Learning Multi-Class Segmentations From Single-Class Datasets

Konstantin Dmitriev and Arie E. Kaufman
Stony Brook University, Stony Brook, NY
kdmitriev@cs.stonybrook.edu

Abstract

Multi-class segmentation has recently achieved significant performance in natural images and videos. This achievement is due primarily to the public availability of large multi-class datasets. However, there are certain domains, such as biomedical images, where obtaining sufficient multi-class annotations is a laborious and often impossible task and only single-class datasets are available. While existing segmentation research in such domains use private multi-class datasets or focus on single-class segmentations, we propose a unified highly efficient framework for robust simultaneous learning of multi-class segmentations by combining single-class datasets and utilizing a novel way of conditioning a convolutional network for the purpose of segmentation. We demonstrate various ways of incorporating the conditional information, perform an extensive evaluation, and show compelling multi-class segmentation performance on biomedical images, which outperforms current state-of-the-art solutions (up to 2.7 %). Unlike current solutions, which are meticulously tailored for particular single-class datasets, we utilize datasets from a variety of sources. Furthermore, we show the applicability of our method also to natural images and evaluate it on the Cityscapes dataset. We further discuss other possible applications of our proposed framework.

1. Introduction

Tremendous progress has been made in deep learning for semantic segmentation, and one of the major factors of such advances is the public availability of large-scale multi-class datasets, such as ImageNet [7], COCO [24], PASCAL VOC [12], and others. Such variety of available datasets not only provides the means to train and evaluate different segmentation models but also to exhibit diverse labels. However, in contrast to natural images, there are certain domains, where despite the critical importance of segmentation research, the generation of ground truth annotations and labeling is extremely costly and remains a bottleneck in advancing research.

Biomedical images is one such domain where the accurate segmentation of various structures is a fundamental problem, especially in clinical research. In traditional clinical practice, segmentation is often omitted during the diagnostic process. However, manual analysis of biomedical images, including measurements, is subject to large variability, as it depends on different factors, including the structure of interest, image quality, and the clinician's experience. Moreover, segmentation is an essential component in various medical systems that support computer-aided diagnosis (CAD) [9, 14] and surgery and treatment planning. Furthermore, early cancer detection and staging often depend on the results of segmentation.

Remarkable progress has been made in the segmentation of radiological images, such as magnetic resonance imaging (MRI) and computed tomography (CT) 3D scans. Radiological images exhibit various objects, such as abdominal organs (Fig. 1a), within a single image. However, creating expert annotations for such images is a time consuming and intensive task, and thus multi-class datasets are difficult to generate. A limited number of segmentation algorithms have been proposed and evaluated on multi-class datasets. These include private or public datasets, such as VISCERAL [20], which has been unavailable due to a lack of funding. Apart from often being private, these multi-class datasets are frequently limited in size (less than 30 volumes) and come from a single institution, where they were generated using the same imaging protocols and imaging devices, leading to the developed segmentation algorithms being sensitive to such imaging parameters. On the other hand, generation of single-class datasets requires less time and effort, and they are often publicly available as part of challenges, such as, Sliver07 [15] (Fig. 1b) and NIH Pancreas [16] (Fig. 1c). Additionally, these single-class datasets come from different institutions and exhibit variability in factors, such as the presence of malignancy, imaging protocols, and reconstruction algorithms.

However, while single-class datasets often contain the same objects within a single image, the ground truth annotations are provided for only a particular class of objects in the form of binary masks, and the sets of images from

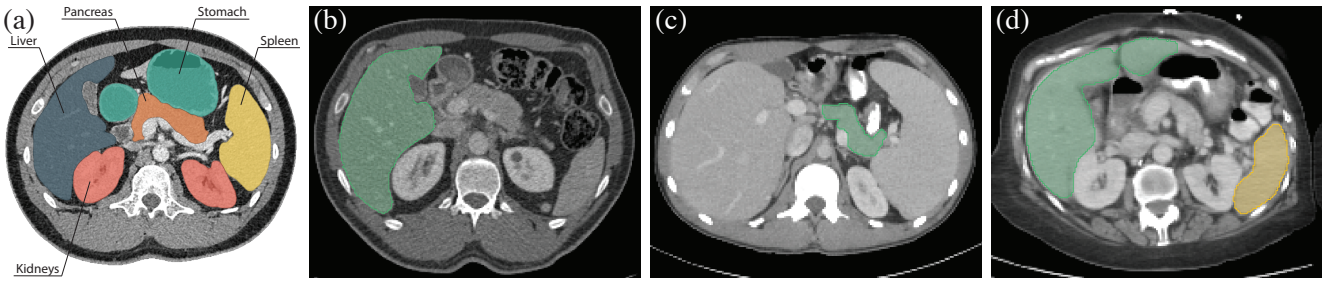

Figure 1: Single-class datasets can be found in various domains, including biomedical images, such as CT scans. While various organs can be seen on a single CT scan (a), the manual generation of the outlines of each organ is an intensive, as often only clinicians can analyze such images, and a time-consuming task, which leads to the lack of comprehensive multi-class datasets. Several single-class datasets have been provided as parts of challenges: (b) a dataset of liver segmentations (Sliver07) [15], (c) a dataset of pancreas segmentations (NIH Pancreas) [16]; while some remain private due to the ethical or legal aspects: (d) a dataset of liver and spleen segmentations. While being the same in nature, the sets of images in these datasets do not overlap, which complicates their simultaneous use for training.

different datasets do not overlap. Thus, it is obstructive to simply combine the datasets to train a single model for multi-class segmentation. Classically, single-class datasets have been used to develop highly tailored solutions for the segmentation of particular classes. In this paper, we introduce a novel and efficient way of training and conditioning a single convolutional network (convnet) for the purpose of multi-class segmentation using non-overlapping single-class datasets for training. Our approach allows the model to share implicitly all of its parameters by all target classes being modeled. This drives the model to effectively learn the spatial connections between objects of different classes and improve its generalization ability.

To the best of our knowledge, our work is the first to describe the use of conditioning a convnet for the purpose of segmentation and to demonstrate the possibility of producing multi-class segmentations using a single model trained on non-overlapping single-class datasets. The **contributions** of our work are: (1) the first application, to the best of our knowledge, of conditioning a convnet for semantic segmentation; (2) the presented conditioning framework enables an efficient multi-class segmentation with a single model trained on single-class datasets, drastically reducing the training complexity and the total number of parameters, in comparison to separate class-specific models; (3) improved state-of-the-art results (up to 2.7%) on publicly available datasets for the segmentation of liver, spleen, and pancreas with significantly reduced computational cost. Moreover, we demonstrate the applicability of our proposed method to natural images and evaluate it on the Cityscapes dataset [5]. Additionally, we discuss the possible extensions and applications of the proposed approach.

2. Related work

The difficulty of collecting large-scale, carefully annotated datasets for semantic segmentation is well acknowledged [37, 39, 46]. A family of approaches has been proposed for learning to perform segmentation using weakly labeled data. Weak annotations, in the form of image labels [22], points and scribbles [1, 18], bounding boxes [6], and their combinations [30, 41] have been explored for learning image segmentation models. While these works in weakly-supervised segmentation are similar in spirit, they are principally different in comparison to our work. They still assume the availability of annotations of every object from a collection of pre-defined target classes if one is present in an image. With regard to CT images, each slice would require a set of annotations for every target organ present on a slice, be it seeds, bounding boxes or labels. However, single-class datasets do not come with such annotations, and provide details for only one particular class.

Segmentation of anatomical structures, especially abdominal organs, is considered a difficult problem, as they demonstrate a high variability in size, position, and shape (Fig. 1). Various convnet-based segmentation methods have been proposed for abdominal organ segmentation. The majority of these methods that utilize single-class datasets are specialized on the segmentation of a particular organ, such as liver [10, 25] or pancreas [13, 34]. Moreover, these works often describe sophisticated and intricate multi-stage approaches [45]. Some more generally applicable convnet-based methods have been proposed and tested on multiple organs [11]. These methods describe models for the segmentation of individual organs, and the separate segmentations are fused together to produce the final outlines. However, while showing state-of-the-art performance, these models must be trained and applied separately for the segmentation of each organ, which manifests inefficient usage of computational resources and additional training time. Moreover, such separately trained models do not embed the spatial correlations among abdominal organs and thus are likely to be overfitted for each particular single-class

dataset. Additionally, these models often also require pre- and post-processing steps, which complicate and particularize the models even more.

Several studies have been proposed for the simultaneous multi-class, or multi-organ, segmentation of anatomical structures in medical images. The majority of these utilize probabilistic atlases [4, 29, 40] and statistical shape models [28]. These methods require all volumetric images in the training dataset to be registered. This pre-processing step is computationally expensive and often imperfect due to the considerable variations in size, shape, and location of abdominal organs between patients. Recently, a few convnet-based solutions [35] were proposed for simultaneous multi-organ segmentation. However, all such methods were developed and evaluated on publicly unavailable multi-class segmentation datasets. Moreover, the used multi-class datasets were acquired by a single institution and exhibit the same image quality and lack chronic abnormalities. In contrast, we leverage diverse single-class datasets and describe a novel way of conditioning a convnet to develop a multi-class segmentation model of high generalization ability.

Conditioning has been widely used in image synthesis. A family of works [23, 38, 42, 43] on generating images conditioned on certain attributes, such as category or labels, have shown successful and compelling results. Ma *et al*. [26] proposed a framework for person image synthesis based in arbitrary poses. Zhu *et al*. [49] modeled a distribution of potential results of the image-to-image translation. Reed *et al*. [32] demonstrated the synthesis of images given the desired content and its location within the image. However, the area of conditional convnets for semantic segmentation has been left untapped, and no application has been explored. In this paper, we describe a method of conditioning a convnet for the purpose of segmentation, evaluate the method on the segmentation of abdominal organs and urban scenes, and discuss a set of other possible applications.

3. Method

As opposed to generating separate models for each object in single-class datasets, we describe a framework that can simultaneously learn multi-class knowledge given a set of single-class datasets. Consider a set of single-class datasets $\{\mathcal{D}_1, ..., \mathcal{D}_K\}$, where each dataset $\mathcal{D}_k = \{(X^k; Y^{k,c_m})\}, k \in \{1, ..., K\}$ contains a set of input images $X^k = \{x_i^k\}$ and a set of corresponding binary segmentation masks $Y^{k,c_m} = \{y_i^{k,c_m}\}$ of object $c_m \in C, m = 1, ..., M$. Additionally, input images X^k in each dataset $\mathcal{D}_k$ exhibit objects of all classes $c_m \in C$. Moreover, we also assume that datasets $\mathcal{D}_k$ do not have the same pairs of $\{(X^k; Y^{k,c_m})\}$, such as $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset, \forall i, j$, and each dataset might have different number of classes. These assumptions greatly relax the initial conditions and attempt to make the description of the problem more general and challenging. The goal is to predict a set of segmentation masks $\{\hat{y}^{c_m}\}, \forall c_m \in C$, given an unseen input image $\hat{x}$.

3.1. Base model

The base component of the proposed framework is a 3D fully-convolutional U-net-like architecture, such as an encoder-decoder with skip connections (Fig. 2a). Additionally, we adopt 3D densely connected convolutional blocks [17, 19], which effectively utilize the volumetric information available in the CT scans. More formally, the model includes densely-connected units of a composite function $F_l(\cdot)$, and the output $\mathbf{x}_l$ of the l^{th} layer is defined as

$$\mathbf{x}_l = F_l([\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{l-1}]), \quad (1)$$

where $[...]$ is a concatenation operation of the feature maps from previous layers. In our experiments, $F_l(\cdot)$ is defined as a leaky rectified linear unit (LReLU [27]) with $\alpha = 0.3$, followed by a $3 \times 3 \times 3$ convolution. The encoder part of the model includes a convolutional layer, followed by six densely connected convolutional blocks, sequentially connected via $2 \times 2 \times 2$ maxpooling layers. The number of feature channels in each dense block is proportional to its depth. The decoder part of the model utilizes transposed convolutions with strides as upsampling layers and is topologically symmetric to the encoder. The last convolutional layer ends with a sigmoid function. See the supplementary material for more details.

3.2. Conditioning

Unlike classic approaches of training separate models for each class $c_m \in C$, our framework is able to infer the segmentations and the relationships of multiple classes from single-class datasets and to learn to generate segmentations for all classes c_m with a single model. To introduce such ability to the model, we propose a novel way of conditioning the base convolutional model with a target class c_m that needs to be segmented. While certain ways of conditioning have been widely used in generative adversarial nets (GANs) [8, 26, 32] for image synthesis, to the best of our knowledge, there have been no attempts to condition a convnet for segmentation.

One of our goals was to keep the base model fully-convolutional, simple, and efficient in order to avoid additional overhead that could negatively affect the performance. To achieve this, we propose to incorporate the conditional information as a part of the intermediate activation signal after performing convolutional operations and before applying nonlinearities. While some examples of conditioned GANs [32] suggest to learn the conditional function, we propose a more computationally efficient approach for the task of segmentation. Specifically, we propose to use the following function:

$$\varphi(c_m, H_j, W_j, D_j) = \boldsymbol{O}^{H_j \times W_j \times D_j} \odot hash(c_m), \quad (2)$$

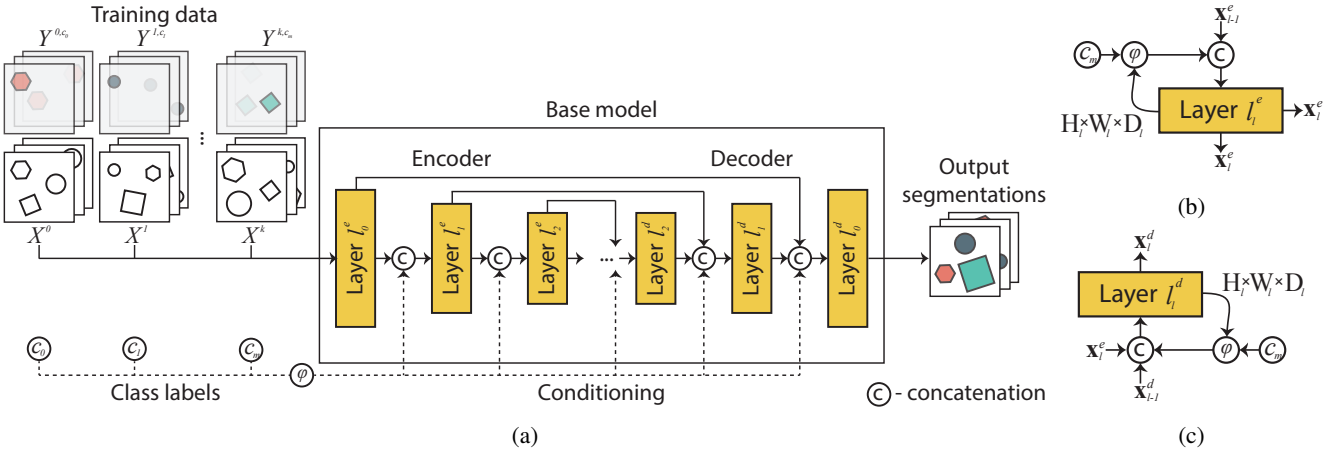

Figure 2: A schematic overview of the proposed framework for conditioning a convnet to perform multi-class segmentation using only single-class datasets during training. (a) The base model uses images from the k single-class datasets and after conditioning on class labels, produces the final segmentation masks. The conditioning can be done for either (b) encoder or (c) decoder layers of the base model, or both.

where $\odot$ is an element-wise multiplication, $\boldsymbol{O}^{H_j \times W_j \times D_j}$ is a tensor of size $H_j \times W_j \times D_j$ with all elements set to 1, and $hash(\cdot)$ is a hash function for a pre-defined lookup table. That is, the function $\varphi(c_m, H_j, W_j, D_j)$ creates a tensor of size $H_j \times W_j \times D_j$ with all values set to $hash(c_m)$. Therefore, the proposed conditioning of the l^{th} layer with input $\mathbf{x}_l$ of size $H_l \times W_l \times D_l$ is defined as

$$\mathbf{x}_l = [\mathbf{x}_{l-1}, \varphi(c_m, H_l, W_l, D_l)] \tag{3}$$

where $\mathbf{x}_{l-1}$ is the output of the previous layer (Fig. 2b, 2c). It is important to note that the proposed conditioning does not depend on the possible attributes of the classes, such as location, shape, etc. It is done to increase the generalization ability of the proposed framework.

During training time, the network is trained on pairs $(x_i^k; y_i^{k,c_m})$ that are randomly sampled from different datasets $\mathcal{D}_k$, while being conditioned on the corresponding class c_m of the binary ground truth segmentation mask y_i^{k,c_m}. During the inference time, the network is sequentially conditioned on all $c_m \in C$ to generate segmentations masks $\{\hat{y}^{c_m}\}$ for all objects in the input image $\hat{x}$. While such an approach of using a pre-defined lookup table maintains the simplicity and austerity of the framework without additional variables to be trained, it also has some practical benefits. In particular, in the event of adding a new target segmentation class c_{M+1}, the framework will only require a new entry to the lookup table and a simple fine-tuning, unlike the more expensive re-training expected if we had learned the conditional function.

However, a natural question arises: given a deep convnet with L layers, where is the best place to perform the conditioning? Conditioning of which layers is the most beneficial? We hypothesize that given an encoder-decoder like architecture, one should expect better performance when the conditioning is done on the layers in the decoder, which could use the provided conditional information and the low-level information present in the encoder feature maps to map them to higher levels within the network. Moreover, we expect that the conditional information directly accessible to multiple layers will make the optimization easier. In Section 4.1, we test our hypothesis and report the performance for a variety of conditioning settings.

4. Experiments

In this section, we describe an extensive analysis of our framework, experiment with different kinds of loss functions and various ways of conditioning, and compare the results to the solutions, which were individually customized for each single-class dataset or designed for multi-class datasets. We show that our conditioned multi-class segmentation framework outperforms current state-of-the-art single-class segmentation approaches for biomedical images. Additionally, we demonstrate the applicability of the proposed approach for the segmentation of urban scenes.

Datasets To evaluate the proposed framework and to test our hypotheses, our work utilizes three datasets of abdominal CT volumes. Particularly, we use 20 volumes of the publicly available Sliver07 dataset [15] of liver segmentations, 82 volumes of the publicly available NIH Pancreas dataset [16] of pancreas segmentations, and 74 volumes from our additional dataset of liver and spleen segmentations. Therefore, in our experiments, $c_m \in C = \{liver, spleen, pancreas\}$. The segmentation masks in the latter dataset have been binarized and stored as separate single-class files. Examples of the CT images and the corresponding ground-truth segmentation masks are illustrated in Fig. 1b, 1c, 1d and Fig. 6 (first column). Following a

common strategy, each dataset was divided into training and testing sets with ratio of 80/20. The size of the volumes in each dataset was $512 \times 512 \times Z_0$, where Z_0 is the number of axial slices. Each dataset was collected at different institutions with different scanners and protocols and incorporates volumes of various inter-slice spacings and, moreover, exhibits various pathologies, such as hepatic tumors and cases of splenomegaly. Such diversity in the datasets allows us to test the proposed approach in a challenging setting.

The input images have been minimally preprocessed: each dataset was sampled with an equal probability, and subvolumes of size $256 \times 256 \times 32$ have been extracted and normalized to create input images. Additionally, all training examples have been augmented with small random rotations, zooms, and shifts.

Training The proposed framework was trained on examples from all used single-class datasets. The framework was optimized with the following objective:

$$\mathcal{L}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \alpha_1\beta_1\mathcal{L}_1(Y^{c_1}, \hat{Y}^{c_1}) + ... + \alpha_n\beta_k\mathcal{L}_k(Y^{c_m}, \hat{Y}^{c_m}), \tag{4}$$

where $\mathcal{L}_i(Y^{c_i}, \hat{Y}^{c_i})$ is a loss function for a single-class dataset $\mathcal{D}_i$, the hyperparameters α_i specify the impact of a particular class c_i on the total loss, and $\beta_i = \{0, 1\}$ specifies the presence of the binary mask for class c_i in the batch.

Inference During the inference time, one can manually specify the target segmentation class c_i. However, to simplify the use of the framework during the inference time, we suggest to automate the process of specifying the target segmentation class by iteratively going through all the entities in the lookup table. Alternatively, specifically for segmentation of abdominal organs, a set of presets can be defined, such as liver and gallbladder, which are often analyzed together by clinicians.

Implementation The proposed framework was implemented using Keras library with TensorFlow backend. We trained our network from scratch using Adam optimizer [21] with the initial learning rate or 0.00005, and $\beta_1 = 0.9, \beta_2 = 0.999$, with a batch size of 2 for 25K iterations.

4.1. Ablation experiments

The predicted segmentation masks are binarized by thresholding them at 0.5. To measure the similarity between binary segmentation masks Y and $\hat{Y}$, we use the common Dice Similarity Coefficient (DSC) metric, which is defined as $DSC(Y, \hat{Y}) = \frac{2\sum Y \odot \hat{Y}}{\sum Y + \sum \hat{Y}}$. We compare our results against the current state-of-the-art segmentation methods, which are proposed specifically for single-class segmentation and are tailored for a particular class. In particular, we compare against the work by Zhou *et al.* [48], which described a two-step coarse-to-fine convnet-based solution for pancreas segmentation, and yielded 82.4% DSC on the NIH Pancreas [16] dataset. We also compare against another convnet-based segmentation work by Yang *et al.* [44], which showed 95% DSC on a private datasets of 1000 CT images of liver. Finally, we compare our results against the two-stage coarse-to-fine multi-organ convnet-based solution by Roth *et al.* [35], which was evaluated on a private multi-class dataset and resulted in 95.4%, 92.8%, and 82.2% DSC for liver, spleen, and pancreas, respectively.

In all experiments described in this section we set $\alpha_i = 1$ and use the DSC-based loss function:

$$\mathcal{L}_i(Y^{c_i}, \hat{Y}^{c_i}) = 1 - \frac{2\sum Y^{c_i} \odot \hat{Y}^{c_i}}{\sum Y^{c_i} + \sum \hat{Y}^{c_i}}. \tag{5}$$

Additionally, we experimented with the binary cross-entropy loss function, which showed significantly worse performance.

We begin our experiments by analyzing the performance of our base model trained separately for each class c_m without the use of conditioning. We refer to this experiment as `indivs` and the learning curves for each model are illustrated in Fig. 3a. We observe that the models failed to get close to the state-of-the-art performance during the first 25K iterations.

Next, we test a naive approach of training a single model on single-class datasets to produce reasonable multi-class segmentation results by predicting a volume of the same dimensions but with three additional channels, each for each class c_m, such as liver, spleen, and pancreas. We refer to this experiment as `no cond` and the learning curves are illustrated in Fig. 3b. The results show that the training does not converge, which was expected and can be explained by the fact that the model struggles to infer multi-class segmentations from the inconsistent binary masks in the training examples. Additionally, this approach is memory-bounded, especially for high-resolution images and volumes, and only a small number of classes can be modeled this way. The examples of the segmentations produced by the `no cond` model can be found in the Appendix.

The next experiments describe the results of the conditioned model. In the experiment `cond-2nd`, we test a simple way of conditioning a model by providing the conditional information as the second channel of the input volume. Particularly, we predefine a lookup table of conditioning variables for each c_m with random real values sampled from [-1, 1]. Specifically, each training 3D subvolume has been augmented in the second channel with a volume of the same size with all elements set to $hash(c_m)$. The learning curves illustrated in Fig. 3c show that the model was able to utilize the provided conditional information and learn to generate multi-class segmentations. However, similarly to the experiment `cond-enc` (see Fig. 3d), where each dense block in the encoder had direct access to the conditional information, the model shows adequate performance but struggles to outperform state-of-the-art approaches. How-

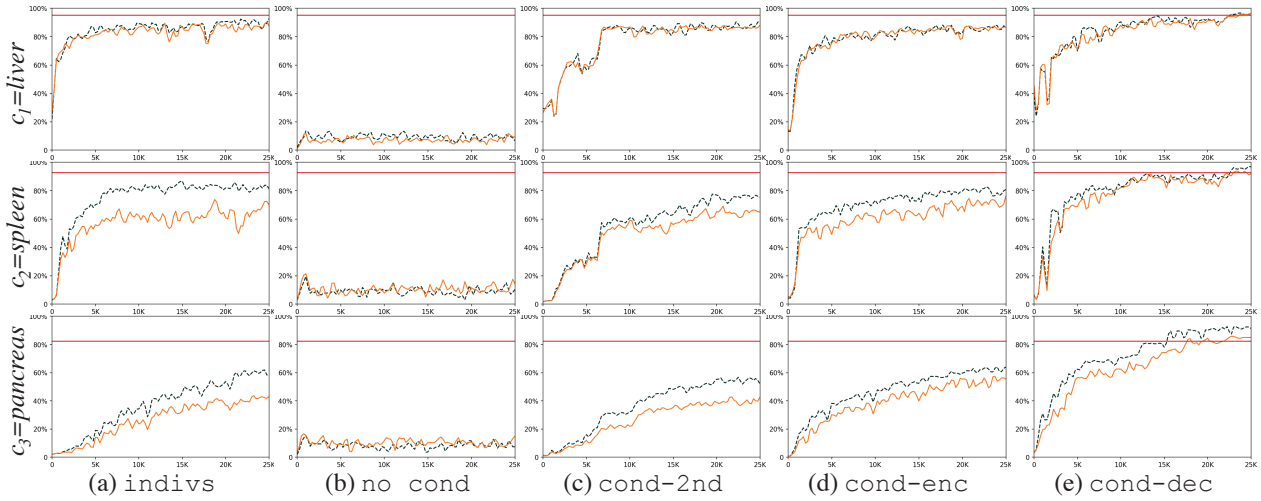

Figure 3: Training curves of various conditioning models generated for each $c_m \in C = liver, spleen, pancreas$ during the first 25K iterations (x-axis). The dashed green line denotes training accuracy (DSC, %) (y-axis), the solid orange line denotes testing accuracy, and the solid red line denotes the current state-of-the-art results.

ever, we notice significantly better generalization performance in these models trained jointly on different datasets while improving training and testing accuracies, compared to `indivs` models trained separately on each dataset.

Finally, we experiment with conditioning the decoder part of the base model. We refer to this experiment as `cond-dec`. The learning curves illustrated in Fig. 3 validate our hypothesis and show a superior segmentation performance. The training in this experiment converges faster than in the other experiments. In addition to outperforming both meticulously tailored solutions for single-class segmentation and multi-class segmentation solutions designed on private datasets (see Table 1), our framework also shows significant generalization ability. Examples of the segmentation results for this experiment are illustrated in Fig. 6. We observe that the model accurately delineates all the target objects even in a difficult case illustrated in Fig. 6 (last row), where due to the imaging protocol all of the organs, besides being congested together, also have similar intensities and their boundaries are hard to differentiate. The reason for such accurate segmentations by this model can be due to (1) a high degree of implicit parameter sharing between all classes being modeled, and (2) the ability of the decoder path to capitalize on the available conditional information and gradually recover the spatial information and sharp boundaries of the target classes.

We also performed additional experiments on conditioning the decoder and encoder at the same time and studied the effects of conditioning only parts of the decoder at various depths. Such approaches yielded no benefits, and the performance in these experiments was inferior compared to when the conditional information was available directly to each layer in the decoder.

Importance of spatial connections between classes To test our hypothesis and to explore the importance of the spatial correlation between classes on the model's performance, we evaluate our `cond-dec` model on corrupted images. For CT images in particular, we compare the baseline performance (Table 1) to the performance on images where different classes were corrupted by randomly replacing 70% of the corresponding voxels with intensity values common for fatty tissue between organs. An example of a corrupted image for $spleen$ is illustrated in Fig. 4. Interestingly, the separate corruption of classes $spleen$ and $pancreas$ had practically no effect on the accuracy of the $liver$ segmentation, which only degraded within the 2% range. However, both the segmentations of $spleen$ and $pancreas$ were significantly affected when other organs were corrupted, dropping the performance on average by 15.3% compared to the baseline. We believe this supports our hypothesis that the model learns and utilizes the spatial correlations between target classes during the inference, and the deprivation of these correlations degrades the performance.

Applicability to natural images The described conditioning technique was developed with the goal of being universally applicable rather than being limited to medical images. To demonstrate the applicability of our method to other domains, we train a model for semantic segmentation of natural images. While datasets of natural images are generally multi-class – i.e., multiple objects in an image are annotated – we believe the validation of our framework on natural images datasets is valuable. We evaluate our method using the challenging urban scene understanding dataset Cityscapes [5]. It contains 2,975 finely-annotated training, 500 validation, and 1,525 test images of 1024×2048 resolution with 19 semantic classes. Additionally, the dataset

Figure 4: An example of a corrupted image, where 70% of *spleen* voxels were replaced with intensity values common for fatty tissue between organs.

Table 1: The comparison of segmentation accuracy (mean DSC, %) for different models for the segmentation of liver, spleen, and pancreas (higher is better).

Model	Liver	Spleen	Pancreas
Yang *et al.* [44]	95.0	-	-
Zhou *et al.* [48]	-	-	82.4
Roth *et al.* [35]	95.2	92.8	82.2
`indivs`	91.5	74.4	42.9
`no cond`	14.7	21.8	18.6
`cond-2nd`	89.7	71.7	44.4
`cond-enc`	88.1	76.9	57.3
`cond-dec`	**95.8**	**93.7**	**85.1**

comes with 20,000 coarsely-annotated training images, although these were not used in this experiment. We selected three essential classes: *road*, *car*, and *person*. To imitate single-class datasets, each multi-class annotation image was converted into a set of three binary masks. We use the same base model described in Section 3.1, but with 2D convolutions and max-pooling layers, and we condition only the decoder part of the model. In addition, to test the sensitivity of the model to c_m values, the lookup table was predefined with values sampled from [-20, 20]. Each image was resized to 512×1024, and the dataset was augmented with random left and right flips and brightness perturbations. The model was trained for 40K iterations using a mini-batch of 4 and a DSC-based loss function (Equation 5). The results were evaluated in terms of class-wise intersection over union (IoU) metric for test images upsampled to the original resolution and are presented in Table 2. Examples of the results are illustrated in Fig. 5. Our model achieves performance close to the state-of-the-art solutions [2, 3, 47] on some classes, without pre-training or post-processing steps, and using only finely-annotated data. While updating the state-of-the-art on this dataset was not our goal in this experiment, given that it is a multi-class dataset, we believe that pre-training the base model on datasets, such as Synthia [33], and using additional annotated data, can improve the

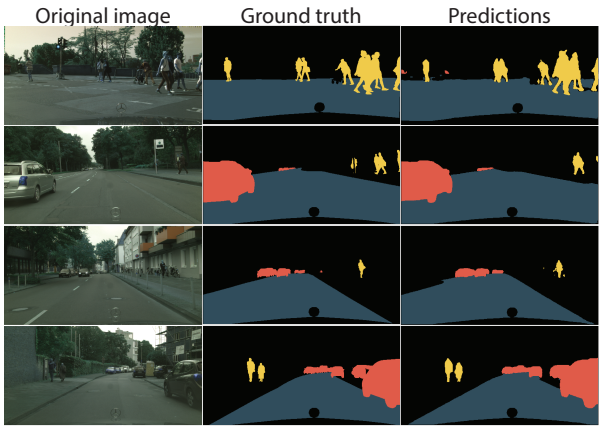

Figure 5: Examples of segmentation results for the Cityscapes validation set.

