# Phase Collaborative Network for Multi-Phase Medical Imaging Segmentation

Huangjie Zheng[1], Lingxi Xie[2], Tianwei Ni[3], Ya Zhang[1], Yan-Feng Wang[1], Qi Tian[4]
Elliot K. Fishman[5], Alan L. Yuille[2]
[1]Shanghai Jiao Tong University [2]The Johns Hopkins University
[3]Peking University [4]Huawei Noahs Ark Lab [5]The Johns Hopkins School of Medicine
{zhj865265, ya_zhang, wangyanfeng}@sjtu.edu.cn {198808xc, twni2016, alan.l.yuille}@gmail.com
tian.qi1@huawei.com efishman@jhmi.edu

## Abstract

*Integrating multi-phase information is an effective way of boosting visual recognition. In this paper, we investigate this problem from the perspective of medical imaging analysis, in which two phases in CT scans known as arterial and venous are combined towards higher segmentation accuracy. To this end, we propose Phase Collaborative Network (PCN), an end-to-end network which contains both generative and discriminative modules to formulate phase-to-phase relations and data-to-label relations, respectively. Experiments are performed on several CT image segmentation datasets. PCN achieves superior performance with either two phases or only one phase available. Moreover, we empirically verify that the accuracy gain comes from the collaboration between phases.*

## 1. Introduction

Semantic segmentation is one of the fundamental problem in computer vision which implies a wide range of applications. Recent years, with the rapid development of deep learning [23, 22, 37, 14], researchers designed a series of segmentation models [29, 5, 4, 6, 8] which are mostly equipped with an encoder-decoder architecture, and have been verified effective in various image domains.

Medical imaging analysis especially organ segmentation is an important prerequisite of computer-assisted diagnosis [33, 34, 49, 47]. In this field, images can appear in more than one *phases*, each of which corresponds to a specific way of data sampling and/or scanning. It has been well acknowledged that incorporating multi-phase information improves visual recognition [35, 45], but nevertheless, this problem has been fewer studied before. There are two possible reasons – One of them lies in the lack of multi-phase training data, and the other refers to the difficulty in aligning multi-phase data and digging complementary information out from them. A typical example is shown in Figure 1.

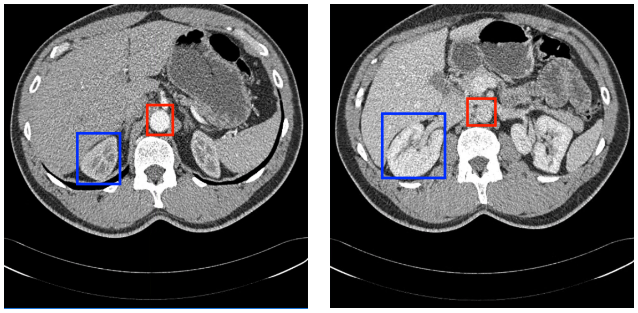
(a) An arterial-phase CT scan (b) A venous-phase CT scan

Figure 1: (*Best viewed in color*) An illustration of the difference between CT scans from arterial (**left**) and venous (**right**) phases. Vessels (marked in red) have larger intensities in the arterial phase while organs (*e.g.*, a *kidney* marked in blue) and soft tissues have larger intensities in the venous phase. These differences are mainly due to the different physical properties of these targets.

Mathematically, in a multi-phase image dataset, each sample is equipped with a semantic annotation $Y$ and a few images $\{X_A, X_B, \ldots\}$, and the goal is to infer $Y$ from all these image data. A straightforward solution is to train a discriminative model $Y = f(X_A, X_B, \ldots)$, but we encounter a major difficulty. In real-world scenarios, *e.g.*, abdominal CT scans, different input images may correspond to different annotations. Thus, the problem becomes inferring $Y$ or any of its instantiation (*e.g.*, $Y_A$) from a few paired data $(X_A, Y_A), (X_B, Y_B), \ldots$, and the major challenge is to build relations between different phases.

To address this issue, we propose an approach named Phase Collaborative Network (PCN). This is an end-to-end framework which formulates the joint distribution of both image data and semantic annotations. The major contrition of this work lies in decomposing this distribution into two parts, namely, a data-to-label relation and a phase-to-phase relation. In practice, the former term is implemented as a discriminative model (*e.g.*, a segmentation network), and the latter one as a generative model (*e.g.*, a GAN which can transfer image style across different phases). PCN, with generators in the back-end and discriminators in the front-end, can be optimized in an end-to-end manner. PCN is evaluated on two sources of data, including our own two-phase datasets and two public single-phase datasets. In the latter scenario, PCN takes advantage of a generator pre-trained from extra unlabeled data, and outperforms the state-of-the-arts using merely single-phase information.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work, and Section 3 describes the Phase Collaborative Network. Experiments are shown in Section 4, and the conclusions drawn in Section 5.

## 2. Related Work

**Semantic segmentation** is a critical problem in computer vision. Conventional methods refer to the graph-based methods [1] and handcrafted local features [42], *etc.* have trend to be replaced by techniques from deep learning, which are typically deep neural networks that can produce higher segmentation accuracy [29, 6]. As various deep network architectures have been proposed [23, 22, 37, 14], the segmentation networks have become more robust and thus been applied to more tasks like video-based segmentation, instance segmentation [31, 12, 30, 36, 39] and more types of data like 3D data [18, 32], *etc.* As the segmentation networks can be extended to more and more tasks, researchers also attempt to apply segmentation networks in medical imaging analysis, where the medical images differ from natural images in that data appear in a volumetric form.

**Medical imaging analysis** is an important pre-requisite of the computer-aided diagnosis (CAD) that can assist human doctors in clinical scenarios. Since medical images contain enormous information such as internal organs, bones, soft tissues vessels, *etc.*, the automatic segmentation of organs or soft tissues from CT volumes is critical for further diagnosis [3, 40, 13, 48]. Researchers often design individualized algorithms in order to capture specific properties of different organs, *e.g.* liver [25, 15], spleen [27], kidney [24], lungs [17] and pancreas [7, 49, 47], *etc.* These methods focus on single phase data and compared to human doctors, who often refer to multi-phase data, face a bottleneck in information. Thus, some studies attempt to bring in multi-phase information to improve segmentation performance [26, 44]. However, due to the limitation of CT scanning process, using multi-phase in segmentation networks faces the problem of alignment between arterial and venous phases. Although there are useful techniques in label fusion [2, 43, 41], they yet face difficulty in fusing data from different phases. The intrinsic problem is there should exist paired arterial and venous data, while we cannot obtain the complete dataset. Thus, we are inspired to model the relation between arterial and venous dataset, so as to achieve multi-phase segmentation.

**Deep generative models** aim to use a parametric distribution to fit the real data distribution in a unsupervised way. By modeling the data distribution, we can achieve the goal of generation. In recent years, generative models like VAE and GAN [20, 11] and their extensions have become popular and have been applied to various scenarios due to their impressive performance. Since the arterial and venous phases form different data distribution, to build a relation between them refers to the domain adaptation using generative models, where Pix2Pix [19], CycleGAN [50] and UNIT [28] *etc.* are typical models in this field. In this research field, the images are considered to consist of elements marked as "content" and "style", and these models focus on style transfer [10, 46]. In our work, we consider the organs as content and their intensity, *etc.* as style, thus to build the relation between original and latent data between phases, we bring in the idea of domain adaptation, which is critical for success in new, unseen environments. Studies like [16] address the scenario where we are provided source data, source labels, and target data to predict target label. In this scenario, we have data and label in both phases, but need to transfer between phases with spatial alignment.

## 3. Phase Collaborative Network

### 3.1. Problem and Motivation

We consider semantic segmentation in the context that multi-phase information is available. Mathematically, we are provided with image data and semantic labels in two phases $A$ and $B$: $\mathcal{A} = \{X_A \in \mathbb{R}^n, Y_A \in \{0,1\}^n\}$ and $\mathcal{B} = \{X_B \in \mathbb{R}^n, Y_B \in \{0,1\}^n\}$. An example comes from medical imaging analysis, which, as shown in Figure 1, may assign different labels in different phases, *i.e.*, $Y_A \neq Y_B$. Thus, a straightforward approach is to learn separate models to deal with each phase. Let the models be $f_A(X_A;\theta_A)$ and $f_B(X_B;\theta_B)$, then the goal of optimization can be formulated as $\theta_A^\star = \arg\min_\theta \|f_A(X_A;\theta) - Y_A\|$ and $\theta_B^\star = \arg\min_\theta \|f_B(X_B;\theta) - Y_B\|$, respectively.

In this paper, we take a better strategy, which considers two phases simultaneously, so that information learned from one phase can assist prediction in the other. This is confirmed by the entropy inequalities [9], $\mathcal{H}(X_A, X_B) \geqslant \max\{\mathcal{H}(X_A), \mathcal{H}(X_B)\}$, where $\mathcal{H}$ indicates the Shannon entropy. The key is to build relations between two phases so that they both benefit from complementary information. Therefore, our goal is to learn a model $f(\cdot)$ that integrates information from both phases: $\theta^\star = \arg\min_\theta \|f(X_A, X_B;\theta) - Y_A\| + \|f(X_A, X_B;\theta) - Y_B\|$.

As the two terms for optimization are symmetric, we will discuss the one related to phase $A$ in the following parts.

## 3.2. The Collaboration between Multiple Phases

To exploit information from both phases, we model the relation between $X_A$, $X_B$ and $Y_A$ as follow:

$$p_{\text{data}}(X_A, X_B, Y_A) = \underbrace{p_{\text{data}}(X_A, Y_A)}_{\text{data-to-label}} \cdot \underbrace{p_{\text{data}}(X_B|X_A, Y_A)}_{\text{phase-to-phase}}.$$

The first term $p_{\text{data}}(X_A, Y_A)$ indicates the relation between the data and label in phase $A$. The second term $p_{\text{data}}(X_B|X_A, Y_A)$ indicates the relation between phase $A$ and $B$. We present the general relation between them using probabilistic graphical model [21] in Figure 2. We can directly model the first term using a segmentation model since they are observed. However, the relation between $X_B$ and $Y_A$ is difficult to model, as $Y_A \neq Y_B$. We conjecture there exists a latent variable $X_B^{(A)} \sim p_{\text{data}}(X_B|X_A, Y_A)$ that carries information from phase $B$, but conditioned on $X_A, Y_A$. To model this relation, a natural solution is to use domain adaptation for generation. Unlike standard domain adaptation methods which model $p_{\text{data}}(X_B|X_A)$ for image-to-image translation, here we need to model $p_{\text{data}}(X_B|X_A, Y_A)$ since $X_B^{(A)}$ has a dependency on both $X_A$ and $Y_A$. Thus, we propose a domain adaptation method that addresses both image translation and spatial alignment to model the phase-to-phase relation. We will present how to model these relations in details in following parts.

### 3.2.1 Building Data-to-Label Relations

As shown in Figure 2, the data-to-label relation involves paired data $X_A, X_B^{(A)}$ and label $Y_A$. To train a model that minimizes $\|f(X_A, X_B^{(A)}; \theta) - Y_A\|$, we suppose two segmentation models, $\mathbf{S}_A : X_A \to Y_A$, $\mathbf{S}_B : X_B \to Y_B$. Given an image $x_A$ and its label $y_A$ in phase $A$, the segmentation loss is:

$$\begin{aligned}&\mathcal{L}_{\text{seg}}(\mathbf{S}_A, \mathbf{S}_B, x_A, y_A)\\&= \lambda\mathcal{L}_{S_A}(\mathbf{S}_A, x_A, y_A) + (1-\lambda)\mathcal{L}_{S_B}(\mathbf{S}_B, x_B^{(A)}, y_A)\\&= \lambda\|\mathbf{S}_A(x_A; \theta_A) - Y_A\| + (1-\lambda)\|\mathbf{S}_B(x_B^{(A)}; \theta_B) - Y_A\|\end{aligned} \tag{1}$$

where $\lambda \in [0, 1]$ is a coefficient that adjusts the information weight from each phase. Note that we jointly use two segmentation models in the modeling of $f(\cdot)$. Here, a question is raised: how to get the unobserved data $x_B^{(A)}$ for the multi-phase segmentation? In the next section, we will present the solution by building the relation between phases.

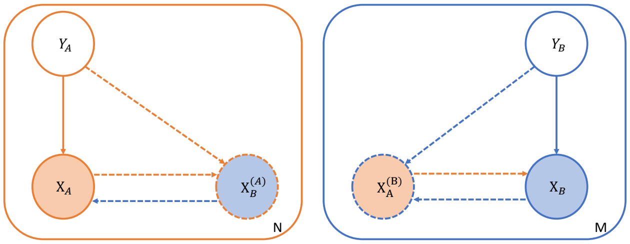

Figure 2: (*Best viewed in color*) The relation between different phases: for data and label in phase $A$ (**left**) and phase $B$ (**right**), there exists latent data variables to be inferred (dashed line), which can be inferred using style transfer.

### 3.2.2 Building Phase-to-Phase Relations

In the previous part, we applied the unobserved data $x_B^{(A)}$ in segmentation to build the data-to-label relation. In this part, we further discuss how to build the phase-to-phase relation so as to infer the unobserved data.

Similar to the concept in domain adaptation, we consider the semantic labels as the content and the images present the content with phase-specific style. Considering $x_B^{(A)} \sim p_{\text{data}}(X_B|X_A, Y_A)$, in order to generate $x_B^{(A)}$, we can transfer $x_A$ to $x_B^{(A)}$, and keep the semantic label invariant to $y_A$. In this sense, suppose we have two segmentation models that respectively work well in phase $A, B$, if we only transfer the style between $A$ and $B$, these segmentation models should be able to make the same prediction. Mathematically, we define two well trained segmentation models in each phase and a translator that only transfer style from $A$ to $B$ as $\mathbf{S}_A(\cdot; \theta_A^\star)$, $\mathbf{S}_B(\cdot; \theta_B^\star)$, $\mathbf{F}(\cdot; \phi_{AB}^\star)$, we should have:

$$x_B^{(A)} = \mathbf{F}(x_A; \phi_{AB}^\star); \quad \mathbf{S}_A(x_A; \theta_A^\star) = \mathbf{S}_B(x_B^{(A)}; \theta_B^\star) \tag{2}$$

In this way, we can build a relation between phase $A$ and $B$. This relation is bidirectional, thus we can also apply another translator $\mathbf{G}$ to transfer from $B$ to $A$.

The loss function in this part involves the generation loss of $\mathbf{F}$, which can be presented in the form of a GAN loss [11] and the segmentation loss using the generated data $\mathbf{F}(x_A)$:

$$\begin{aligned}&\mathcal{L}_F(\mathbf{F}, \mathbf{S}_B, x_A, x_B, y_A)\\&= \mathcal{L}_{\text{GAN}}(\mathbf{F}, \mathbf{D}_B, x_A, x_B) + \mathcal{L}_{S_B}(\mathbf{S}_B^\star, x_B^{(A)}, y_A)\end{aligned} \tag{3}$$

where $\mathbf{D}_B$ is an adversarial discriminator that aims to distinguish $\mathbf{F}(x_A)$ from real samples $x_B$. Note that the phase-to-phase relation partially depends on the data-to-label relation. Plus, the segmentation model $\mathbf{S}_B$ and the translator $\mathbf{F}$ are supposed to be well trained. Thus in the training stage we need train them separately at the beginning, which we will discuss in section 3.3.

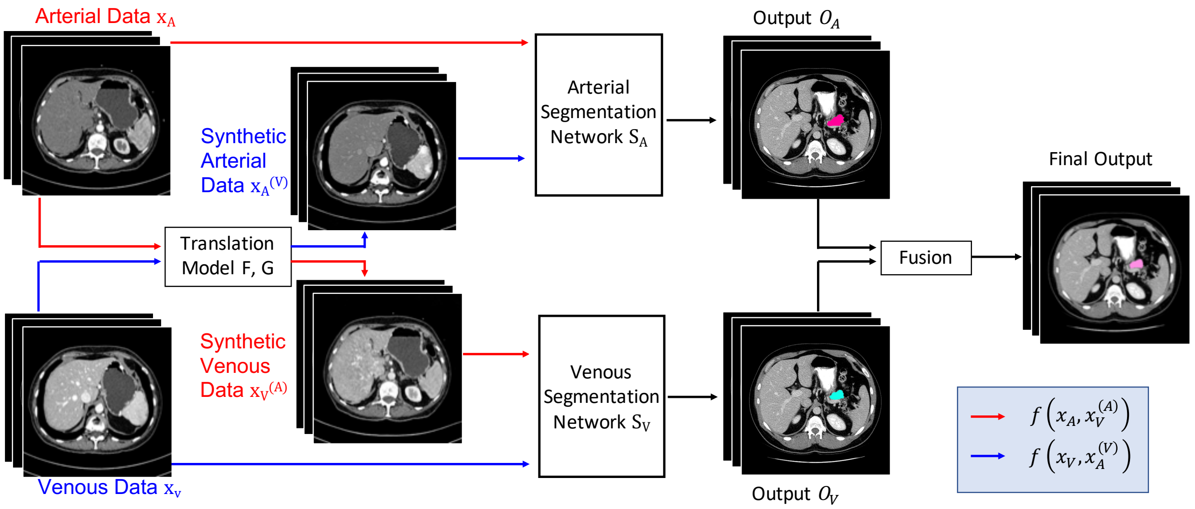
Arterial Data $x_A$
Synthetic Arterial Data $x_A^{(V)}$
Arterial Segmentation Network $S_A$
Output $O_A$
Final Output
Translation Model F, G
Fusion
Synthetic Venous Data $x_V^{(A)}$
Venous Segmentation Network $S_V$
Output $O_V$
Venous Data $x_v$
$f\left(x_A, x_V^{(A)}\right)$
$f\left(x_V, x_A^{(V)}\right)$

Figure 3: (*Best viewed in color*) Illustration of the whole framework in the medical imaging scenario. The red (*resp.* blue) line indicates the forward process supervised with input in the arterial (*resp.* venous) phase. The fuse operation is presented in Equation 1.

### 3.2.3 The Overall Framework

Combining the above components yields the proposed Phase Collaborative Network (**PCN**), which consists of four parts, namely, two discriminative models $\mathbf{S}_A$ and $\mathbf{S}_B$ for semantic segmentation in phases $A$ and $B$, as well as two generative models ($\mathbf{F}$ and $\mathbf{G}$) for translation between phases $A$ and $B$, respectively, as shown in Figure 3.

Combining Equation (1), (3) and considering both phases $A, B$, the overall optimization goal is:

$$\begin{aligned}&\mathcal{L}(\mathbf{S}_A, \mathbf{S}_B, \mathbf{F}, \mathbf{G}, x_A, y_A, x_B, y_B)\\ &= \mathcal{L}_{\text{seg}}(\mathbf{S}_A, \mathbf{S}_B, x_A, y_A) + \mathcal{L}_{\text{seg}}(\mathbf{S}_A, \mathbf{S}_B, x_B, y_B)\\ &+ \mathcal{L}_{\text{GAN}}(\mathbf{F}, \mathbf{D}_B, x_A, x_B) + \mathcal{L}_{\text{GAN}}(\mathbf{G}, \mathbf{D}_A, x_A, x_B)\end{aligned} \tag{4}$$

## 3.3. Optimization

In this paper, we instantiate PCN using two state-of-the-art approaches, namely, RSTN [47] as a discriminative model and CycleGAN [50] as a generative model. RSTN is a coarse-to-fine segmentation model that consists of two segmentation networks for coarse and fine level segmentation. These two models are separately optimized at the first two stages and jointly optimized at the final stage. CycleGAN is an unsupervised generative model that works in a cyclic manner with a pair of GAN losses. Jointly training these two models is equivalent to optimizing the objective in Equation (4).

**The training phase** aims at training the framework with data in both phases and involves two stages. As we take CycleGAN [50] as the translator, given the data in phase $A$ and $B$ ($x_A, y_A, x_B, y_B$), the full objective in Equation (4) can be specified as:

$$\begin{aligned}&\mathcal{L}(\mathbf{S}_A, \mathbf{S}_B, \mathbf{F}, \mathbf{G}, x_A, y_A, x_B, y_B)\\ &= \mathcal{L}_A(\mathbf{S}_A, \mathbf{S}_B, x_A, y_A, x_B^{(A)}) + \mathcal{L}_B(\mathbf{S}_A, \mathbf{S}_B, x_B, y_B, x_A^{(B)})\\ &+ \mathcal{L}_{\text{CycleGAN}}(\mathbf{D}_A, \mathbf{D}_B, \mathbf{F}, \mathbf{G}, x_A, x_B)\end{aligned}$$

where $\mathcal{L}_{\text{CycleGAN}}(\mathbf{D}_A, \mathbf{D}_B, \mathbf{F}, \mathbf{G}, x_A, x_B)$ is the CycleGAN objective that consists of a pair of GAN loss and cycle-consistency loss [50].

Note that in Equation (3), building the phase-to-phase relation needs the segmentation network to be well trained; meanwhile, the translator should also provide proper transferred image so that the training of segmentation network will not collapse. The training phase thus consists of two stages:

- **The separate stage**: we set $\lambda = 1$ in Equation (1) for PCN's discriminative models. Thus, the segmentation models are only trained with real data $x_A, y_A$ in phase $A$ and $x_B, y_B$ in phase $B$; the transfer model is trained in a unsupervised way as CycleGAN.
- **The joint stage**: we set $\lambda \in (0, 1)$. In this case, PCN starts to incorporate information in different phases in its discriminative module to make prediction; meanwhile, the generative module starts to address the semantic consistency to ensure spatial alignment between images in different phases.

**The testing phase** follows the flowchart as in Figure 3 to predict the final result using information from both phases, but we don't need data from both phases. PCN takes the input $x_A$ (*resp.* $x_B$), the generative module generates the unobserved data $x_B^{(A)}$ (*resp.* $x_A^{(B)}$) and make a prediction with its discriminative module as shown in Equation (1).

**In a single-phase dataset**, we only have data and label in one phase (*e.g.*, $A$). PCN is still able to train a segmentation model in phase $B$ to enable phase-specific information to be incorporated. We assume that, with a translation model $\mathbf{F}^\star$ that well models $p_{\text{data}}(X_B|X_A, Y_A)$, the generated data $x_B^{(A)}$ can well represent the information from phase $B$. Therefore, for a single-phase dataset, we need a wellpre-trained PCN translation module to help the training for $\mathbf{S}_A$ and $\mathbf{S}_B$.

### 3.4. Connection to Domain Adaptation

Domain adaptation is critical to deal with problems in unseen environments. In recent years, adversarial adaptation models are widely applied and achieved great success in transferring knowledge from a known domain to a unknown domain. As our approach brings in the idea of domain adaptation, in this part we discuss the relation of our approach to some typical domain adaptation methods.

CycleGAN [50] is a widely applied method that addresses the issues of unpaired image-to-image translation. This model assumes there exists a translation between data distribution in two domains $p_{\text{data}}(X_A)$ and $p_{\text{data}}(X_B)$. Thus a pair of GAN models [11] are applied to model $p_{\theta_A}(X_A|X_B)$ and $p_{\theta_B}(X_A|X_B)$ for the translation, and the cycle-consistency is proposed to ensure the content is invariant in translation. Compared to CycleGAN, our method models $p_{\theta_A}(X_B|X_A, Y_A)$ and $p_{\theta_B}(X_A|X_B, Y_B)$ to ensure the content invariance and spatial alignment.

CyCADA [16] aims to improve the content invariance by bring in the task-level consistency. This approach addresses in a scenario where we are given data and label in source domain, but in target domain we only have data. With this domain adaptation method, great success is achieved in transferring knowledge from source domain to assist the task in target domain. Compared to CyCADA, our method can improve the task performance in both directions. That is, we can not only assist the task in the target domain, but can also apply the information from target domain back to source domain. If PCN models the phase-to-phase relation only from source domain to target domain, the PCN will degenerate to a model equivalent to CyCADA. Moreover, PCN can be used for single-domain data to help the tasks on most of existing datasets.

### 3.5. Application to Medical Imaging Segmentation

Now, we investigate the real-world scenario that motivates this formulation, *i.e*, medical imaging analysis, or specifically, organ segmentation in abdominal CT scans. In this problem, multi-phase information is often useful, while most existing approaches were based on a single (*e.g.*, venous) phase [47]. This is mainly due to two reasons. First, most public datasets appear in a single phase. Second, even if multiple phases are available, both image data and annotations are difficult to be aligned, as shown in Figure 1.

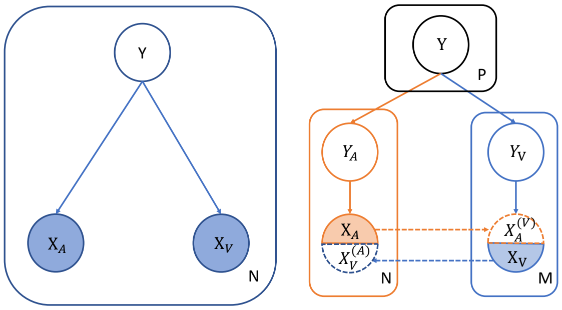

Figure 4: (*Best view in color*) Plate notation of the ideal multi-phase data generative process (**left**) and the real-world multi-phase data generative process (**right**). The dash line marks the data should exist but cannot be sampled due to the real-world scanning mechanism.

Figure 4 shows the problem in the plate notation. Ideally, the observations in both phases are supposed to be well aligned (left part in Figure 4), and we aim at training a model $f$ that can infer the semantic label from observations $X_A$ and $X_V$: $\hat{Y} = f(X_A, X_V; \theta)$. Unfortunately, as shown in the right part of Figure 4, the observed data are often unpaired due to the time gap between arterial and venous scans. Since $Y_A$ and $Y_V$ are independently sampled from the semantic information distribution $p_{\text{data}}(Y)$, it is difficult to guarantee $Y_A = Y_V$. Intuitively, during the CT scanning process, as the abdominal organs and vessels dilate or contract from time to time, we can hardly get paired $x_A$ and $x_V$ that share the same annotation $y$.

Theoretically, there should exist images in another phase that correspond to the current semantic annotation, but we cannot observe these images due to the time interval between arterial and venous phase. As shown in Figure 4, the dashed semicircles present the data that cannot be scanned in real world. To fix the gap, we can apply PCN to build relations between these variables for the multi-phase segmentation $\hat{Y}_A = f(X_A, x_V^{(A)}; \theta)$ or $\hat{Y}_V = f(X_V, x_A^{(V)}; \theta)$. Even for medical datasets in single phase, a nice property of PCN is that it can be applied to these single-phase datasets to assist the medical imaging analysis.

## 4. Experiments

### 4.1. Datasets, Evaluation and Details

We collect a multi-phase dataset that contains CT scans with 11 abdominal organs and 5 vessels on 200 normal and 200 pancreas-abnormal patients in the arterial and the venous phase. Moreover, we also evaluate our approach on

Table 1: Comparison of test segmentation accuracy (DSC) by RSTN [47] and our approach on our dataset

| | Organ | *Adrenal G.* | *Gallbladder* | *I. V. C.* | *Kidney L.* | *Kidney R.* | *Pancreas* | *S. M. A.* |
|---|---|---|---|---|---|---|---|---|
| Normal | RSTN-A | 59.40% | 87.20% | 73.62% | 94.28% | 95.13% | 84.58% | 74.48% |
| | RSTN-V | 56.11% | 87.19% | 78.77% | 94.26% | 92.10% | 86.94% | 71.67% |
| | PCN-A | **64.96%** | 87.58% | 77.09% | **94.44%** | **95.81%** | 84.89% | **79.30%** |
| | PCN-V | 64.00% | **88.16%** | **81.42%** | 93.33% | 95.49% | **88.20%** | 74.36% |
| Abnormal | RSTN-A | 58.14% | 80.59% | 73.20% | 92.09% | 94.45% | 80.32% | 66.28% |
| | RSTN-V | 52.60 | 86.10% | 78.19% | 92.78% | 90.26% | 75.89% | 62.72% |
| | PCN-A | **63.57%** | 86.19% | 74.70% | 93.01% | **94.58%** | **81.48%** | **69.75%** |
| | PCN-V | 59.90% | **89.28%** | **79.43%** | **95.51%** | 94.00% | 79.82% | 64.40% |

two public datasets, which are in single phase: the NIH pancreas segmentation dataset [34] and the medical decathlon pancreas dataset[1], which respectively contain 82 and 282 contrast-enhanced abdominal CT volumes and corresponding annotations. The resolution of each scan is $512 \times 512 \times L$, where $L$ is the number of slices along the long axis of the body, and $L \in [381, 1089]$ (Our dataset), $L \in [181, 466]$ (NIH) and $L \in [37, 751]$ (Medical Decathlon) .

Following the standard cross-validation strategy, we split the dataset into 4 folds, each of which contains approximately the same number of samples. We apply cross validation, i.e., training the models on 3 out of 4 subsets and testing them on the remaining one. We measure the segmentation accuracy by computing the Dice-Sørensen coefficient (DSC) [38] for each sample, and report the average and standard deviation over all cases.

To be fair in comparison, we follow the original papers to construct CycleGAN [50] and two RSTN [47] for both phase. With a learning rate of $10^{-5}$, we set PCN in the separate training stage and run 80,000 iterations to train RSTN in its separate C2F stage; then run 10,000 iteration for RSTN's joint C2F stage. After this, we run PCN in the joint training stage for another 30,000 iteration, and every 10,000 steps we apply a learning iteration decay with a scale of $0.8$. Typically, it spend three to four times longer in the joint training stage than in the separate stage. To train a PCN model on 200 patient cases, normally it takes 40 hours.

### 4.2. Segmentation on Multi-phase Datasets

We take RSTN [47] as the baseline. Respectively we train and evaluate RSTN and our framework on our dataset. We choose targets including the abdominal organs and blood vessels that represent different degree of difficulty in segmentation. Each organ is trained and tested individually.

The results are summarized in Table 1. We can note that the segmentation models trained in the venous phase achieve better results for some organs like Pancreas, Inferior. V. C., *etc.*, while for some organs like Duodenum, Superior. M. A., *etc.*, the segmentation accuracy is better if the model is trained in the arterial phase. As for our framework, we can observe for all the organs and vessels, either in the arterial phase or the venous phase, better segmentation results are achieved, which verifies that our framework can incorporate the information from another phase to improve the segmentation accuracy.

Table 2: Test segmentation accuracy (DSC) comparison between our approach and the state-of-the-arts on NIH pancreas dataset and Medical Decathlon pancreas dataset.

| Data | Approach | Average | Max | Min |
|---|---|---|---|---|
| NIH | Roth et al. [34] | 78.01% | 88.65% | 34.11% |
| | Zhou et al. [49] | 82.37% | 90.85% | 62.43% |
| | Yu et al. [47] | 84.50% | 91.02% | 62.81% |
| | PCN | **85.15%** | **94.68%** | **68.89%** |
| MD | Yu et al. [47] | 73.38% | **87.54%** | 35.53% |
| | PCN | **76.59%** | 86.23% | **57.29%** |

### 4.3. Segmentation in Single-phase Datasets

To justify our approach is applicable to most of existing datasets in the real world, we evaluate the effectiveness of PCN on the NIH pancreas segmentation dataset [34] and the medical decathlon pancreas dataset. As these public datasets are in venous phase, we apply a pre-trained generator $G_{\mathrm{V2A}} : V \rightarrow A$ on our dataset to PCN and fix it during the training. The discriminative modules of PCN are trained from scratch.

In Table 2, we show that our approach works better than the baseline models, i.e., the state-of-the-art approaches in single-phase [34, 49, 47]. As shown, the average improvement over 82 cases on NIH pancreas dataset is 0.65%. With our case-by case study, compared to RSTN [47], which achieves the best results among the baseline methods, PCN outperforms than RSTN on 74 cases.

To investigate how PCN benefits from the arterial phase, we studied the prediction of the segmentation network that takes generated input. On average, the segmentation network taking generated arterial data provides 4814 voxels as auxiliary prediction, where 2258 voxels possess intersection

[1] http://medicaldecathlon.com/index.html

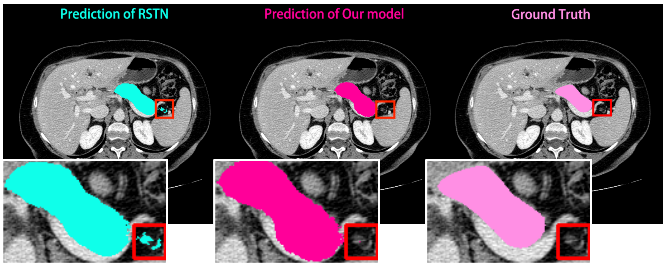

Figure 5: (*Best view in color*) Visualization of the segmentation in the axial view.

with labeled voxels, and 553 effective voxels are accepted in the final prediction. On another public dataset, the Medical Decathlon Pancreas Dataset, we compare with the results with RSTN [47]. Our approach shows an improvement of 3.21% on this dataset. The segmentation network taking the generated arterial data provides 4224 effective auxiliary prediction voxels on average, 929 of which are accepted in the final prediction.

From these experiments, we justify the effectiveness of our approach on single-phase dataset. With a well trained PCN translation module, our approach is able to generate the data in the style of another phase, thus benefit the training of the segmentation network in another phase to achieve the multi-phase segmentation.

## 4.4. Diagnostic Experiments

### 4.4.1 Quality of Image Translation

In this section, we apply the statistical diagnosis on the quality of latent data inference. That is to say, we justify: whether the style transfer provides the information from another phase.

As described in Section 3.2.2, the translation model aims to approximate the distribution of data in a specific phase $p_{\text{data}}(X_A)$ and $p_{\text{data}}(X_V)$. To evaluate the quality of translation, we use ground-truth segmentation mask on the training and testing data to obtain HU distributions of different organs on original/generated-arterial/venous data. For each organ (*e.g.* pancreas), we take all voxels labeled as this organ, and plot out the corresponding distribution. Each organ have 4 distributions: the distribution in arterial phase $p_{\text{data}}(X_A, Y_A)$, the distribution in venous phase $p_{\text{data}}(X_V, Y_V)$, the generated arterial data distribution $p_G(X_A|X_V, Y_V)$ and the generated venous data distribution $p_F(X_V|X_A, Y_A)$. With these data, we can understand whether the current transfer model learns these transfer information correctly.

In Figure 6, we present the pancreas HU distribution and the visualization of the image translation. We can observe that for the origin data in both phases, the distributions shows different appearance. Correspondingly, the arterial and venous CT scans possess different characteristics. From the distribution, we can observe the transfer model has succeeded in approximating the distribution to the original data distribution. From the CT scans, we can also observe that the translation models indeed captures some phase-specific information, for example, the intensity of liver (the organ on the upper-left corner) is higher in the venous phase and lower in the arterial phase.

### 4.4.2 Qualitative Visualization

In Figure 5, we show a typical example on how multi-phase information improves segmentation. Our single-phase baseline, RSTN [47], produces a false-positive area in *pancreas* segmentation, and it is filtered out while we take complementary information from the other phase into consideration. In some other organs or soft tissues, the complementariness of multi-phase information is also significant, *e.g.*, in the context of blood vessels, *arteries* are often easier to recognize in the arterial phase, while *veins* are easier in the venous phase. PCN provides a practical solution to fuse multi-phase information especially when there is not sufficient information in one phase.

### 4.4.3 Difference from Data Augmentation

A natural question comes here: for our approach, is there any difference or relation to the data augmentation? In this section, we give a quick discussion on this question.

We first apply 100 case arterial and venous data to train a RSTN model, respectively; then we randomly pick out 50 cases from both phases to form a new dataset for training; finally we combine all data together to form a dataset of 200 cases to form another training set. The test set is the same as in section 4.2. Table 3 shows the test results of models trained on different training sets. We can observe: (1) as the training sets consist both arterial and venous phase data, the test results are not necessarily improved *w.r.t.* the result from model trained on arterial/venous phase data; (2) when we extend the training set size, although the test results are better, compared to the results in Table 1, these test results are not comparable to those produced by models trained on single-phase data. Thus, incorporating information from another phase is not equivalent to the data augmentation.

We observed that combining data from different phase, if the size of training set is enlarged, the performance of the model might be improved. However, this also increases the variance of the training set, which requires the model to be more powerful to handle the data variance. Thus, our approach differs simply combining data from different phase for data augmentation, but to incorporate the useful information so as to improve the segmentation.

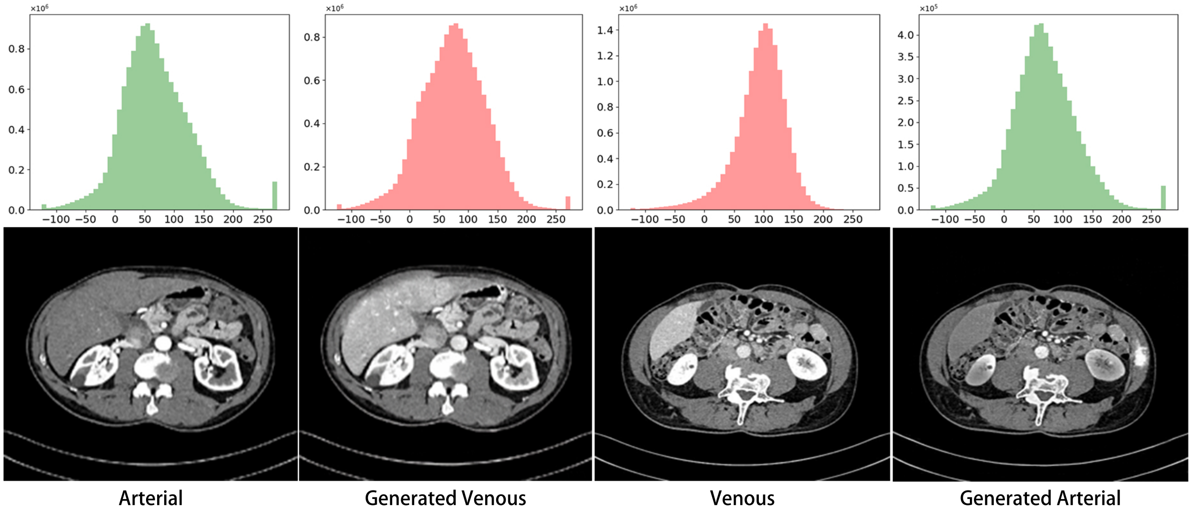

Figure 6: (*Best viewed in color*) HU distribution of pancreas area and visualization on real and generated CT scans from arterial/venous phases.

Table 3: Test segmentation accuracy (DSC) comparison of RSTN with different training set size.

| Organs | *I.V.C.* | *Pancreas* | *S.M.A* |
|---|---|---|---|
| Arterial (100 cases) | 67.5% | 79.3% | 68.4% |
| Venous (100 cases) | 70.9% | 78.3% | 64.6% |
| Ar+Ve (50+50 cases) | 67.3% | 74.5% | 65.4% |
| Ar+Ve (100 + 100 cases) | 74.5% | 81.2% | 72.7% |

### 4.4.4 Ablation: Unsupervised Domain Adaptation

We further investigate the difference between our approach and the unsupervised domain adaptation approach [50, 16].

We pick two organs where PCN has large improvement in segmentation accuracy. We use four different training configurations: (1) PCN on multi-phase dataset; (2) PCN on single-phase dataset (we suppose the data on another phase is unknown, thus we apply a pretrained PCN translation module and fix it during training); (3) we fix PCN in separate stage during training, which means the translation module is trained in an unsupervised way (marked as UDA-multi); (4) we apply a pretrained CycleGAN to PCN's translation model for single-phase datasets.

The results are summarized in Tabel 4. We compare PCN with standard unsupervised domain adaptation, and PCN outperforms than the unsupervised adaptation methods in modeling $p_{\text{data}}(X_B|X_A, Y_A), p_{\text{data}}(X_A|X_B, Y_B)$ in both multi-phase and single-phase scenario, since PCN is a more attention on semantic consistency.

## 5. Conclusions

In this paper, we present Phase Collaborative Network (PCN) to deal with the difficulty in multi-phase segmentation especially in the subarea of medical imaging analysis. From a theoretical analysis, we argue that the difficulty mainly lies in the gap between images from different views. Then, we formulate this problem into modeling two relations, namely, data-to-label relations and phase-to-phase relations, for which we design discriminative and generative models, respectively. The entire network is optimized in an end-to-end manner, and achieves satisfying performance on both single-phase and multi-phase datasets. Confirmed by the radiologists in our team, these segmentation results are helpful to computer-assisted clinical diagnoses.

Table 4: Test segmentation accuracy (DSC) comparison with different training configuration.

| Phase | Arterial | | Venous | |
|---|---|---|---|---|
| Organs | *I.V.C* | *S.M.A* | *I.V.C* | *S.M.A* |
| PCN-multi | 77.09% | 79.30% | 81.42% | 74.36% |
| PCN-single | 74.32% | 75.28% | 78.96% | 73.62% |
| UDA-multi | 73.85% | 74.43% | 78.80% | 73.14% |
| UDA-single | 73.14% | 72.01% | 72.66% | 68.11% |

The success of our approach lays the foundation of integrating multi-phase data into various vision problems. However, as a preliminary study, PCN still suffers some drawbacks. For example, PCN currently adopts a GAN-based method as the generative model, but such model can be heavily constrained by the domains in training data, which limits its application to other organs, *e.g.*, a generative model pre-trained in abdominal CT scans is unlikely to transfer to brain CT scans. Also, PCN assumes that both phases contain part of known information, but, in real applications, possibly, one or more phases are not equipped with any annotations. It is possible to integrate PCN with weakly-supervised or semi-supervised approaches in these scenarios. All these topics are left for future research.

## Acknowledgements

This paper was supported by the Lustgarten Foundation for Pancreatic Cancer Research. We thank Qing Liu and Zihao Xiao for enormous help in improving the quality of this paper. We also appreciate Wei Shen, Seyoun Park, Yan Wang, Yuyin Zhou, Chenxi Liu, Zhuotun Zhu, Yingda Xia, Fengze Liu, Jie Chang, Xu Chen, Maosen Li and Jiangchao Yao for instructive discussions.

## References

[1] A. M. Ali, A. A. Farag, and A. S. El-Baz. Graph cuts framework for kidney segmentation with prior shape constraints. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 384–392. Springer, 2007.

[2] A. J. Asman and B. A. Landman. Non-local statistical label fusion for multi-atlas segmentation. *Medical image analysis*, 17(2):194–208, 2013.

[3] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.

[4] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4545–4554, 2016.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[7] C. Chu, M. Oda, T. Kitasaka, K. Misawa, M. Fujiwara, Y. Hayashi, Y. Nimura, D. Rueckert, and K. Mori. Multi-organ segmentation based on spatially-divided probabilistic atlas from 3d abdominal ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 165–172. Springer, 2013.

[8] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016.

[9] A. Dembo, T. M. Cover, and J. A. Thomas. Information theoretic inequalities. *IEEE Transactions on Information Theory*, 37(6):1501–1518, Nov 1991.

[10] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423. IEEE, 2016.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014.

[13] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. 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] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.

[17] S. Hu, E. A. Hoffman, and J. M. Reinhardt. Automatic lung segmentation for accurate quantitation of volumetric x-ray ct images. *IEEE transactions on medical imaging*, 20(6):490–498, 2001.

[18] J. Huang and S. You. Point cloud labeling using 3d convolutional neural network. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2670–2675. IEEE, 2016.

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[20] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[21] D. Koller, N. Friedman, and F. Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[23] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[24] D.-T. Lin, C.-C. Lei, and S.-W. Hung. Computer-aided kidney segmentation on abdominal ct images. *IEEE transactions on information technology in biomedicine*, 10(1):59–65, 2006.

[25] H. Ling, S. K. Zhou, Y. Zheng, B. Georgescu, M. Suehling, and D. Comaniciu. Hierarchical, learning-based automatic liver segmentation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[26] M. G. Linguraru, J. A. Pura, A. S. Chowdhury, and R. M. Summers. Multi-organ segmentation from multi-phase abdominal ct via 4d graphs using enhancement, shape and

location optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 89–96. Springer, 2010.

[27] M. G. Linguraru, J. K. Sandberg, Z. Li, F. Shah, and R. M. Summers. Automated segmentation and quantification of liver and spleen from ct images using normalized probabilistic atlases and enhancement estimation. *Medical physics*, 37(2):771–783, 2010.

[28] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[29] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[30] P. H. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *31st International Conference on Machine Learning (ICML)*, 2014.

[31] P. O. Pinheiro, R. Collobert, and P. Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015.

[32] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1(2):4, 2017.

[33] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[34] H. R. Roth, L. Lu, A. Farag, H. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[35] B. Sandberg, S. H. Kang, and T. F. Chan. Unsupervised multiphase segmentation: A phase balancing model. *IEEE Transactions on Image Processing*, 19(1):119–130, Jan 2010.

[36] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell. Clockwork convnets for video semantic segmentation. In *European Conference on Computer Vision*, pages 852–868. Springer, 2016.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, 5:1–34, 1948.

[39] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Deep end2end voxel2voxel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–24, 2016.

[40] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[41] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):611–623, 2013.

[42] Z. Wang, K. K. Bhatia, B. Glocker, A. Marvao, T. Dawes, K. Misawa, K. Mori, and D. Rueckert. Geodesic patch-based segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 666–673. Springer, 2014.

[43] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

[44] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert. Automated abdominal multi-organ segmentation with subject-specific atlas generation. *IEEE transactions on medical imaging*, 32(9):1723–1730, 2013.

[45] Y. Yang, Y. Zhao, B. Wu, and H. Wang. A fast multiphase image segmentation model for gray images. *Computers & Mathematics with Applications*, 67(8):1559 – 1581, 2014.

[46] Z. Yi, H. R. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2868–2876, 2017.

[47] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. *Computer Vision and Patter Recognition*, 2018.

[48] Y. Zhou, L. Xie, E. K. Fishman, and A. L. Yuille. Deep supervision for pancreatic cyst segmentation in abdominal ct scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 222–230. Springer, 2017.

[49] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille. A fixed-point model for pancreas segmentation in abdominal ct scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017.

[50] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.