 5: Examples of real mask (right) vs generated mask (right) with fine-tuned pix2pix model; (left) 30 epochs of decaying learning rate; (right) 30 epochs: 10 with steady, 20 with decaying learning rate.

5. Discussion and conclusions:

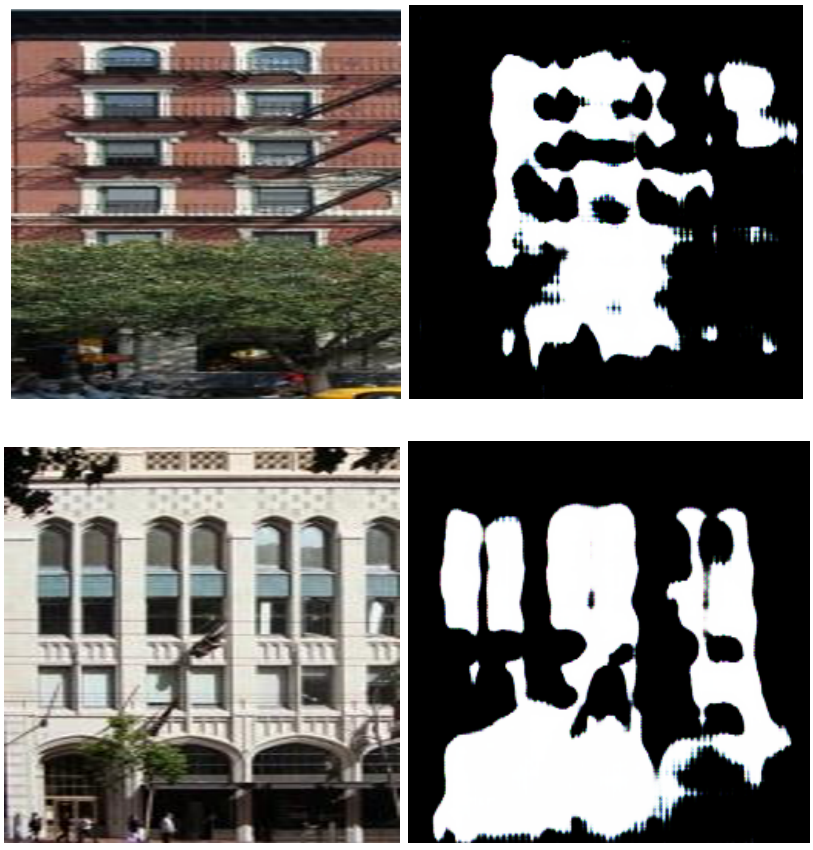
We achieved quite good segmentation results and successfully illustrated that transfer learning is a much faster, cheaper (in terms of time and GPU resources) approach than scratch-training, also in the GAN setting. Transfer learning is especially useful in the medical setting where training data is often scarce. We were able to transfer knowledge between two very different domains (outdoor, multicolored facade images to greyscale x-rays) with reasonable accuracy. As a benchmark, we refer to (Islam and Zhang, 2018) for state of the art results on the Montgomery Chest X-ray database. They achieved a DICE score of 0.98. Training these models from scratch would likely have taken several days on multiple parallel GPUs, whereas we achieved reasonable results in 1.5 hrs (30 epochs), on a single moderately powerful GPU. Our best performing model, CycleGAN trained for 120 epochs (6 hrs) performed the best, only 3 points below SOTA (Islam and Zhang, 2018) who trained on a more powerful NVIDIA TITAN GPU for 200 epochs.

It would be an interesting experiment to transfer in the opposite direction (x-ray to outdoor facade segmentation. Plausibly it would have worse results as a model trained only on grey-scale images would be entirely missing a representation of color, and of the more diverse shapes present in eg the facades dataset.

The pix2pix model performs better with a decaying learning rate than a fixed one, as this prevents the model from converging too quickly on a suboptimal solution.

5.1 Forgetting

Below are some results from an experiment testing the fine-tuned pix2pix model again on the dataset it was originally trained on (facade segmentation). The model changed to a binary segmentation and was unable to change back, but even with that, it is quite clear that the original task has been almost entirely forgotten. This illustrates the need for models adapted for “lifelong learning” eg learning sequential tasks without forgetting the previous ones.



5.1.1 A proposed solution to forgetting - Piggyback

A proposed solution to the problem of catastrophic forgetting is a method called Piggybacking. It was initially proposed by Mallya et al., (2018) and extended to GANs by Zhai et al., (2021). Piggyback works on the notion of a filter-bank that is continually expanded. This extends the original model with additional parameters, but these parameters are a fraction of what would be required to train an entirely new model from scratch or make a copy of the original model that is trained by transfer learning.

In Piggyback training, the adaptation of a network trained for task T1 to task T2 is achieved through two key modifications. For each layer in task T2, a set of filters is constructed by factorizing the filters from task T1 with a learned weight matrix, denoted the piggyback weight matrix, these are referred to as constrained filters. In addition to this, a set of new, unconstrained filters is added to capture unique features inherent to the new task. Thus the new layers for task T2 consist of a concatenation of constrained filters from task T1 and unconstrained filters. The network can be extended to task T3,...Tn in a similar manner. For each new task, the additional parameters are only the piggyback weight matrix and the unconstrained filters. This piggyback weight matrix allows the model to adapt

its weights to a new task without changing the weights themselves (thus “forgetting” its original task) and the added unconstrained filters allow the model to learn unique features of the new task not present in the first.