Table 2: The comparison of segmentation accuracy (per class IoU, %) on Cityscapes test set for different classes (higher is better).

Model	Road	Car	Person
Chen *et al.* [2]	98.7	96.5	88.2
Chen *et al.* [3]	98.6	96.3	87.6
Zhao *et al.* [47]	98.7	96.2	86.8
`cond-dec`	96.4	91.0	76.2

performance of our method on this dataset, as was shown in other works [2, 3, 36, 47].

5. Discussion

In this paper, we described a framework for learning multi-class segmentations from single-class datasets by a novel way of conditioning a convnet for the purpose of multi-class segmentation. We performed an extensive experimental evaluation of the various ways of conditioning the model and found that providing each layer in the decoder a direct access to the conditional information yields the most accurate segmentation results. The proposed framework was evaluated on the task of segmentation of medical images, where the problem of single-class datasets naturally arises. While being significantly more computationally efficient, the method outperforms current state-of-the-art solutions, which were specifically tailored for each single-class dataset. Additionally, we demonstrated the applicability of our method to the semantic segmentation of natural images using the Cityscapes dataset.

While our work has been validated using radiological CT scans and natural images of urban scenes, our idea can be easily expanded to other applications in various domains. In particular, one can imagine how our framework can be applied for the detection of cancer metastases in pathology

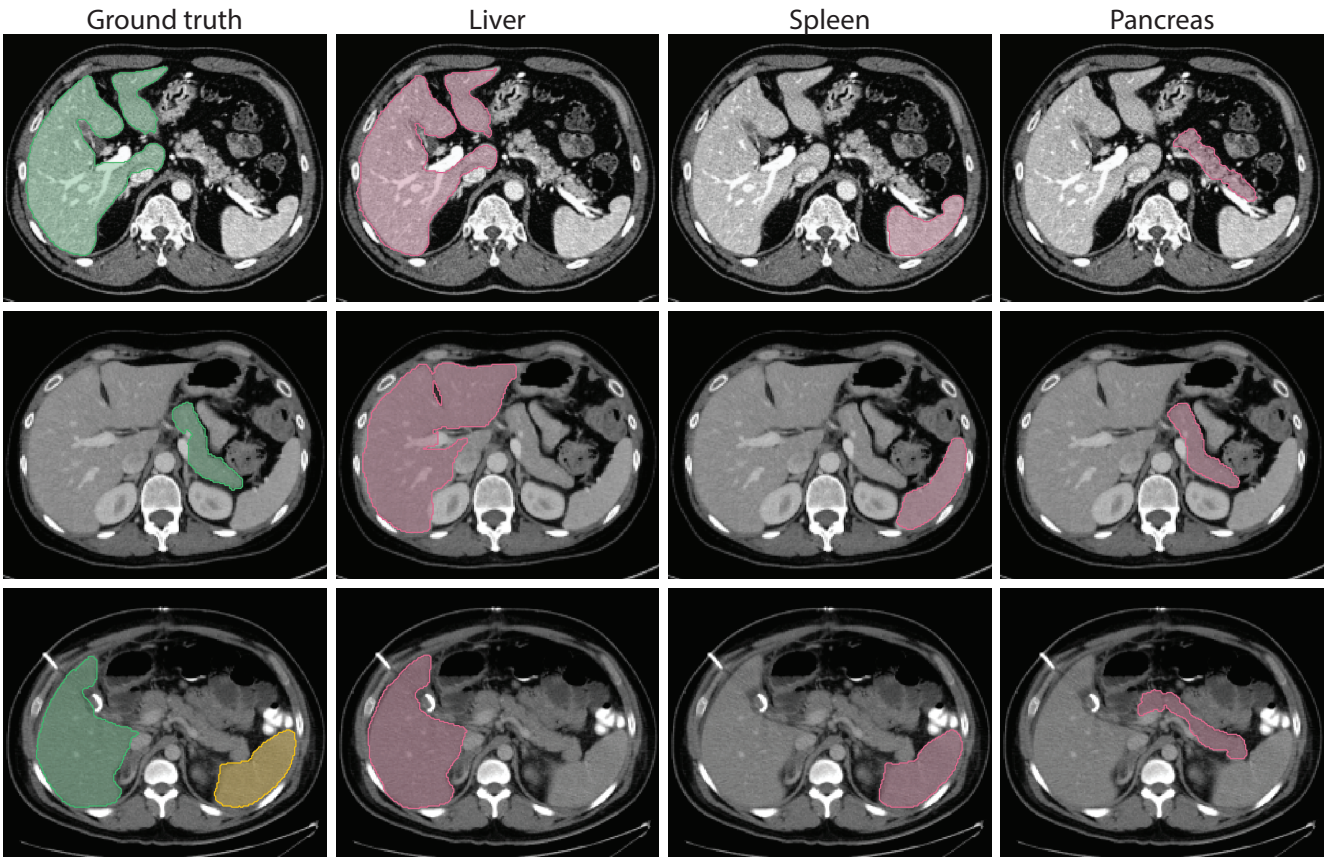

Figure 6: Examples of segmentation results for CT images from different datasets generated by the `cond-dec` model. The results are presented in 2D for illustrative purposes, but actual results are 3D. Rows from top to bottom: Sliver07 [15], NIH Pancreas [16], our own additional dataset of liver and spleen segmentations. From left to right: available ground truth outlines in the datasets (green and yellow), and segmentation results conditioned on each $c_m \in C = \{liver, spleen, pancreas\}$, which are outlined in purple. Although the segmentation outlines for our additional dataset are shown together (green and yellow), they were generated and stored separately in a form of binary masks.

images. Pathology datasets show similar fragmentation – a unified database of pathology images of various biological tissues, such as brain or breast, currently does not exist and research focuses on separate subproblems. Similarly to our experiments, a convnet can be conditioned on the target type of metastasized cancel cells in different tissue samples. Moreover, one can also imagine similar applications of conditioning a convnet for the purpose of instance-level segmentation, where each instance can be conditioned on certain attributes, such as size, color, etc, or something more sophisticated, such as species or kind. Furthermore, Rebuffi *et al*. [31] have described a method of learning data representations in multiple visual domains for the purpose of classification. Our framework can augment such works for the purpose of segmentation.

Our future work will focus on expanding the problem and incorporating images from different domains. Particularly for radiological images, it will be interesting to see if a model trained on a mixture of CT and MRI images will be able to infer and transfer classes marked in one imaging modality to another.

Acknowledgements

We thank Le Hou and Dimitris Samaras for their valuable input and advice. This work has been partially supported by National Science Foundation grants NRT1633299, CNS1650499, the Marcus Foundation, and National Heart, Lung, and Blood Institute of NIH under Award Number U01HL127522. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Additional support was provided by the Center for Biotechnology, a New York State Center for Advanced Technology; Stony Brook University; Cold Spring Harbor Laboratory; Brookhaven National Laboratory; the Feinstein Institute for Medical Research; and

the New York State Department of Economic Development under Contract C14051.

References

[1] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. Whats the point: Semantic segmentation with point supervision. *Proc. of European Conference on Computer Vision (ECCV)*, pages 549–565, 2016.

[2] Liang-Chieh Chen, Maxwell D Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jonathon Shlens. Searching for efficient multi-scale architectures for dense image prediction. *arXiv preprint arXiv:1809.04184*, 2018.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[4] Chengwen Chu, Masahiro Oda, Takayuki Kitasaka, Kazunari Misawa, Michitaka Fujiwara, Yuichiro Hayashi, Yukitaka Nimura, Daniel Rueckert, and Kensaku Mori. Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images. *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 165–172, 2013.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[6] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pages 1635–1643, 2015.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[8] Emily L Denton, Soumith Chintala, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 1486–1494, 2015.

[9] Konstantin Dmitriev, Arie E Kaufman, Ammar A Javed, Ralph H Hruban, Elliot K Fishman, Anne Marie Lennon, and Joel H Saltz. Classification of pancreatic cysts in computed tomography images using a random forest and convolutional neural network ensemble. *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 150–158, 2017.

[10] Qi Dou, Hao Chen, Yueming Jin, Lequan Yu, Jing Qin, and Pheng-Ann Heng. 3D deeply supervised network for automatic liver segmentation from CT volumes. *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 149–157, 2016.

[11] Michal Drozdzal, Gabriel Chartrand, Eugene Vorontsov, Mahsa Shakeri, Lisa Di Jorio, An Tang, Adriana Romero, Yoshua Bengio, Chris Pal, and Samuel Kadoury. Learning normalized inputs for iterative estimation in medical image segmentation. *Medical image analysis*, 44:1–13, 2018.

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[13] Amal Farag, Le Lu, Holger R Roth, Jiamin Liu, Evrim Turkbey, and Ronald M Summers. A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling. *IEEE Transactions on Image Processing*, 26(1):386–399, 2017.

[14] Michael Götz, Christian Weber, Bram Stieltjes, Klaus Maier-Hein, and K Maier. Learning from small amounts of labeled data in a brain tumor classification task. *Proc. of Neural Information Processing Systems (NIPS)*, 2014.

[15] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, 2009.

[16] Roth Holger, Farag Amal, Turkbey Evrim, Lu Le, Liu Jiamin, and Summers Ronald. Data from pancreas – CT. *Cancer Imaging Archive*, 2016.

[17] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1(2):3, 2017.

[18] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7014–7023, 2018.

[19] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1175–1183, 2017.

[20] Oscar Jimenez-del Toro, Henning Müller, Markus Krenn, Katharina Gruenberg, Abdel Aziz Taha, Marianne Winterstein, Ivan Eggel, Antonio Foncubierta-Rodríguez, Orcun Goksel, András Jakab, et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks. *IEEE Transactions on Medical Imaging*, 35(11):2459–2475, 2016.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. *Proc. of European Conference on Computer Vision (ECCV)*, pages 695–711, 2016.

[23] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. *arXiv preprint arXiv:1705.04098*, 2017.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. *Proc. of European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[25] Fang Lu, Fa Wu, Peijun Hu, Zhiyi Peng, and Dexing Kong. Automatic 3D liver location and segmentation via convolutional neural network and graph cut. *International Journal of Computer Assisted Radiology and Surgery*, 12(2):171–182, 2017.

[26] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 405–415, 2017.

[27] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. *Proc. of ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[28] Toshiyuki Okada, Marius George Linguraru, Masatoshi Hori, Ronald M Summers, Noriyuki Tomiyama, and Yoshinobu Sato. Abdominal multi-organ segmentation from CT images using conditional shape–location and unsupervised intensity priors. *Medical Image Analysis*, 26(1):1–18, 2015.

[29] Bruno Oliveira, Sandro Queirós, Pedro Morais, Helena R Torres, João Gomes-Fonseca, Jaime C Fonseca, and João L Vilaça. A novel multi-atlas strategy with dense deformation field reconstruction for abdominal and thoracic multi-organ segmentation from computed tomography. *Medical Image Analysis*, 45:108–120, 2018.

[30] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. *Proc. of IEEE International Conference on Computer Vision (CVPR)*, pages 1742–1750, 2015.

[31] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 506–516, 2017.

[32] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 217–225, 2016.

[33] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016.

[34] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 556–564, 2015.

[35] Holger R Roth, Hirohisa Oda, Yuichiro Hayashi, Masahiro Oda, Natsuki Shimizu, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. Hierarchical 3D fully convolutional networks for multi-organ segmentation. *arXiv preprint arXiv:1704.06382*, 2017.

[36] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] Tong Shen, Guosheng Lin, Chunhua Shen, and Ian Reid. Bootstrapping the performance of webly supervised semantic segmentation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1363–1371, 2018.

[38] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 4790–4798, 2016.

[39] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9049–9058, 2018.

[40] Robin Wolz, Chengwen Chu, Kazunari Misawa, Kensaku Mori, and Daniel Rueckert. Multi-organ abdominal CT segmentation using hierarchically weighted subject-specific atlases. *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 10–17, 2012.

[41] Jia Xu, Alexander G Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3781–3790, 2015.

[42] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2016.

[43] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. *Proc. of European Conference on Computer Vision (ECCV)*, pages 776–791, 2016.

[44] Dong Yang, Daguang Xu, S Kevin Zhou, Bogdan Georgescu, Mingqing Chen, Sasa Grbic, Dimitris Metaxas, and Dorin Comaniciu. Automatic liver segmentation using an adversarial image-to-image network. *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 507–515, 2017.

[45] Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K. Fishman, and Alan L. Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[46] Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle- and shapeconsistency generative adversarial network. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9242–9251, 2018.

[47] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *Proc.*

of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2881–2890, 2017.

[48] Yuyin Zhou, Lingxi Xie, Wei Shen, Yan Wang, Elliot K Fishman, and Alan L Yuille. A fixed-point model for pancreas segmentation in abdominal CT scans. *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 693–701, 2017.

[49] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 465–476, 2017.

Learning Multi-Class Segmentations From Single-Class Datasets: Supplementary Material

Konstantin Dmitriev and Arie E. Kaufman
Stony Brook University, Stony Brook, NY
kdmitriev@cs.stonybrook.edu

Base model

Below are the details on the architecture of the base model.

	Input
Encoder	Conv 2
Encoder	DenseBlock 16x3 + MaxPooling
Encoder	DenseBlock 32x3 + MaxPooling
Encoder	DenseBlock 64x3 + MaxPooling
Encoder	DenseBlock 128x3 + MaxPooling
Encoder	DenseBlock 256x3 + MaxPooling
Encoder	DenseBlock 512x3 + MaxPooling
Decoder	TransConv 512 + DenseBlock 256x3
Decoder	TransConv 256 + DenseBlock 128x3
Decoder	TransConv 128 + DenseBlock 64x3
Decoder	TransConv 64 + DenseBlock 32x3
Decoder	TransConv 32 + DenseBlock 16x3
Decoder	TransConv 16 + DenseBlock 8x3
Decoder	Conv 32
Decoder	Conv 1
	Output

Table 1. Architecture details of the base model. We utilize 2D or 3D convolutional and transposed convolutional ($TransConv$) layers, depending on the experiment. Each $DenseBlock\ X \times N$ contained N densely connected convolutional layers with X filters each.

References

[1] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, 2009.

[2] Roth Holger, Farag Amal, Turkbey Evrim, Lu Le, Liu Jiamin, and Summers Ronald. Data from pancreas – CT. *Cancer Imaging Archive*, 2016.

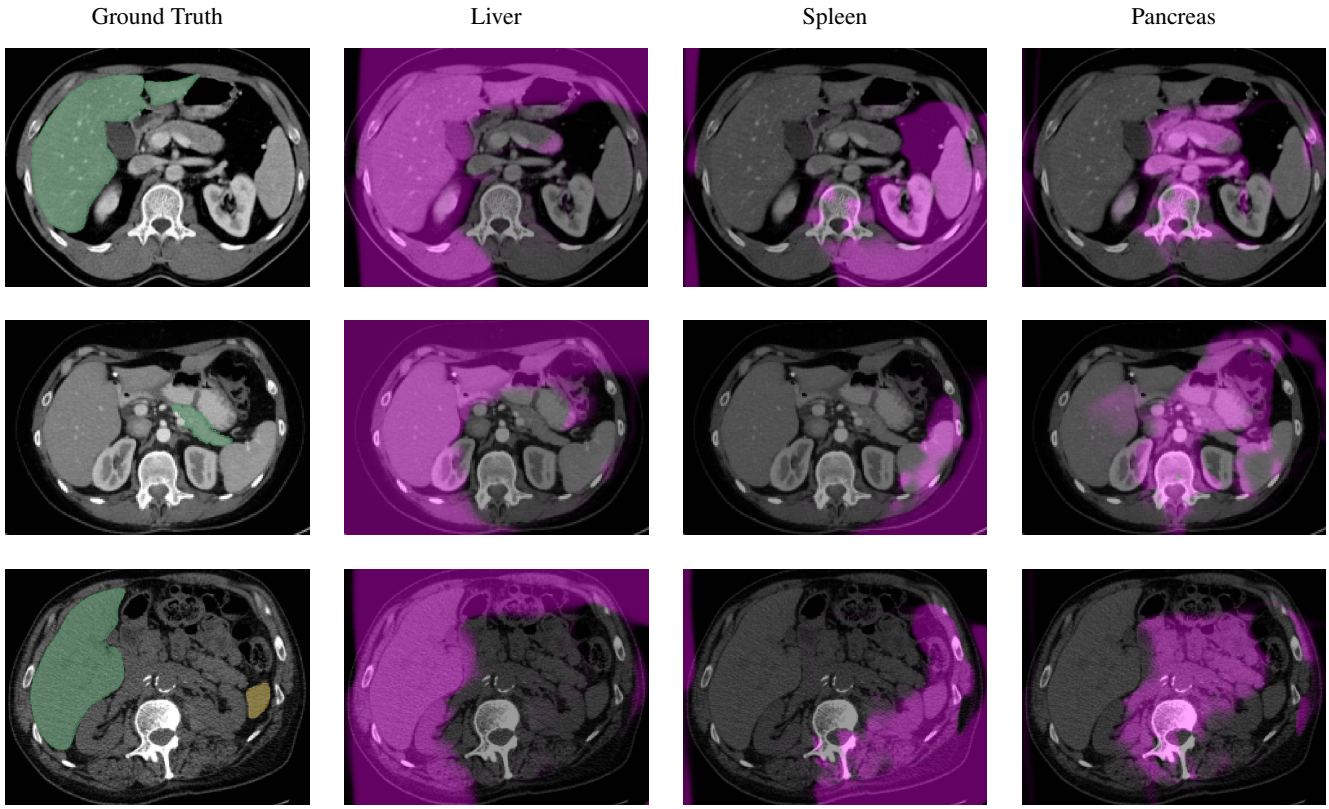

Figure 1: Examples of segmentation predictions for CT images from different testing sets generated by the `no cond` model. The results are presented in 2D for illustrative purposes, but actual results are in 3D. Rows from top to bottom: Sliver07 [1], NIH Pancreas [2], our own additional dataset of liver and spleen segmentations. From left to right: available ground truth outlines in the datasets (green and yellow), segmentation results from each additional channel. Although the segmentation outlines for our additional dataset are shown together (green and yellow), they were generated and stored separately in a form of binary masks.