Piggyback training has been shown to successfully allow a network to perform well on multiple tasks, while only being trained on one at a time. (Zhai et al., 2021)

6. Statement of individual contribution:

Nattapon Jaroenchai: Data collecting and curating in addition to exploring transfer learning.

Edward Tang: Running tests and collecting data from test results, exploring hyperparameter fine-tuning.

Srivardhan Sajja: Exploring transfer learning, formulating results, and writing reports.

Beatrice Lovely: Writing data loader and exploring efficient lifelong learning, code for computing aggregated metrics, and writing report (discussion, methodology and portions of results).

References

Agarwala, S., Nandi, D., Kumar, A., Dhara, A. K., Sadhu, S. B. T. A., & Bhadra, A. K. (2017, November). Automated segmentation of lung field in HRCT images using active shape model. In *TENCON 2017-2017 IEEE Region 10 Conference* (pp. 2516-2520). IEEE.

Chen, G., Xiang, D., Zhang, B., Tian, H., Yang, X., Shi, F., ... & Chen, X. (2019). Automatic pathological lung segmentation in low-dose CT image using eigenspace sparse shape composition. *IEEE transactions on medical imaging*, *38*(7), 1736-1749.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. ArXiv.org. https://arxiv.org/abs/1604.01685

Dong, H., Yang, G., Liu, F., Mo, Y., & Guo, Y. (2017). Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks. ArXiv.org. https://arxiv.org/abs/1705.03820

Goksel, O., Foncubierta-Rodríguez, A., del Toro, O. A. J., Müller, H., Langs, G., Weber, M. A., ... & Hanbury, A. (2015, May). Overview of the VISCERAL Challenge at ISBI 2015. In *VISCERAL Challenge@ ISBI* (pp. 6-11).

Harrison, A. P., Xu, Z., George, K., Lu, L., Summers, R. M., & Mollura, D. J. (2017, September). Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images. In *International conference on medical image computing and computer-assisted intervention* (pp. 621-629). Springer, Cham.

Hu, S., Hoffman, E. A., & Reinhardt, J. M. (2001). Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE transactions on medical imaging*, *20*(6), 490-498.

Iglesias, J. E., & Sabuncu, M. R. (2015). Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, *24*(1), 205-219.

Islam, J. and Zhang, Y., 2018. Towards Robust Lung Segmentation in Chest Radiographs with Deep Learning. Machine Learning for Health, NeurIPS.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125-1134).

Korfiatis, P., Skiadopoulos, S., Sakellaropoulos, P., Kalogeropoulou, C., & Costaridou, L. (2007). Combining 2D wavelet edge highlighting and 3D thresholding for lung segmentation in thin-slice CT. *The British journal of radiology*, *80*(960), 996-1004.

Mallya, A., Davis, D., & Lazebnik, S. (2018). Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights. ArXiv.org. https://arxiv.org/abs/1801.06519

Mansoor, A., Bagci, U., Foster, B., Xu, Z., Papadakis, G. Z., Folio, L. R., ... & Mollura, D. J. (2015). Segmentation and image analysis of abnormal lungs at CT: current approaches, challenges, and future trends. *Radiographics*, *35*(4), 1056-1076.

Oakden-Rayner, L., Carneiro, G., Bessen, T., Nascimento, J. C., Bradley, A. P., & Palmer, L. J. (2017). Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework. *Scientific reports*, *7*(1), 1-13.

Organization for Economic Cooperation. (2013). *Health at a glance 2013: OECD Indicators*. OCDE.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345-1359.

Pulagam, A. R., Kande, G. B., Ede, V. K. R., & Inampudi, R. B. (2016). Automated lung segmentation from HRCT scans with diffuse parenchymal lung diseases. *Journal of digital imaging*, *29*(4), 507-519.

Rajaraman, S., & Antani, S. K. (2020). Modality-specific deep learning model ensembles toward improving TB detection in chest radiographs. IEEE Access, 8, 27318-27326.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2), 227-244.

Sofka, M., Wetzl, J., Birkbeck, N., Zhang, J., Kohlberger, T., Kaftan, J., ... & Zhou, S. K. (2011, September). Multi-stage learning for robust lung segmentation in challenging CT volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 667-674). Springer, Berlin, Heidelberg.

Stein, J. M., Walkup, L. L., Brody, A. S., Fleck, R. J., & Woods, J. C. (2016). Quantitative CT characterization of pediatric lung development using routine clinical imaging. *Pediatric radiology*, *46*(13), 1804-1812.

Tyleček, R., & Šára, R. (2013). Spatial Pattern Templates for Recognition of Objects with Regular Structure. *Lecture Notes in Computer Science*, 364–374. https://doi.org/10.1007/978-3-642-40602-7_39

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1), 1-40.

Wiens, J., Guttag, J., & Horvitz, E. (2014). A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4), 699-706

Yang, J., Veeraraghavan, H., Armato III, S. G., Farahani, K., Kirby, J. S., Kalpathy‐Kramer, J., ... & Sharp, G. C. (2018). Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017. *Medical physics*, *45*(10), 4568-4581.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11), e1002683.

Zhai, M., Chen, L., He, J., Nawhal, M., Tung, F., & Mori, G. (2021). Piggyback GAN: Efficient Lifelong Learning for Image Conditioned Generation. ArXiv.org. https://arxiv.org/abs/2104.11939

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision